Slide 1

Approccio infrastrutturale di una applicazione Web per la gestione di testi di notevole difficoltà interpretativa e loro traduzioni

Andrea Bozzi

Premessa.

La linguistica e la filologia computazionali si sono occupate solo molto di recente del problema della gestione comparativa computerizzata di testi e loro traduzioni, soprattutto in seguito alla spinta che è stata data alla ricerca in tal senso dal tema del riconoscimento, grazie a sistemi automatizzati, dei fenomeni di plagio (*plagiarism recognition)*. Tecniche specifiche dedicate non solo, ma principalmente, allo studio dei testi in traduzione sono state sviluppate per scoprire eventuali stili (e, di conseguenza, autori) diversi che si celano sotto la traduzione di un medesimo testo. Tali indagini sono per lo più fondate su tecniche statistiche o stocastiche[1] per le quali è possibile realizzare specifici algoritmi computabili da una macchina, con risultati spesso molto soddisfacenti.

Il fatto che si parli di tecniche statistiche o stocastiche farebbe pensare ad una connaturata applicabilità su domini molto vasti, compresi quelli delle scienze testuali e letterarie. Si potrebbe, cioè, inferire che l'astratto approccio matematico risulti indipendente da contenuti testuali specifici e dai condizionamenti diacronici che caratterizzano opere e traduzioni. La mia esperienza personale mi permette di affermare che, se da un lato questi metodi hanno decisamente un ruolo importante nella Linguistica computazionale (si pensi, per esempio, al *machine learning*), non altrettanto si può dire qualora siano adoperati per analisi specialistiche di testi antichi, soprattutto quelli caratterizzati da notevole difficoltà interpretativa. Quei modelli e quelle tecniche producono risultati spesso poco affidabili per gli studi di filologi, editori critici o storici del pensiero antico.

I due ultimi grandi progetti che ho avuto la fortuna di ottenere prima di lasciare l'Istituto di Linguistica Computazionale del CNR in occasione del mio pensionamento mi hanno confermato quanto già avevo avuto modo di constatare: in certe condizioni il *Computing in the Humanities* non può fare a meno di alternare procedure automatiche, affidando dei compiti al software, a sistemi semiautomatici o, nel peggiore dei casi, a pratiche completamente manuali affinché la comunità degli utenti specialisti non resti delusa e non rifiuti pregiudizialmente la tecnologia.

[1] I modelli stocastici (stocastico = dovuto al caso, aleatorio, dal greco stochastikòs=congetturale) tengono in considerazione le variazioni (causali e non) delle variabili di input, e quindi forniscono risultati in termini di "probabilità". È importante sottolineare che ciò che differenzia i modelli deterministici da quelli stocastici è che in questi ultimi si tiene conto della variabilità dei dati di input. In genere i modelli stocastici hanno una struttura più complessa di quelli deterministici. Di maggiore complessità sono i calcoli, che vengono eseguiti sempre con l'ausilio del computer. Esistono anche applicazioni dedicate specificamente a questo scopo. Ovviamente i modelli stocastici sono anche più affidabili in quanto, proprio perché tengono conto del caso, sono capaci di fornire risultati più aderenti alla realtà.

Questa visione prudenziale si adatta bene ad una modalità innovativa nella produzione di applicativi Web: l'approccio infrastrutturale. Esso differisce, come vedremo, dalle architetture un po' rigide che sottostanno ai programmi di gestione dei testi, la maggior parte dei quali sono realizzati per svolgere un compito ben preciso con preclusione sia verso una loro più vasta applicabilità sia all'utilizzo di specifiche interfacce che favoriscono l'intervento diretto da parte dell'utente specialista al quale l'ambiente computazionale è offerto.

Slide 2

Ma perché adottare un approccio infrastrutturale nella progettazione di una applicazione Web per la gestione di testi e di traduzioni?

- In primo luogo perché esso garantisce la possibilità di far interagire fra di loro i membri di una stessa comunità scientifica nel settore delle scienze umane allo stesso modo di come avviene in altri settori del sapere, come, per esempio, nella biologia molecolare o nella astrofisica. Non a caso questi àmbiti sono stati i primi a proporre progetti di tipo infrastrutturale all'Unione Europea, progetti che sono poi stati inseriti nella *Roadmap* per le *Research Infrastructures* dell'Unione (ESFRI). Lo scopo che si intende raggiungere è quello di mettere dati, software e sistemi prodotti da alcuni a disposizione di molti, ovvero di intere comunità di studiosi, al fine di impedire dispersione di risorse economiche e di facilitare il raggiungimento di risultati condivisi.
- Tale approccio, inoltre, garantisce anche una più stretta relazione fra membri di diverse comunità di ricercatori facenti parte del variegato mondo delle scienze umane tradizionalmente marcate da confini disciplinari molto netti e oltrepassati solo in situazioni molto particolari. Per esempio: un testo o un corpus di ricette medico-farmaceutiche medievali in lingua occitana può essere di estremo interesse anche per ebraisti medievali in quanto in quei dati non è escluso che si possano trovare soluzioni di lettura ed interpretazione di termini tecnici (medici, farmaceutici, botanici) non riconosciuti perché rari e non vocalizzati nella fonti manoscritte.
- In terzo luogo una prospettiva infrastrutturale rende interoperabili dati e sistemi prodotti in progetti diversi e realizzati per scopi diversi. Per esempio, un algoritmo di attribuzione di POS (*part of speech*) o un sistema di analisi morfologica e di lemmatizzazione può essere applicato per lingue flessive diverse.
- Ancora, esso garantisce riusabilità dei componenti realizzati per scopi specifici in altri àmbiti delle *Digital Humanities.* Per esempio, un sistema di annotazione del testo eventualmente affiancato da uno schema di classificazione ontologica (concettuale, semantica) di ogni singola annotazione trova una vasta gamma di utilizzi in filologia, archeologia, filosofia, ecc.

- Infine, l'infrastruttura di ricerca è sempre aperta ad accogliere nuovi interventi e integrazioni che aumentano il valore e l'estensione applicativa di tutti i componenti dei quali essa è costituita.

In questo seminario vorrei presentarvi il modello[2] che ho disegnato per la progettazione di un'applicazione Web di *Textual scholarship* da utilizzare nello studio di opere di particolare difficoltà interpretativa, con l'ulteriore complicazione di dover eventualmente trattare, oltre al testo originale, una traduzione di esso che: a) sia stata già realizzata; b) si stia o si debba ancora effettuare.

Slide 3

Il superamento della specificità e la garanzia di ottenere un'applicazione non limitata, ma estendibile si ottiene se il modello soddisfa almeno 5 condizioni:

- modularità: interazione di più moduli software in una architettura a componenti;
- condivisione: utilizzazione da parte di più studiosi operanti in forma collaborativa;
- flessibilità: utilizzazione da parte di ricercatori che operano su testi trasmessi da più fonti, su documenti unici, su manoscritti di autori moderni e contemporanei e, infine, su testi a stampa;
- standardizzazione: utilizzo di sistemi di sviluppo del software e di marcatura dei testi riconosciuti come standard internazionali o ad essi riconducibili;
- produzione *open source* del codice sorgente, condizione indispensabile per poter realizzare le fasi di verifica anche da parte di utilizzatori diversi da coloro che hanno contribuito allo sviluppo del progetto.

In sintesi, il progetto si basa su un modello di sviluppo fondato su cinque principi: modularità, funzione collaborativa, flessibilità, uso (o adeguatezza a) di sistemi di marcatura e linguaggi standard, filosofia di sviluppo basata sul principio dell' *open source*. Grazie ad essi e grazie alla continua evoluzione della tecnologia correlata al Web si prefigura la costituzione di una batteria di strumenti di elaborazione, che potremmo definire col termine di infrastruttura tecnologica per le *digital humanities*.

[2] In questa mia esposizione io adopero il termine modello in un'accezione che richiama la prima parte di quanto detto al proposito da von Neumann ne *I fondamenti della fisica quantistica* (1932): "Le scienze non cercano di spiegare, a malapena tentano di interpretare, ma fanno soprattutto modelli", ma si discosta dalla seconda parte ("Per modello si intende un costrutto matematico che, con l'aggiunta di interpretazioni verbali, descrive dei fenomeni osservati. La giustificazione è soltanto e precisamente che funzioni, descriva correttamente i fenomeni in un'area abbastanza ampia, e soddisfi dei criteri estetici, cioè deve essere piuttosto semplice"), in nome della specificità non meccanicistica e non algoritmizzabile delle discipline filologico-letterarie, storiche e filosofiche. Si noti l'importanza attribuita da von Neumann al modello quando è in grado di descrivere fenomeni che entrano in gioco in un'area abbastanza ampia, fenomeni che, nel nostro caso, si organizzano vantaggiosamente in una dimensione infrastrutturale, coerente con la varietà e molteplicità delle componenti convergenti negli studi della filologia del testo.

Slide 4

Vorrei soffermarmi su una rappresentazione grafica del sistema infrastrutturale e dei moduli attualmente previsti, alcuni dei quali sono stati effettivamente realizzati per due importanti progetti che coinvolgono attività di ricerca su testi di particolare difficoltà interpretativa e delle loro traduzioni, dei quali parlerò in seguito.

Al centro del grafico si colloca il progetto che partecipa ed usufruisce dei componenti (o moduli) infrastrutturali ad esso funzionali. E' bene precisare a questo proposito che:

- lo schema attuale indica un buon numero di componenti i quali, tuttavia, non esauriscono tutte le esigenze di gestione computerizzata dei dati che utenti di tipo diverso possono richiedere per raggiungere compiutamente i risultati previsti;
- lo schema attuale, in base al principio di modularità, presenta componenti attivi, componenti latenti ed è predisposto per accogliere altri eventuali componenti purché il loro sviluppo rispetti l'altro principio di base rappresentato dalla standardizzazione e dalla filosofia dell'*open source*.

Fatta questa premessa, esaminiamo ora i singoli componenti visibili sulla slide.

Text manager. Non mi soffermo su questo aspetto e rinvio ai paragrafi 5, 6 e 7 delle fotocopie che vi sono state consegnate. Esse sono tratte dal Capitolo 8 del recentissimo *Manuel de la Philologie de l'édition*, a cura di David Trotter, pubblicato da De Gruyter nella serie "Manuals of Romance Linguistics". Vorrei solo ribadire qui due osservazioni.

La prima riguarda la necessità di valutare con grande attenzione quali siano gli elementi che valga realmente la pena di marcare affinché il sistema informatico possa effettuare selezioni e ricerche mirate. Non esiste, infatti, un criterio generale che stabilisca come e che cosa si debba codificare: il processo di marcatura è una variabile dipendente dalle finalità della ricerca.

La seconda osservazione consiste in una conferma della validità nell'utilizzare sistemi di codifica standardizzati una volta che si sia deciso quali elementi sia indispensabile marcare. Non è obbligatorio l'utilizzo di XML-TEI; è tuttavia consigliabile, qualora vi siano ragioni per l'adozione di altri tipi di marcatura, rendere questi ultimi compatibili con il linguaggio XML-TEI che si è diffuso e affermato a livello internazionale.

Slide 5, 6, 7

Per esempio, il testo marcato si riferisce ad un progetto di moltissimi anni fa relativo allo spoglio elettronico del corpus dei *Grammatici Latini Antichi* (ed. Keil e altre): il sistema di codifica rappresentato in alto era basato sulla procedura di spoglio dell'ILC-CNR di Pisa, in seguito automaticamente trasformato nel linguaggio RTF (al centro), a sua volta ricodificato con procedure automatiche in XML-TEI da colleghi di un istituto del CNRS di Parigi (*Laboratoire d'Histoire des Théories Linguistiques*) ai quali l'archivio testuale fu ceduto, su accordo con l'Università degli Studi

di Torino. Grazie a queste marcature il sistema di indicizzazione dei dati testuali consente di effettuare interrogazioni sui vari livelli nei quali il testo è stato organizzato (testo propriamente detto, autori citati, titoli di opere, citazioni in prosa, citazioni in poesia, ecc.). La procedura è stata utilizzata da filologi latini anche per facilitare il riconoscimento di citazioni implicite che spesso i grammatici inserivano senza indicarne l'origine.

Slide 8

Image manager. Il grande sviluppo della tecnologia digitale e della riproduzione delle fonti manoscritte o librarie da parte di Enti pubblici e privati ha aperto una serie di nuove possibilità di indagine agevolando, in molti casi, la lettura e l'interpretazione dei testi. Anche a questo riguardo non intendo soffermarmi sugli aspetti di *digital image processing* sui quali trovate riflessioni al paragrafo 4 del citato mio contributo sul manuale di De Gruyter. Vorrei solo sottolineare il fatto che i sistemi tecnologicamente avanzati di segmentazione delle immagini, soprattutto relative a libri a stampa antichi e a fonti manoscritte nelle quali la distribuzione del testo segua un andamento piuttosto regolare, offrono la possibilità di ottenere concordanze testo/immagine molto interessanti.

Tali procedure, infatti, consentono di riconoscere, soprattutto nelle liste di parole a frequenza 1, errori di lettura facilmente rilevabili confrontando la parola trascritta con la corrispondente zona nell'immagine digitale.

Slide 9 e 10

In questi due esempi, tratti, il primo, da un lavoro sperimentale svolto alcuni anni fa per l'Istituto papirologico G. Vitelli di Firenze nell'ambito di un progetto MiUR/MiBAC e, il secondo, da un progetto per l'Università di Pisa e Gottinga, si può valutare concretamente il vantaggio offerto da un modulo che faccia cooperare l'elaborazione del testo e l'elaborazione delle immagini. Anche in questo caso, tuttavia, è bene ripetere quanto abbiamo già osservato a proposito della marcatura e della codifica: è indispensabile valutare il rapporto costi e benefici nell'attivazione di un simile componente i cui risultati possono essere soggetti a notevoli interventi umani, soprattutto in situazioni complesse come nel caso della papirologia o dei documenti che hanno subito gravi danni alla superficie scrittoria.

Slide 11

Computer assisted translation. Questo modulo è quello sviluppato più di recente e si può affermare che la sua realizzazione ha comportato anche l'inserimento di altri componenti nell'infrastruttura che sono ad esso strettamente connessi, ma che, per il criterio generale esposto in precedenza, possono essere attivati ciascuno indipendentemente dagli altri. Si tratta di: ***Advanced search***, ***Linguistic analysis***, ***Semantic annotation***, e ***Desktop publishing exporter***.

Negli esempi che seguiranno nelle prossime slide, cercherò di mettere in risalto, nell'ambito di uno sguardo di insieme, i contributi dei singoli moduli, in special modo quello delle annotazioni (*Semantic annotation*).

Trascuro qui di parlare del modulo *optical character recognition* perché su di esso ho potuto effettuare solo esperimenti che non ebbero un seguito per interruzione dei finanziamenti da parte dei Progetti finalizzati del CNR, anche se riuscii a brevettare un sistema di intelligenza artificiale basato su reti neurali chiamato LAperLA (*Lettore Automatico per Libri Antichi*). Informazioni dettagliate a questo proposito si leggono in un volume curato da A. Bozzi, *Computer-aided recovery and analysis of damaged text documents*, Bologna, CLUEB, 2000.

Potrei affermare che il modulo di assistenza per i testi in traduzione ai quali si affianchi l'originale costituisce una sotto-infrastruttura della quale vengono a far parte la maggioranza dei moduli utilizzati per l'analisi computerizzata dei testi e altri moduli specifici per l'analisi del testo tradotto.

É opportuno sottolineare che questo aspetto del *text processing* e delle Linguistica computazionale non è stato oggetto di studi da parte dei maggiori Centri che operano nel settore delle *Digital Humanities* e del *Textual Scholarship*. Solo di recente, come anticipato all'inizio, il tema ha conosciuto un notevole sviluppo nell'ambito delle indagini stilometriche utili anche per rilevare la eventuale presenza di più mani nel corpo di una stessa traduzione.

All'ILC-CNR di Pisa ebbi modo di ampliare lo spettro di funzionalità che avevo previsto per l'infrastruttura Web (poi denominata *TS_app - Textual Scholarship Web App*) grazie alla partecipazione attiva a due importanti progetti.

- **Il primo progetto** (ERC Advanced Grant "Greek into Arabic. Philosophical Concepts and Linguistic Bridges") richiedeva una piattaforma Web per la gestione comparativa del testo di alcuni capitoli della Enneadi di Plotino e quello corrispondente della traduzione araba (circa IX sec.) trasmessa col nome di *Teologia di Aristotele*.
- **Il secondo progetto**, invece, ci impegnava per un'applicazione Web grazie alla quale circa 50 fra traduttori e revisori avrebbero cooperato, anche simultaneamente, per rendere in italiano contemporaneo il testo del *Talmud* babilonese in ebraico/aramaico.

Dunque, due progetti dei quali il primo presentava una traduzione già eseguita molti secoli fa in una lingua non occidentale, il secondo una traduzione da doversi ancora realizzare sulla base di un testo ebraico/aramaico, sedimentatosi in uno spazio temporale molto vasto e con la necessità

di estrarne glossari tematicamente distinti per soggetti (amministrazione, medicina, giurisprudenza, botanica, ritualità/liturgia, religiosità, ecc.).

Il modello infrastrutturale che avevo progettato e che vi ho mostrato prima si rivelò perfettamente idoneo a questo doppio compito grazie, soprattutto, al principio della modularità e flessibilità che già abbiamo discusso.

In particolare, per le attività del primo progetto, i componenti principali attivati sono evidenziati nella slide seguente:

Slide 12

I moduli nei riquadri colorati indicano che l'infrastruttura mette a disposizione le funzioni da essi espresse per le attività richieste dagli scopi del progetto. Gli elementi di dettaglio e le specificità sono stati discussi e realizzati in stretta collaborazione con il PI (*Principal Investigator*) del progetto, nella persona di Prof. Cristina D'Ancona, Università di Pisa e del Prof. Endress dell'Università di Bochum (Associated partner).

Vediamo ora alcune schermate della TS_app relativi ai due progetti citati.

Primo progetto: "Greek into Arabic"

Tutte le schermate che vi mostrerò sono ricavate direttamente dalla applicazione e sono dettagliatamente commentate in Marchi S., *Greek into Arabic, a research infrastructure based on computational modules to annotate and query historical and philosophical digital texts. Part II. System components and features*, in Bozzi, A. (ed.), *Digital Texts, Translations, Lexicons in a multi-modular Web Application: Methods and Samples*, Firenze, Olschki, 2015, pp. 43-56. I criteri e il modello generale, invece, sono descritti in Bozzi, A., *Greek into Arabic, a research infrastructure based on computational modules to annotate and query historical and philosophical digital texts. Part I. Methodological aspects*, in Bozzi, A. (ed.), *Digital Texts, Translations, Lexicons in a multi-modular Web Application: Methods and Samples*, Firenze, Olschki, 2015, pp. 27-42.

Io mi limiterò qui a descriverne gli aspetti principali.

Slide dalla 13 alla 19

Commentare le slide facendo notare in particolare le funzioni del modulo ***Advanced search*, *Linguistic analysis, Semantic annotation***.

Slide 20

Nel progetto "Greek into Arabic", inoltre, è stato realizzato anche il modulo **Computational Lexicon**. A differenza di quello che il nome potrebbe suggerire a chi non ha esperienze nel settore

della linguistica computazionale o della filosofia del linguaggio che, in questo caso specifico, risultano entrambi coinvolte, si tratta di un modulo per la descrizione ontologicamente strutturata dei valori semantici espressi dai lemmi. Dal momento che questo tema esula dagli argomenti del seminario, invito gli eventuali interessati a leggere il contributo di Ruimy e Piccini nel volume di A. Bozzi (ed.), *Digital Texts, Translations, Lexicons in a multi-modular Web Application: Methods and Samples*, Firenze, Olschki, 2015. Ho dato una copia del volume al Prof. Mordenti.

Lasciare in evidenza ancora la slide 20

Per quanto riguarda, invece, il modulo **Linguistic analysis** non bisogna considerare che esso svolga solo le funzioni di un analizzatore morfologico (con attribuzione dei POS alle forme considerate) e/o di lemmatizzatore, funzioni sulle quali non mi soffermo in quanto esse sono ormai note a tutti coloro che abbiano fatto un po' di esperienza su archivi testuali computerizzati.

Merita, tuttavia, di sottolineare qui un ruolo non molto noto, ma altrettanto importante, che esso può svolgere se integrato con un componente di tipo statistico-linguistico grazie al quale fornisce ipotesi utili all'integrazione di parole frammentarie presenti in un testo danneggiato.

Slide dalla 21 alla 24

Questo ulteriore modulo, da me sperimentalmente realizzato molti anni fa per una collaborazione con l'Istituto Papirologico Vitelli di Firenze, potrebbe ora venire ingegnerizzato dall'ILC-CNR ed inserito come modulo aggiuntivo nella stessa infrastruttura per un nuovo progetto ERC che ci si augura possa venire approvato dopo le prime fasi positive di valutazione.

Secondo Progetto: Traduzione Talmud babilonese

Il modello infrastrutturale dimostra la propria validità anche per il secondo grande progetto sulla prima traduzione integrale in italiano del Talmud babilonese. Riguardo alle problematiche generali del Talmud in rapporto alle attività computazionali previste, si rimanda a Bellusci, A., *Towards a translation platform as a bridge between ancient and modern languages. Part I. The Babylonian Talmud: a Web of knowledge between past and present*, in Bozzi, A. (ed.), *Digital Texts, Translations, Lexicons in a multi-modular Web Application: Methods and Samples*, Firenze, Olschki, 2015, pp. 57-67.

Slide 25

Tutti i moduli, tranne quelli contrassegnati dalle avvertenze "work in progress" e "future work", sono attivati. In particolare i moduli *Computer Assisted translation, Semantic annotation* e *Desktop publishing exporter* sono indispensabili. Si tratta di un progetto nato in collaborazione fra il CNR, il MiUR, la Presidenza del Consiglio dei Ministri, l'UCEI (Unione delle Comunità ebraiche italiane), il Collegio rabbinico ed è presieduto dal rabbino capo di Roma Rav Riccardo Di Segni.

Tutte le schermate riprodotte sono ricavate direttamente dall'applicazione e sono dettagliatamente commentate in Bellandi A., *Towards a translation platform as a bridge between ancient and modern languages. Part II. A research infrastructure for translation and interpretation of ancient texts*, in Bozzi, A. (ed.), *Digital Texts, Translations, Lexicons in a multi-modular Web Application: Methods and Samples*, Firenze, Olschki, 2015, pp. 69-83. Io mi limiterò qui a descriverne gli aspetti principali.

Slide dalla 26 alla 30

Commentare le slide (sottolineare il modulo *Semantic annotation*)

Slide 26

Per ciascun componente del sistema (i riquadri in blu) implementa specifiche funzioni dedicate a soddisfare le esigenze di differenti tipologie di utilizzatori. Traduttori e revisori sono assistiti nel processo di traduzione da tecnologie di *Computer assisted translation*, compresi Indicizzatori e *Translation memory* (TM); filologi e linguisti esperti di Talmud possono inserire note, commenti, annotazioni semantiche e riferimenti bibliografici; i curatori dei volumi possono produrre l'edizione a stampa della traduzione del Talmud in maniera semplicissima, sistemando traduzioni e note in formati standard che sono gestiti da software per il *desktop publishing*; gli esperti di dominio, ovvero chi possiede conoscenze specialistiche sui temi e i contenuti testuali del Talmud, sono in grado di organizzare i termini significativi in glossari e, nel caso sia necessario, in ontologie di dominio; infine, i ricercatori possono effettuare interrogazioni complesse su base linguistica (significante), semantica (significato) e ontologica (concetto).

Slide 27

Qui viene presentata l'interfaccia grafica principale per gli operatori delle traduzioni. Il traduttore seleziona il capitolo del trattato che gli è stato assegnato e che egli deve tradurre (parte a). Una parte del testo originale viene selezionata e copiata nella *translation table* (parte b). Il sistema acquisisce la stringa da tradurre, interroga la TM e suggerisce le traduzioni italiane delle stringhe più simili ad essa. A destra (parte c) si trova, segnalata dal sistema con 5 stelline, la migliore traduzione dopo che esso ha consultato la TM: il traduttore la sceglie ed essa viene collocata nella *translation table* (si veda la parte highlighted). 5 stelle: suggerimento perfetto; 4 stelle: poche correzioni sono richieste al traduttore per migliorare il suggerimento; ecc. Si noti che ogni correzione effettuata dal traduttore sui suggerimenti dati dal sistema viene aggiunta alla TM che, di conseguenza, accresce l'insieme delle traduzioni disponibili per nuovi suggerimenti.

Era molto importante, inoltre, permettere ai traduttori di rispettare nel testo italiano la struttura interna del Talmud. Per questa ragione la *translation table* presenta campi differenziati secondo la seguente struttura: trattato, capitolo (verde), blocco (pesca), unità logica (azzurro), stringa o segmento (bianco).

I traduttori/revisori/ricercatori, inoltre, possono distinguere le traduzioni letterali (caratterizzandole col neretto) dalle parti esplicative o aggiuntive (stile normale).

Slide 28

Il sistema consente anche l'inserimento di vari livelli di annotazioni grazie alle quali gli esperti talmudisti possono rendere la traduzione italiana più comprensibile ai non esperti. Attualmente è disponibile un gruppo di 7 classi predefinite di annotazioni (ciascuna evidenziata da un colore), come si legge nel box dell'interfaccia grafica: nell'esempio, alle parole evidenziate in verde il traduttore ha attribuito l'annotazione semanatica di "concetto", mentre "Rabbì Elazar" è stato classificato con l'etichetta "rabbino". Questa funzione consente, inoltre, di costituire glossari specializzati; quando sarà effettuata la completa implementazione di tale componente, si potrà procedere alla annotazione automatizzata e alla conseguente creazione automatica dei glossari.

Slide 29

Inoltre abbiamo previsto, sulla base degli elementi concettuali espressi grazie al sistema di annotazione semantica, una *Talmudic knowledge base* (base di conoscenza semantico-concettuale talmudica, ovvero una ontologia talmudica) usando un linguaggio di rappresentazione formale della conoscenza. A tale scopo, è in fase di sviluppo ormai avanzato uno strumento di tipo grafico per assistere gli esperti del dominio (il Talmud) nel definire le relazioni gerarchiche e associative fra le parti di testo semanticamente annotate, le quali denotano specifiche entità, mettendo a loro disposizione un sistema collaborativo per la rappresentazione grafica della struttura della conoscenza.

In **questa slide** si vede una prova di questa rappresentazione grafica. La figura si riferisce a uno strumento che permette di:

- costruire una base di conoscenza semantica talmudica, da parte di un esperto;
- consultare il testo sulla base della conoscenza semantica creata precedentemente, sia da parte di esperti sia da un più vasto pubblico in possesso di inferiori livelli di competenze (per es.: studenti delle scuole rabbiniche).

Tramite le annotazioni effettuate su porzioni di testo sulla base di alcune classi precostituite e la creazione di relazioni semantiche tra esse, si consente ad uno studioso del Talmud di navigare il testo su base semantica oltrepassando i limiti delle ricerche effettuate con parole o parti di parole. Nell'esempio in figura ci si riferisce al *Trattato dei sogni* che descrive l'interpretazione dei sogni dei rabbini fornita loro dai Maestri. Un operatore di una certa esperienza ha annotato porzioni di quel testo, marcandole con le opportune classi semantiche, per esempio:

- Il testo che descrive il sogno è stato annotato con la classe "Sogno";

- Il testo che descrive l'interpretazione del sogno è stato annotato con la classe “Interpretazione”;
- I nomi dei rabbini sono stati annotati con la classe “Rabbino”
- I nomi dei Maestri sono stati annotati con la classe “Maestro”;
- Le parti del corpo nel testo sono state annotate con la relativa classe (naso, braccio, ecc.);
- Sono state create relazioni gerarchiche tra le parti del corpo;
- Sono state create relazioni associative tra rabbini e maestri (discepolo, ha detto a, ecc....);
- Sono state create relazioni tra sogni e persone (visto in sogno, interpretazione del sogno, ecc.).

E’ stato quindi sviluppato un piccolo prototipo di navigazione grafica di tale conoscenza, che ha permesso di partire da una classe semantica, "esplodere" le proprie relazioni, sceglierne una, continuare la navigazione semantica, esplodere altre relazioni e così via, creando dinamicamente un grafo in base ai percorsi semantici interessati.

Partendo, per esempio, da un nodo che rappresenta una parte del corpo, collegando ad esso un nodo “sogno” e a quest'ultimo un nodo “interpretazione” e poi un nodo “maestro” (con le relative relazioni semantiche) e accedendo al testo del nodo d'arrivo, si ottengono tutte le parti del testo relative alla descrizione delle interpretazioni di sogni fatte da un certo maestro e che parlano di una specifica parte del corpo.

Slide 30

Il modulo di **Desktop publishing** merita una serie di osservazioni. L’infrastruttura, che ora possiamo chiamare Textual Scholarship Research Infrastructure, considera sempre indispensabile l’esportazione dei dati e dei risultati dell’attività di ricerca su documenti cartacei oltre che, naturalmente, sul Web. A tale scopo la marcatura in linguaggio XML soggiacente viene interpretata dal XSL (eXtensible Stylesheet Language) che produce la trasformazione dei file XML originari in un altro documento (per esempio, un file PDF) secondo le desiderate indicazioni di formattazione e layout. Questa strategia permette di esportare direttamente il file della traduzione italiana ed eventuali altre parti (commento, glossario, note, ecc.) in un formato di stampa *camera ready*.

Slide 31

Scolarly editing manager. Questo modulo meriterebbe un seminario a parte perché è forse il settore che presenta un grado di complessità di progettazione e di sviluppo maggiore rispetto agli altri. Esso si riferisce ad un vero componente per l’editoria critica digitale con un campo di applicazioni non limitato alla filologia antica o medievale, ma che potrebbe articolarsi in sotto-moduli per *la critique génetique* e la filologia del testo a stampa. La difficoltà consiste principalmente nel disegnare un modello di riferimento che non derivi direttamente da quelli

studiati da Lachmann o Bedier, ma che ad essi faccia comunque riferimento affinché rimangano vivi almeno due elementi fondamentali, almeno per me, della critica del testo: lo *stemma codicum* e la *constitutio textus*. Il fatto che le fonti sulle quali il filologo intende operare siano in formato digitale non ci autorizza a demolire i paradigmi teorici sui quali la filologia del testo si basa. Ritengo molto discutibili, nonostante l'interesse che hanno rappresentato negli anni passati, le esperienze di Peter Robinson o di Cerquiglini: il primo ipotizzò la sostituzione dello *stemma* e, di conseguenza, la *constitutio* con una ragnatela di relazioni spazio-temporali fra i codici, rappresentabili mediante un programma di tipo filogenetico derivato da metodiche della biologia molecolare. Il secondo, facendo leva sulla incommensurabile memoria della tecnologia informatica, propose di immettere in rete immagini e trascrizioni di tutti i testimoni di un testo, considerando praticamente superate le basilari operazioni di *collatio*, *divinatio*, *eliminatio codicum descriptorum*, valutazione e analisi delle eventuali interpolazioni, ecc.. Egli, in sostanza, riteneva un inaccettabile artificio quello di risalire al testo sulla base dei testimoni che nel tempo e nello spazio lo hanno trasmesso inserendovi modificazioni conscie o inconscie ed errori meccanici di varia origine. Quello che colpiva Cerquiglini, ma non era l'unico, era il fenomeno del bifidismo degli stemmi che, secondo il suo parere, ripetendosi quasi sempre nella storia della tradizione manoscritta delle opere medievali, faceva insorgere il dubbio di trovarsi di fronte ad un difetto metodologico, solo parzialmente corretto dall'ipotesi di lavoro di Bédier. Quest'ultimo, infatti, aveva optato per la scelta di un *bon manuscrit* sulla base del quale effettuare il confronto con quanto trasmesso dagli altri testimoni recensiti. (Si veda *Cerquiglini B., l'Éloge de la variante. Histoire critique de la philologie*, del 1989).

Il modello fino ad oggi da me disegnato consiste in una posizione in un certo senso intermedia fra i due sistemi, come descritto in molti miei lavori. Qui vorrei analizzare solo una schermata estratta da un modulo sperimentale che sul quel modello si basa e sul quale il gruppo di ricerca sulla Filologia digitale, che avevo a suo tempo costituito presso l'ILC, sta ancora lavorando. In seguito ad una indispensabile fase di sperimentazione su dati effettivi da parte di giovani editori critici (l'invito è aperto a eventuali dottorandi interessati) il componente potrà essere inserito nell'infrastruttura.

Slide 32, 33. Commentare le slide.

Slide da 34 a 37: Bibliografia citata

Osservazioni conclusive.

Il modello descritto per una infrastruttura di ricerca per lo studio filologico del testo, del testo e sue traduzioni, del testo e la rappresentazione digitale delle immagini sulle quali esso è documentato ed, infine, per alcune analisi di tipo linguistico che possono essere condotte con l'ausilio di uno strumento computazionale, costituisce una prospettiva tecnologicamente molto innovativa nel settore delle *Digital Humanities*. Il fatto che al modello siano seguiti concreti

componenti software effettivamente realizzati e funzionanti su progetti di notevole impegno ed importanza mi confermano che la strada intrapresa è quella corretta.

Ora è necessario aumentare la sperimentazione dei moduli previsti nell'IR, ma non ancora compiutamente realizzati. Il primo su cui credo valga la pena dedicarsi è lo *Scholarly editing manager*; per questo invito chiunque di voi fosse interessato a partecipare attivamente alla sperimentazione a mettersi in contatto con l'attuale responsabile di questo àmbito di ricerca presso l'ILC-CNR di Pisa, il Dr. Emiliano Giovannetti (Emiliano.giovannetti@ilc.cnr.it). Sulla base di accordi istituzionali fra la vostra Università ed il CNR è certamente possibile far partecipare giovani studiosi con esperienza nell'editoria critica al perfezionamento di questa parte dell'IR non ancora sufficientemente definita e messa alla prova dei fatti.