How to reference this article
Bozzi, A. (2020). PTTB e DiTMAO: modularità di alcune applicazioni per le Digital Humanities. *Italica Wratislaviensia*, 11(1), 101–121.
DOI: http://dx.doi.org/10.15804/IW.2020.11.1.04

Andrea Bozzi
già Direttore ILC-CNR-Pisa, Italia
andrea.bozzi@cnr.it
ORCID: 0000-0002-4627-7174

PTTB E DITMAO: MODULARITÀ DI ALCUNE APPLICAZIONI PER LE DIGITAL HUMANITIES

PTTB AND DITMAO: THE MODULARITY OF SOME TOOLS FOR DIGITAL HUMANITIES

Abstract: For several years now, computational linguistics has been addressing the problems of and developing technological tools for automatic translation, with its important economic implications. At the same time, projects dedicated to facilitating translations of ancient works, which are often fraught with considerable hermeneutical difficulties, are far rarer. The PTTB system, which was designed and constructed at the Institute for Computational Linguistics (National Research Council) in Pisa, enables a group of about fifty scholars to translate the entire Babylonian Talmud, written in Aramaic and Biblical Hebrew, more quickly and uniformly. While the language and structure of the textual corpus made the development of machine translation algorithms impossible, translation memory and edit distance techniques have produced excellent results. Based on them, the system offers scholars a high percentage of correct translations, accessible through a very intuitive graphic user interface. The results are easily exportable to xml files suitable for the final editing and printing operations. So far, these innovations have made it possible to publish four treatises in six printed volumes with translations, annotations and thematic indexes within a relatively short time. Several other volumes have already been processed and are currently being edited. Various perspectives open up for the use of the digital Talmud in Italian. One of the most interesting options involves using machine learning and named entity recognition techniques to associate semantic or conceptual values (Talmud Ontological Framework) with and make cross-references among portions of the text that report or discuss similar themes. This will help various groups of (general and specialised) users to browse this vast and heterogeneous textual archive on the semantic basis. The strategy adopted here is also aligned with the Dictionnaire des Termes Médico-botaniques de l'Ancien Occitan (DiTMAO), another ongoing lexicographical project. It will enable users to semantically navigate within an extensive medical-pharmaceutical and botanical textual corpus in medieval Occitan. For these reasons, PTTB and DiTMAO can be regarded as two instances of one innovative technological infrastructure for linguistic and philological research in the field of digital humanities.

Keywords: Talmud, computational linguistic, digital philology, TM (translation memory), Old Occitan, ontologies

Received: 01/02/2020; Accepted: 15/06/2020; Published: 31/07/2020
ISSN 2084-4514 e-ISSN 2450-5943

1. DIGITAL HUMANITIES E LINGUISTICA COMPUTAZIONALE

La tecnologia digitale si è diffusa in modo sempre più capillare anche nel settore delle scienze umane tanto che sono ormai molti anni che si parla di *Digital Humanities*. Questa definizione non rappresenta un settore disciplinare specifico, dal momento che essa copre una molteplicità di àmbiti dalla storia alla filologia, dalla linguistica ai beni librari, dall'archeologia alla paleografia, e così via per molti altri ancora. Certamente un passo significativo verso l'innovazione tecnologica e l'adozione di strumenti di elaborazione elettronica dei dati fu fatto dalla linguistica già a partire dagli anni '50 del secolo scorso. Essa si mosse in questa direzione, fra gli altri motivi, anche per: – costituire grandi archivi di testi, le cosiddette banche di dati testuali; – produrre strumenti di analisi del testo, soprattutto letterario e in dimensione sia sincronica che diacronica (*Machine Readable Textual Archives* e *Concordances*); – rilevare fenomeni morfologici o morfo-sintattici utili per verificare ipotesi o teorie relative al funzionamento dei meccanismi di produzione delle parole (*Morphological Analyzer*) o degli enunciati (*Formal Grammars*, *Transition Networks*).

Il principale interesse era rivolto, come avviene tuttora, alle lingue moderne anche perché i risultati ottenuti hanno consentito di progettare algoritmi di traduzione automatica, settore considerato strategico per evidenti ragioni di opportunità, dalla Commissione Europea, e pertanto finanziato per un lungo periodo. I corpora paralleli di testi in due o più lingue hanno in tal senso svolto un ruolo di primaria importanza, nonostante che gli sforzi economici dedicati a questa tematica non siano probabilmente giustificabili in rapporto ai risultati prodotti.

A fronte del dinamismo qui esposto in maniera limitata e assai schematica, l'innovazione tecnologica applicata a lingue e culture antiche, nonostante il pionieristico lavoro di Padre Busa s.J. per l'*Index Thomisticus*, ha dovuto attendere l'affermazione planetaria del Web per affrontare i problemi di carattere critico testuale, talvolta molto complessi, che le singole filologie (latina, greca, romanza, slava, semitica, ecc.) pongono agli studiosi. In effetti, le tecniche ben collaudate nella Linguistica Computazionale (LC) come, per esempio, i programmi per

concordanze, indici, analisi morfologica e lemmatizzazione, pur producendo materiale di lavoro indispensabile, costituiscono una porzione molto limitata di attività nel complesso armamentario di cui un filologo necessita, soprattutto quando egli si appresti a studiare un testo antico e le fonti che lo tramandano. Se poi all'edizione critica si accompagnano commenti esegetici propedeutici alla traduzione, e se tutto questo vede la presenza di più ricercatori cooperanti in rete nel seno di uno stesso progetto, allora è indispensabile studiare soluzioni tecnologiche di notevole complessità.

2. IL PROGETTO TALMUD

Un caso emblematico che vede intrecciati aspetti di LC, Filologia computazionale (FC) e sviluppo di applicazioni collaborative per il Web è rappresentato dal Progetto Traduzione del Talmud Babilonese (PTTB)[1]. Si tratta di un progetto finanziato dal Ministero dell'Università e della Ricerca (MiUR), sulla base di un protocollo di intesa fra la Presidenza del Consiglio dei Ministri, il MiUR stesso, il Consiglio Nazionale delle Ricerche (CNR), l'Unione delle Comunità Ebraiche Italiane/ Collegio Rabbinico Italiano (UCEI/CRI), che ha lo scopo di eseguire la traduzione integrale di tutti i trattati del *Talmud* (TB)[2]. L'edizione curata da Adin Steinsaltz e disponibile in formato digitale costituisce il testo aramaico/ebraico di riferimento[3].

---

[1] Il *Talmud* (TB) è la *Torah* orale, in quanto registra le discussioni che avvennero nelle accademie di Babilonia tra il III e il V sec. della nostra era, discussioni basate sulla *Mishnah*, che a sua volta è la codificazione scritta della legislazione ebraica prodotta fra il II e il III sec.. Esiste anche un *Talmud* di Gerusalemme, più vecchio, composto in Terra di Israele occupata dai Romani. Per una dettagliata, chiara ed esaustiva storia del *Talmud*, si veda Freedman, 2014.

[2] Si veda https://www.talmud.it/ (ultimo accesso: dicembre 2019).

[3] *The Talmud: the Steinsaltz edition*, New York: Random House, 1989–2000; Paris: Ramsay, 1994–2010. In lingua italiana esistono solo antologie la più significativa delle quali è A. Cohen (a cura di). *Il Talmud*, Laterza, Roma–Bari, 1999 (8°edizione 2018), realizzata sulla base di quella inglese dello stesso Cohen, *Everyman's Talmud*, London, J.M. Dent & Sons, ltd, 1932. Il progetto TB ha preso avvio nel 2011 dopo che il Governo Italiano stanziò un finanziamento per contribuire alla realizzazione di

La traduzione in italiano di TB presenta alcuni aspetti particolari: a) essa è affidata ad un cospicuo numero di specialisti, circa 50, attivi contemporaneamente sul corpus memorizzato in un *server* sempre accessibile via Web[4]. Ne consegue che vi è un concreto rischio di disomogeneità nella traduzione anche di passi simili o identici; b) TB presenta un linguaggio formulare talvolta criptico, ricco di espressioni standardizzate e di citazioni interne ed esterne, cosa che comporta l'inserimento di parti aggiuntive per rendere comprensibile ad un lettore italiano non specialista la traduzione comunque eseguita letteralmente; c) i passi che presentano identica la parte di traduzione letterale possono differire a causa delle informazioni contestuali integrative introdotte da traduttori diversi; d) TB è, per tali ragioni, intraducibile con sistemi di traduzione automatica, compresi quelli basati su sistemi di intelligenza artificiale che adottano modelli e tecnologie a reti neurali supervisionate[5].

---

un'iniziativa culturalmente molto rilevante; prima dell'avvio dei lavori, le parti coinvolte si sono dotate di una struttura organizzativa che garantisse da un lato il rispetto dei tempi di sviluppo, dall'altro la validità scientifica dei risultati ottenuti. Per questi motivi si sono costituiti organismi di controllo gestionale: un Consiglio di Amministrazione e Direzione del progetto (Presidente: Rav Riccardo Di Segni, rabbino capo di Roma; Direttore: Clelia Piperno) e un Comitato di coordinamento (Presidente: Riccardo di Segni, Cinzia Caporale, Vincenzo di Felice, Alessio Gorla, Davide Jona Falco, Evelina Milella, Anna Nardini, Fiamma Nirestein). Sono stati istituiti anche un Comitato tecnico–scientifico (Presidente: Andrea Bozzi, Giacomo Ferrari, Marco Mancini, Nicoletta Maraschio, Tito Orlandi, Abramo Piattelli, Giacomo Saban, Oliviero Stock) e un Comitato d'onore (Presidente: Gianni Letta, Antonio Catricalà, Adin Even Israel (Steinsaltz). Alberto Melloni, Giacomo Moscati, Ministro Università e Ricerca, Presidente del CNR, Presidente UCEI).

[4] Non è opportuno qui soffermarci a descrivere le modalità di controllo degli accessi e di permanenza dei dati, tutti aspetti particolarmente rilevanti in questo progetto, ma che sono di carattere esclusivamente tecnologico.

[5] Le regole e i modelli che sono alla base della traduzione automatica non si possono applicare, con la speranza di ottenere risultati apprezzabili, alla lingua delle opere talmudiche, in conseguenza degli aspetti di estrema sinteticità che la caratterizzano e che presuppongono una conoscenza pregressa dei temi e delle situazioni contingenti. Inoltre, non sono al momento disponibili analizzatori morfosintattici e lemmatizzatori dell'aramaico ed ebraico talmudico, in grado di attribuire alle singole parole le informazioni grammaticali grazie alle quali un sistema evoluto di traduzione automatica possa effettuare opportune operazioni di confronto con le corrispondenti informazioni

L'estensione del corpus, che oltrepassa le 7.000 pagine, il problema del contenimento dei tempi di realizzazione e la necessità di ottenere traduzioni quanto più possibili omogenee, nonostante l'elevato numero di operatori, il fatto che non siano utilizzabili sistemi di traduzione automatica, tutto ciò ha persuaso ad adottare metodi di *Computer Assisted Translation* (CAT) (Giovannetti, Albanesi, Bellandi, Benotto, 2017) comprendendovi sia indicizzatori sia, soprattutto, tecniche di *Translation Memory* (TM). Ciò costituisce una fra le principali componenti di TRADUCO, il modulo di PTTB che ha il compito di agevolare ed automatizzare il processo traduttivo (Albanesi A., Bellandi A., Benotto G., Giovannetti E., 2015; Bellandi, 2015; Bozzi, 2017).

2.1. Traduzione assistita

Traduzione assistita. TM può essere definito come un insieme di programmi che svolgono funzioni coordinate, la prima delle quali consiste nel fatto che tutte le porzioni di testo (chiamate stringhe) e tutte le relative traduzioni eseguite da tutti i traduttori sono organizzate in una base di dati opportunamente strutturata. Un secondo componente valuta la ripetizione di certe strutture linguistiche o verifica la eventuale identicità

---

della lingua italiana e, di conseguenza, avanzare ipotesi di traduzione. A questo proposito si ricorda che il *Morphological Analyser* by MILA (*Knowledge Center for Processing Hebrew*; cf. http://www.mila.cs.technion.ac.il/eng/tools_analysis.html, ultimo accesso: dicembre 2019) e lo HEBMORPH (*Morphological Analyser and Disambiguator for Hebrew Language*. Cf. http://code972.com/hebmorph: ultimo accesso: dicembre 2019) sono stati realizzati solo per l'analisi linguistica dell'ebraico moderno e risultano, pertanto, inefficaci per i testi ebraici più antichi, come TB, caratterizzati da un alto numero di varianti arcaiche. Ricordiamo, inoltre, RESPONSA, il primo sistema di indicizzazione dell'ebraico (Choueka, 1980). Il progetto ebbe inizio nella metà degli anni '60 ed aveva lo scopo di indicizzare più di 500.000 documenti religiosi, appartenenti ad un arco temporale di 1400 anni. Per quanto riguarda le reti neurali, un settore dell'Intelligenza Artificiale (IA), esse consistono di uno o più livelli di unità di calcolo di un computer (neuroni) che memorizzano parti di testo originale e corrispondenti parti di traduzione effettuate da uno specialista umano, attivano cicli di apprendimento grazie a modelli statistici (fasi di *Training*) e propongono la traduzione di una parte del testo non tradotta (fase di richiamo). Informazioni aggiornate sulle tecniche di *Natural Language Processing* anche dell'ebraico si trovano in Zitouni, 2014, il quale, tuttavia, tratta sempre di analisi automatica di lingue semitiche moderne.

di intere frasi. A tale proposito si osservi che il compito, per un *server* con buone capacità di calcolo, non è molto gravoso perché le stringhe sono piuttosto corte, la formularità del testo le riproduce con una certa frequenza e il lessico non è ricco, se paragonato a opere filosofiche o letterarie antiche di altri àmbiti linguistici (per esempio, il greco classico).

In una fase successiva, il programma esegue il confronto fra la stringa selezionata da chi si sta accingendo a tradurla e tutte quelle uguali o più simili già tradotte e memorizzate nella base di dati. Il confronto e i risultati relativi al rapporto di somiglianza viene effettuato mediante un algoritmo di *Edit Distance* (ED). Esso prende in considerazione solo le stringhe depositate in memoria che risultino identiche o più simili a quella selezionata ancora da tradurre, le valuta con un punteggio di similarità espresso numericamente (ma che poi l'interfaccia-utente rappresenta in forma di stelle, da un massimo di cinque ad un minimo di una, rispettivamente per indicare che la stringa in questione è identica o presenta soltanto alcuni elementi di similarità rispetto a quelle trovate in memoria e confrontate), ed infine propone le relative traduzioni in ordine decrescente al punteggio attribuito. A parità di punteggio, offre per prima la versione eseguita dal traduttore più accreditato[6].

L'algoritmo di ED confronta due stringhe alla volta, procedendo parola per parola, e calcola quante trasformazioni sono necessarie per ottenere la stringa 2 (S2) partendo dalla 1 (S1); costruisce quindi una matrice *M* dove ciascun elemento rappresenta il numero minimo delle mutazioni richieste per trasformare S1 in S2.

Nel corso dello sviluppo del progetto è stato possibile perfezionare l'algoritmo, introducendo elementi che ne hanno aumentato le potenzialità e l'affidabilità; il più importante calcola il peso della similarità anche in base a relazioni paradigmatiche (per esempio, la sinonimia), conseguenti ad una fase di estrazione automatica di coppie di termini sinonimici dal corpus. Ne consegue che la diversità fra termini aramaici che nei testi risultano semanticamente sinonimi ha un impatto pari

[6] La graduatoria di competenza viene stabilita dalla presidenza e direzione del progetto nonché dal coordinatore dei traduttori (Rav Gianfranco Di Segni, CNR).

a zero sull'algoritmo di ED, poiché nessuna trasformazione viene effettuata per passare dall'uno all'altro.

Nella figura 1 si vede l'interfaccia realizzata per i traduttori. Ciascuno di essi ha ricevuto la lista dei capitoli dei singoli trattati che deve interpretare e tradurre in italiano; egli sceglie la stringa (la parte evidenziata nella colonna **a**) la copia e la incolla nella parte sinistra della colonna centrale (**b**). La stringa viene numerata dal sistema (2-2a-1) e si attiva la ricerca di stringhe simili nella base di dati mediante il confronto eseguito da parte dell'algoritmo di ED. I risultati ottenuti compaiono nella zona "Suggerimenti" della colonna (**c**). disposti in ordine decrescente in base al numero delle stelle di similarità rilevata fra la stringa selezionata e quelle trovate nella base di dati.

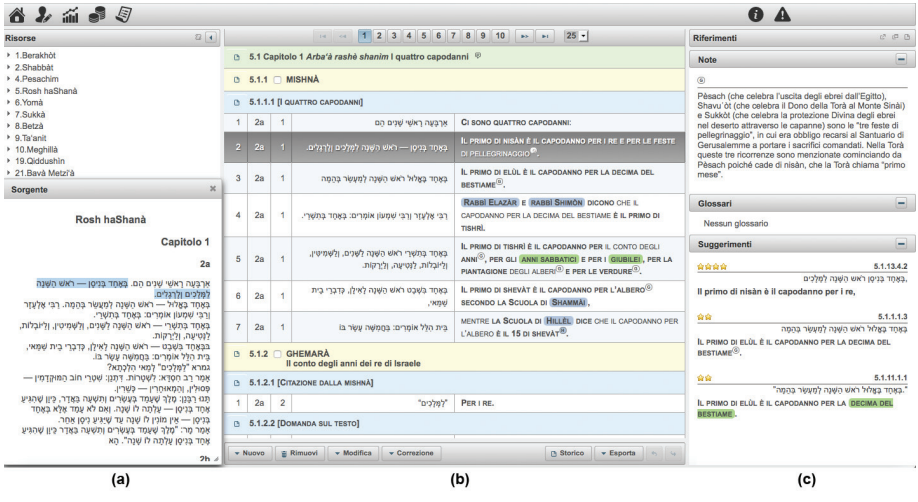

Fig. 1.

La traduzione contrassegnata dal maggior numero di stelle corrisponde alla stringa aramaica (già tradotta dallo stesso traduttore o da un qualsiasi altro traduttore in un qualsiasi testo del corpus) più simile a quella selezionata. Se la traduzione suggerita (Il primo di nisàn è il capodanno per i re) viene accolta, è sufficiente porvi sopra il cursore del *mouse* e premere il tasto di invio; essa viene automaticamente inserita nella zona riservata alla traduzione nella parte destra della colonna (**b**), ove

agiscono le regole di formattazione stabilite a priori, in base alle quali le parti relative alla *Mishnà* sono tutte in maiuscolo ed in neretto (**IL PRIMO DI NISÀN È IL CAPODANNO PER I RE**), perché questa è traduzione letterale del testo aramaico. Come si vede da questo esempio, tuttavia, il traduttore ha inserito una parte aggiuntiva, resa graficamente in stile normale, non in neretto (E PER LE FESTE DI PELLEGRINAGGIO), giudicata utile per una migliore comprensione del passo.

Questa strategia ha confermato la propria efficacia: la percentuale delle traduzioni corrette, suggerite dal sistema, è cresciuta, facilitando e velocizzando il processo interpretativo e, nello stesso tempo, garantendo omogeneità di resa in italiano nei passi aramaici simili o ripetitivi.

2.2. Annotazioni

Annotazioni. Un secondo componente di TRADUCO che merita di essere commentato riguarda le annotazioni. Come si è detto, il testo del *Talmud* in lingua italiana presenta integrazioni graficamente ben distinguibili rispetto a quello originale, che aiutano la comprensione delle parti dei singoli trattati con significati inespressi o impliciti. Si sono presentati comunque numerosi casi nei quali perfino un traduttore esperto ha ritenuto insufficiente lo strumento che gli consente di introdurre ed evidenziare le integrazioni al testo e ha pertanto richiesto un ulteriore spazio ove esprimere, ancora più compiutamente, il senso di un brano di particolare complessità. A questo fine il sistema gli mette ora a disposizione un'area dell'ambiente di lavoro, quella che in figura 1 compare in alto a destra, dedicata a annotazioni e commenti. Essi vengono valutati dal coordinamento del progetto e, se confermati, costituiranno le note al testo da convertire nella versione XML predisposta per la stampa sui volumi cartacei.

Un altro metodo per introdurre annotazioni è stato studiato e realizzato per definire una maggiore granularità semantica, indispensabile agli specialisti che saranno chiamati ad effettuare indagini su tematiche particolari, considerata la varietà e la vastità degli argomenti distribuiti sui numerosi trattati di TB.

Tali annotazioni sono costituite, per ora, da 7 classi semantiche, visibili in un menu a tendina, e ciascuna di esse è graficamente evidenziata

da un colore. Il traduttore è invitato a marcare singole parole o parti del testo tradotto che si riferiscono a: «Concetto», «Espressione linguistica», «Idioma», «Rabbino/Maestro», «Misura», «Natura», «Nome». La lista, naturalmente, può essere ampliata a seconda delle esigenze che emergano durante il lavoro.

Il fatto che le parole afferenti alla medesima classe semantica siano evidenziate con lo stesso colore, presenta due vantaggi: – favorisce una percezione immediata delle parti di testo afferenti ad uno specifico tema; – permette al sistema di produrre glossari tematici[7].

Un'ulteriore attività prevede, infine, che ricercatori di varie discipline, utilizzando specifiche funzioni di ricerca sull'archivio digitale in rete, svolte in particolare da componenti di analisi linguistica (indicizzatori, lemmatizzatori, analizzatori morfo-sintattici, ecc.) e di ricerca avanzata (concordanze, ricerca di co-occorrenze, ecc.), ottengano rapidamente i contesti che trattano situazioni o argomenti (medicina, farmacopea, diritto, religione, storia, ecc.) dei quali sono competenti, affinché vi possano associare commenti che determinano valore aggiunto a tutto il corpus, trasformandolo in un sapere talmudico enciclopedico.

2.3. Prospettive

Prospettive. La versione digitale di un corpus tanto vasto consente, come abbiamo già accennato, di effettuare una molteplicità di indagini che sarebbero inimmaginabili se lo scopo del progetto fosse stato solo quello di produrre una traduzione unicamente per essere stampata e distribuita su volumi cartacei. Anche se, in effetti, questa fu la richiesta iniziale avanzata da UCEI/CRI, parve evidente ed opportuno presentare una prospettiva di sviluppo scientifico sui dati raccolti nel corso degli anni, al fine di valorizzarne l'utilizzo mediante tecniche sempre più sofisticate di analisi linguistica e filologica computazionali. TRADUCO, pertanto, rappresenta solo una parte del progetto di ricerca, in quanto una dimensione orientata al formato digitale del vastissimo corpus pro-

[7] Si vedano, per esempio, le voci classificate come «Concetti» e quelle come «Maestri», nella sezione Glossari del trattato *Rosh haShanà* (Talmud Babilonese, 2016, p. 344 e p. 358).

ietta l'utilizzo del *Talmud* nel futuro e secondo direttrici diversificate. Vediamo, solo per fare alcuni esempi, quali prospettive sono attualmente allo studio e quali risultati esse si propongono di ottenere.

Ricordiamo che:

- TB ripropone frequentemente temi e situazioni in uno stesso trattato o in più trattati diversi;
- solo grande esperienza consente di ritrovare e leggere parti che si riferiscono a situazioni analoghe o simili e che meritino, se necessario, di essere studiate e valutate comparativamente;
- una concordanza per parole o per lemmi, che sarà comunque realizzata nella versione digitale, potrebbe non essere sufficiente per raccogliere tutte le porzioni dei testi che si trovino in questa situazione, ovvero che veicolino il medesimo significato o tràttino lo stesso tema;
- in molti casi i contesti potrebbero differire solo a causa delle diverse espressioni linguistiche adottate, anche se il valore semantico/concettuale da esse espresso è, appunto, il medesimo (cf. per esempio, il caso dei termini sinonimici, citata a proposito dell'algoritmo di ED)[8].

Ne consegue che si debba orientare la ricerca verso la rappresentazione concettuale delle molteplici forme di conoscenza che sono sottese al testo del *Talmud* e, per tale ragione, seguendo un orientamento affermatosi nel settore del *Semantic Web*, avviare la progettazione di adeguate strutture ontologiche[9], servendosi di gruppi di specialisti che pos-

---

[8] Si veda *supra*, paragrafo *Traduzione assistita*.

[9] A differenza del significato che la parola ontologia possiede nella terminologia filosofica a partire dalla *Metafisica* aristotelica, in informatica essa denota la concettualizzazione esaustiva ed esplicita di un qualsiasi dominio della conoscenza (per esempio, per il medioevo, l'anatomia umana, la botanica, la mineralogia, ecc.). Si tratta generalmente di una struttura dati gerarchica che contiene tutte le entità rilevanti (in forma di una tassonomia semantico-concettuale), le relazioni esistenti fra di esse, le regole, gli assiomi ed i vincoli specifici del dominio. Gli aspetti di formalizzazione derivano in buona parte dal pensiero logico matematico di Gottlob Frege, il quale elabora la propria ontologia in: *Funzione e concetto* (*Funktion und Begriff*, 1891), *Concetto e oggetto* (*Über Begriff und Gegenstand*, 1892) e *Senso e significato* (*Über Sinn und Bedeutung*,1892) consultati in: G. Frege, *Senso, funzione e concetto. Scritti filosofici*,

siedono idonee competenze. Tali strutture sono da organizzare in forma di domini di conoscenza differenziati grazie alla competenza di esperti, i quali possono essere gli stessi che abbiano già operato per l'inserimento delle annotazioni specialistiche (giurisprudenza, amministrazione, medicina, farmacopea, morale, ecc.). Questo lavoro renderà possibile la raccolta e la valutazione sinottica di tutte quelle parti del testo che presentino tratti semantici comuni (ovvero, trattino argomenti identici o molto simili), indipendentemente dalla terminologia adoperata nei sin-

---

a cura di C. Penco, E. Picardi, Laterza, Roma–Bari, 2001. Ci si è serviti anche di C. Penco, *Frege*, Carocci editore, Roma, 2010. Frege distingue rigorosamente i ***concetti*** dagli ***oggetti***. Ogni concetto è una funzione caratteristica che divide l'insieme degli oggetti in due: quelli che appartengono al concetto (*sussumono lo stesso concetto*) e quelli che non vi appartengono. Rifacendosi ad una metafora della chimica, i concetti sono qualche cosa di ***insaturo***, e vengono saturati dagli ***oggetti*** (espressi sotto forma di nomi). Similmente in informatica si dice che i concetti, nella struttura ontologica di un dominio di conoscenza, vengono popolati da nomi (*gli oggetti di Frege*) che ne costituiscono le istanze (*ovvero, li saturano*). La bibliografia relativa al tema delle ontologie nella scienza dell'informazione e nell'intelligenza artificiale è sterminata. Qui si indicano solo: un lavoro che, pur riferendosi ad uno specifico modello ontologico (DOLCE – *Descriptive Ontology for Linguistic and Cognitive Engineering*) fornisce anche indicazioni introduttive sul tema della rappresentazione della conoscenza (Gaio, Borgo, Masolo, Oltramari, Guarino, 2010), e un articolo che tratta della rappresentazione della conoscenza giuridica (Jori, Sartor, 2016). Qualcuno potrebbe a buon diritto obiettare che quanto è stato detto a proposito della concettualizzazione di alcuni termini significativi del *Talmud* non avrebbe direttamente a che fare con l'organizzazione ontologica dei dati afferenti ad uno o più domini della conoscenza (il dominio dei Maestri/rabbini; quello del mondo vegetale; ecc.), ma più semplicemente costituirebbe una modalità di organizzazione di parole in forma tassonomica al fine di costruire indici e glossari: si tratterebbe, insomma, di una specie di nomenclatura. In effetti, almeno in questa fase del progetto, non si sono volutamente evidenziate e rese esplicite le relazioni fra i termini considerati. La modalità fino ad ora seguita, tuttavia, oltre a consentire il raggruppamento di parole che richiamano un medesimo concetto, si configura come un primo passo per definire in seguito, con maggiori dettagli, i rapporti che i concetti individuati si trovano ad avere nelle singole tematiche, discusse ripetutamente e con accenti diversi, nei vari trattati del *Talmud*. Inoltre, una successiva e reale struttura ontologica, se ben definita, si prevede che sia in grado di effettuare operazioni di *reasoning* (inferenze) in base alla conoscenza acquisita e proporre quindi, in maniera del tutto autonoma, connessioni di parti del testo che siano coerenti dal punto di vista logico e semantico.

goli trattati, ovvero indipendentemente dalle parole che siano state usate dai traduttori per rendere concetti che, nell'originale aramaico, figurano come identici o simili. In altre parole, la competenza di alcuni specialisti consente di strutturare una griglia di concetti, sotto-concetti e relazioni esplicite fra concetti che di fatto rappresenta il mondo, la conoscenza spalmata tra i trattati del *Talmud*, ma di reperimento estremamente difficile per chi non abbia un'esperienza talmudica eccezionale e una memoria prodigiosa. Per facilitare la comprensione di questa prospettiva di ricerca, simuliamo qui l'ipotesi, semplice ed elementare, di voler visualizzare tutti i contesti che contengano i seguenti elementi:

concetto «rabbino» + concetto «decima» + concetto «albero da frutto»

Come si ricorderà, alcuni termini ed espressioni linguistiche sono già stati classificati dai traduttori/revisori, nelle annotazioni strutturate, mediante i concetti elencati in una lista che è disponibile in un menu a tendina. Dobbiamo tuttavia considerare che tecniche specifiche di linguistica computazionale (come, per esempio, la *Named Entity Recognition* o la semantica distribuzionale) possono effettuare tali classificazioni in maniera molto più vasta, realizzando in tal modo la correlazione fra termini rilevanti presenti nei testi e valore semantico-concettuale da essi veicolato[10].

[10] Per l'estrazione automatica dei termini italiani è stato utilizzato lo strumento T2K, sviluppato presso l'ILC (Dell'Orletta, Venturi, Cimino, Montemagni, 2014). T2K prevede che il testo sia analizzato linguisticamente (attribuendo ad ogni parola la sua parte del discorso) e opera in due fasi: i) estrae candidati termini (mono e polirematici) corrispondenti a *pattern* morfo-sintattici predeterminati (es. nome, nome-preposizione-nome, nome-aggettivo) e, successivamente, ii) applica una misura statistica ai termini polirematici per determinare il grado di associazione delle singole parole che li compongono. Sono stati estratti 4166 termini dai quattro trattati già tradotti (*Berakhòt*, *Rosh haShanà*, *Ta'anìt* e *Qiddushìn*), 210 dei quali sono stati esclusi perché derivanti da errori di attribuzione delle parti del discorso o perché considerati non rilevanti. I restanti 3956 termini sono stati quindi ordinati sulla base della loro rilevanza all'interno dei singoli trattati e in relazione a specifiche porzioni di trattato. La rilevanza è stata calcolata attraverso la TF-IDF (*Term Frequency – Inverse Document Frequency*), una funzione solitamente utilizzata nell'ambito del recupero d'informazione (*Information Retrieval*) per misurare l'importanza di un termine rispetto a un documento o a un

Per rimanere all'esempio indicato sopra, sulla base delle classificazioni ad oggi effettuate a mano dai traduttori, il sistema proporrebbe i seguenti contesti:

1. Rosh haShanà, 1,15a: "*Rabbà* disse: un *cedro* nato nel sesto anno, rimasto sull'*albero* fino all'inizio del settimo anno, è esente dal prelievo della *decima*, come tutti i prodotti del settimo anno …";
2. Rosh haShanà, 1,15a: "*Abbayè* gli disse: l'ultima frase è accettabile, … da un lato di rendere obbligatoria l'eliminazione del *frutto* in quanto appartenente al settimo anno, e dall'altro si esenta dalla *decima* ….";
3. Rosh haShanà, 1,15b: "*Rabbì Yochanàn* …: parlavo riguardo alla *decima* del *carrubo*, istituita dai *Maestri*, e tu mi contraddici riportando la regola del settimo anno, che deriva dalla Torà?"

Si evince che, anche grazie all'uso della grafica (i concetti, infatti, potrebbero essere identificati ciascuno con un colore diverso da quello degli altri), si ottiene un immediato colpo d'occhio sulle similarità tematiche presenti nei testi anche qualora uno stesso senso sia veicolato da parole o espressioni diverse. Nell'esperimento effettuato, le parole Rabbà, Abbayè, Rabbì, Yochanàn e Maestri sono evidenziate dal colore assegnato al concetto di «rabbino», mentre un colore differente marca Cedro, Albero, Carrubo e Frutto, assimilati concettualmente al senso di «albero da frutta/frutta»; infine, un colore ancora diverso connota il termine Decima.

Come detto, questa organizzazione per concetti è elementare, ma costituisce una base sperimentale per arrivare, a passi graduali e successivi, ad una maggiore granularità che consenta al sistema informatico di fornire risposte più dettagliate e precise a colui che consulterà il corpus digitale.

A tale proposito si può affermare che non esiste un criterio universalmente valido per stabilire il grado di analiticità e di dipendenza ge-

---

corpus (Bolasco, Pavone, 2008). Infine, il repertorio terminologico, una volta estratto secondo le modalità descritte, riceve una formalizzazione ed una descrizione semantica strutturata sul modello del *Dizionario esplicativo e combinatorio* (Mel'čuk, Clas, Polguère, 1995).

rarchica con cui strutturare i concetti e i sotto-concetti nella rappresentazione ontologica di un dominio[11]; essa, infatti, è strettamente dipendente sia dai dati che documentano il dominio stesso (nel nostro caso si tratta di dati diacronicamente molto estesi e di interpretazione talvolta molto difficile), sia dagli scopi che si intendono raggiungere utilizzando quei dati. Se sappiamo bene ormai che le strategie messe in atto per il *Semantic Web* tendono a fornire informazioni di tipo economico, commerciale e perfino di "profilazione" politica degli utenti della rete, non è invece possibile stabilire a priori le motivazioni di tipo culturale e scientifico che spingono varie comunità di specialisti a operare su vasti corpora digitali, in specie quando questi si riferiscono a testi antichi, caratterizzati da notevole difficoltà interpretativa. A tale fine è importante che siano gli stessi ricercatori umanisti, grazie alla propria competenza, a valutare in anticipo fino a quale livello di dettaglio intendano organizzare la struttura gerarchica dei termini del dominio e le loro reciproche relazioni, affinché i risultati che il sistema informatico sarà in grado di offrire alle *query* di una vasta comunità di utenti corrispondano in maniera coerente ed esaustiva alle loro richieste.

3. IL PROGETTO DITMAO

Un esempio concreto serve a chiarire questo aspetto del problema: ci si riferisce al progetto DiTMAO (*Dictionnaire des Termes Médico-botaniques de l'Ancien Occitan*)[12] che studia la medicina e la farmacopea

[11] Non è il caso qui di distinguere fra ontologia fondazionale, che è assimilabile ad un generico glossario di base, usualmente gerarchizzato in un numero molto ristretto di livelli entro i quali tutto il resto deve essere descritto, e ontologia di dominio la quale, al contrario, presenta una struttura molto più analitica e dettagliata, in grado di rappresentare la conoscenza di un intero dominio come, per esempio, la scienza giuridica, la anatomia umana, la filosofia medievale, ecc. (Per questi aspetti si veda ancora Gaio, Borgo, Masolo, Oltramari, Guarino, 2010).

[12] Il progetto (*An XML-based lnformation System for Old Occitan Medical Terminology*) oltre a prevedere la stampa del lessico su volumi cartacei, sta realizzando anche un sistema informativo basato su ontologie per la terminologia medico-botanica in occitano antico e, più in generale, per la medicina medievale. Si tratta di un'iniziativa scientifica congiunta tra la Georg-August-Universität Göttingen, il Dip. Filologia,

medievale occitaniche: – per pubblicare su volumi a stampa il lessico di riferimento; – per facilitare il raccordo fra il lessico appartenente a tale àmbito semantico e il corpus dei testi nei quali esso è documentato; – per fornire risposte puntuali e mirate alle interrogazioni, sul corpus lessicale disponibile in rete, effettuate da specialisti.

A differenza di quanto fino ad oggi realizzato per la traduzione del *Talmud*, in DiTMAO la struttura che organizza i concetti e le relazioni fra i concetti è molto più granulare (si veda un dettaglio nella figura 2, che riguarda il concetto di «testa»). Lo studio che ha portato a descrivere l'anatomia umana medievale secondo un criterio così analitico è giustificato dal fatto che queste informazioni sono variamente distribuite nei testi utilizzati dal progetto per la redazione del lessico (Corradini, 2014; Corradini, 2016). Esse sono accompagnate da altre informazioni relative alle malattie e alle cure in forma di ricette, spesso ripetute e con varianti più o meno significative. É stata pertanto una scelta obbligata quella di

---

Letteratura, Linguistica dell'Università di Pisa, la Universität zu Köln e l'ILC-CNR di Pisa, sulla base di un finanziamento erogato dalla DFG (*Deutsche Forschungsgemeinschaft*). Gli strumenti lessicografici a disposizione dei linguisti e filologi romanzi, in particolare occitanisti, sono ormai datati e in essi la registrazione delle voci è principalmente condotta sullo spoglio di opere poetiche trobadoriche, ad eccezione del DAO (*Dictionnaire onomasiologique de l'Ancien Occitan*. Cf. Baldinger, 1975–1996 e Baldinger suppl., 1980–2000) e del DAG (*Dictionnaire onomasiologique de l'Ancien Gascon*. Cf. Baldinger, 1975–1998), redatti su una base lessicale più vasta e organizzati secondo una struttura onomasiologica. A partire dagli anni '90 del secolo scorso, soprattutto a cura di Maria Sofia Corradini dell'Università di Pisa, sono state realizzate edizioni critiche di fonti inedite, riguardanti vari testi della *fachliteratur* occitanica, molte delle quali documentano ricettari medico–farmaceutici e trattati di anatomia animale ed umana (Corradini, 1997). La provenienza di queste opere si colloca nei centri della Francia meridionale, spesso in zone di contatto fra Provenza e Catalogna medievali, e tutte presentano sia elementi di medicina popolare, sia influssi consistenti delle scuole mediche di Salerno e Montpellier, con stretti rapporti con la ricca tradizione scientifica greca, latina ed araba. Degno di nota è il fatto che alcune opere presentano liste sinonimiche nelle quali il termine medico o botanico occitanico è reso con grafia ebraica, fenomeno di estremo interesse per mettere in luce la circolazione linguistica e culturale in area mediterranea, nonché il lavoro di traduzione eseguito per rendere disponibili i rimedi atti a conservare o restituire la salute (Bos, Mensching, Hussein, Savelsberg, 2011 e Bos, Mensching, Zwink, 2017).

organizzare la conoscenza di un dominio vasto e frammentato in una forma organica, che definiamo struttura ontologica, in modo da avere, da un lato, la descrizione della medicina e farmacopea occitaniche come si evince dalle fonti utilizzate; dall'altro, di offrire la possibilità di "popolare" i singoli elementi dell'ontologia (i concetti) con le parole o le espressioni presenti nel corpus, le quali ereditano tutte le proprietà (correlazioni e dipendenze) possedute, nella struttura, dai concetti a cui sono associate; infine, di offrire la possibilità di interrogare il corpus adoperando, come chiavi di ricerca, i concetti e non soltanto, come di solito avviene, le forme linguistiche o, nella migliore delle ipotesi, i lemmi.

La conseguenza di queste scelte è rilevante: chi intenda visualizzare, per esempio, tutte le ricette relative alle cure per danni o malattie alla testa, otterrà i contesti desiderati perché il sistema sfrutta le relazioni presenti nella struttura ontologica che legano il concetto «testa» con tutti gli altri concetti che ne ereditano le proprietà e le relazioni (per esempio: «cranio», «faccia», «fronte», «parte laterale», «mento», ecc.; e, a un livello sempre più analitico, troviamo «gota» e «tempia» che si pongono, ad un livello più analitico, rispetto a «parte laterale»). L'utilizzatore potrà leggere anche tutti i passi nei quali, come capita di frequente nei trattati medievali afferenti a questo àmbito, le parole utilizzate nel corpus, ed associate al medesimo concetto, siano diverse in quanto sinonimi (es.: antico occitano *ventre* e *estomach*), o appartenenti a lingue diverse (latino: *abrotonum*; occitano: *alambroze, brona, bretonia*), o scritte in alfabeti diversi (caratteri latini per l'antico occitano *malva*; caratteri ebraici: אבלמ = traslitterato MLB').

La figura 2 mostra un piccolo dettaglio dell'ampio grafico in corso di realizzazione con lo scopo di organizzare la conoscenza anatomica medievale occitanica in tutte le sue componenti. Tale dominio ontologico è stato originariamente espresso da medievalisti e linguisti romanzi, partecipanti al progetto DiTMAO, in linguaggio UML (*Unified Modeling Language*) (Bozzi, Luzzi, 2016), ed è, oggi, in fase di riscrittura nel più diffuso *Protegé*, editore di ontologie basato sullo standard OWL (*Ontology Web Language*) riconosciuto a livello internazionale.

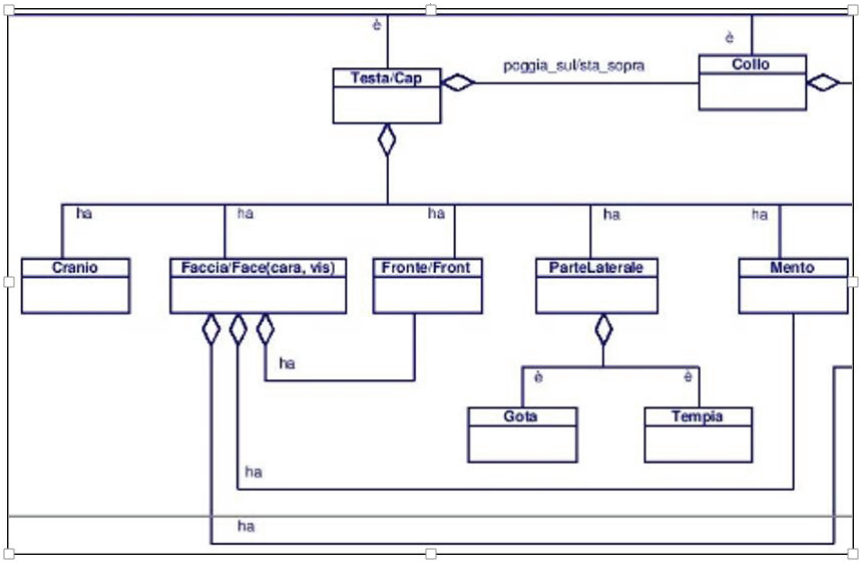

Fig. 2.

Una rappresentazione della conoscenza così dettagliata e articolata risulterebbe del tutto inappropriata per il *Talmud*, corpus eterogeneo e pluri-tematico per eccellenza; si è deciso, pertanto, di provare, mediante le già citate tecniche di *information extraction* (si veda *supra* la nota 10), a raggruppare solo le parole denotanti un ristretto nucleo di concetti e, sulla base di questi dati preliminari, costruire di volta in volta gruppi sempre più vasti corredati dell'indicazione esplicita del tipo di relazione che raccorda i termini fra di loro. Esperimenti in tal senso e con risultati promettenti sono in corso, per esempio, per tentare di definire un'ontologia dei «Rabbini/Maestri». L'idea di base è stata quella di associare al nome (o ai nomi, nel caso ne abbia più di uno) di un Maestro l'elenco dei termini che co-occorrono all'interno di una porzione testuale di lunghezza predefinita a priori, e si assume che vi siano accettabili probabilità che essi, o solo una parte, abbiano a che fare con gli argomenti che quel Maestro tratta. Naturalmente questo non sarà sempre vero: per tale motivo, fintanto che le informazioni estratte non saranno state opportunamente validate, le associazioni inserite nell'ontologia saranno formalmente definite tra Maestri e termini co-occorrenti e non

tra Maestri e i relativi argomenti di discussione, fase, quest'ultima, che necessita di successive azioni e appropriate tecnologie[13].

4. CONCLUSIONE

Il progetto PTTB, secondo quanto sopra schematicamente descritto, ha dunque consentito di rispondere a due esigenze complementari. Da un lato viene garantita assistenza ad una numerosa comunità di esperti che cooperano *online* alla traduzione italiana di un corpus vasto e complesso sia per lingua che per contenuto. Dall'altro, PTTB sta offrendo un'occasione privilegiata per progettare, realizzare e validare tecnologie linguistico-computazionali innovative, capaci di soddisfare molte esigenze di studio da parte di talmudisti esperti, e molte curiosità di argomento storico e culturale richieste da lettori non specialisti. Il *Talmud* digitale, mediante le forme di accesso attuali e prospettate per un prossimo futuro, viene ad assumere le fattezze di un Maestro virtuale che, in linea con una tradizione secolare, sarà strumento importante di formazione culturale e religiosa.

Si è voluto qui, inoltre, sottolineare l'aspetto relativo all'importanza della riusabilità ed adattabilità di almeno un componente tecnologico concepito per PTTB. Il modello di rappresentazione ontologica del lessico, in fase di adattamento alle caratteristiche dei dati presenti nel DiTMAO, consentirà una navigazione semantica e concettuale più adeguata alle esigenze di consultazione delle molte parti del vasto corpus di

---

[13] Alcuni termini che, per esempio, sono stati estratti dal sistema in relazione al rabbino *Yochanàn*, sono (tra parentesi compare un indice di rilevanza TF–IDF – si veda ancora *supra* la nota 10 – che è generato dall'algoritmo statistico): appezzamento (71.386): Qiddushìn 062a, Qiddushìn 062b; cherubino (53.509): Rosh haShanà 031a–031b; impasto (44.256): Qiddushìn 080a, Qiddushìn 080b, Qiddushìn 080a–080b; pensiero (43.380): Qiddushìn 059a–059b; frutti del ventre (36.122): Berakhòt 051b; redenzione (33.639): Berakhòt 004b; azione (33.203): Qiddushìn 059b, Qiddushìn 059a–059b; anno (31.793): Rosh haShanà 010a, Rosh haShanà 010b, Rosh haShanà 015b, ecc.; frutti (30.710): Berakhòt 051b, Berakhòt 044a, Rosh haShanà 010a, Ta'anìt 007a, ecc. Per avere informazioni su questi aspetti del problema e sugli esperimenti già condotti sul testo di TB, si veda: Giovannetti, Bellandi, Del Grosso, Marchi, Pecchioli, Piccini (in corso di stampa).

ricette ove ricorrono situazioni, malattie o rimedi simili o affini, ma che, in mancanza di questo nuovo strumento tecnologico, sarebbe impossibile raccogliere e valutare comparativamente.

BIBLIOGRAFIA

Albanesi, A., Bellandi A., Benotto, G., & Giovannetti, E. (2015). *Translation, Annotation and Knowledge Modelling of the Babylonian Talmud: the Traduco System*. Paper presented at the conference Digital Humanities 2015, (Sydney, 29/06–03/07/2015).

Baldinger, K. (1975–1996). *Dictionnaire onomasiologique de l'ancien occitan (DAO)*, fasc. 1–7. Tübingen: Niemeyer.

Baldinger, K. (1980–2000). *Dictionnaire onomasiologique de l'ancien occitan (DAO)*, fasc. 1–7, suppl. Tübingen: Niemeyer.

Baldinger, K. (1975–1998). *Dictionnaire onomasiologique de l'ancien gascon (DAG)*, fasc. 1–9. Tübingen: Niemeyer.

Bellandi, A. (2015). 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 A. Bozzi (Ed.). *Digital texts, translations, lexicons in a multi-modular Web application: methods and samples* (pp. 69–84). Firenze: Leo S. Olschki.

Bolasco, S., & Pavone, P. (2008). Multi-class categorization based on cluster analysis and TFIDF. In S. Heiden, & B. Pincemin (Eds.), *JADT 2008 : 9es Journées internationales d'Analyse statistique des Données Textuelles*. Lyon: Presses universitaires de Lyon. Retrieved from *http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.659.3472&rep=rep1&type=pdf*.

Bos, G., Mensching, G., Hussein, M., & Savelsberg, F. (2011). *Shem Tov Ben Isaac of Tortosa, Sefer ha-Shimmush, Book 29. Medical synonym lists from Medieval Provence: Part 1: Edition and Commentary of List 1 (Hebrew—Arabic— Romance/Latin).* Leiden: Brill.

Bos, G., Mensching, G., & Zwink, J. (2017). *Medical Glossaries in the Hebrew Tradition: Shem Tov Ben Isaac, Sefer Almansur*. Leiden: Brill.

Bozzi, A., & Luzzi, D. (2016). Un'ontologia per il DiTMAO (Dictionnaire des Termes Médico-botaniques de l'Ancien Occitan). In D. Trotter, A. Bozzi, & C. Fairon (Eds.). *Actes du XXVII Congrès international de linguistique et de philologie romanes (Nancy, 15–20 juillet 2013).*

Section 16: Projets en cours; ressources et outils nouveaux (pp. 55–63). Nancy: Atilf., ATILF. Retrieved from *www.atilf.fr/cilpr2013/actes/section-16/CILPR-2013-16-Bozzi-Luzzi.pdf*.

Bozzi, A. (2017). TRADUCO. Linguistica e filologia computazionali nella traduzione del *Talmud*. In R.S. Di Segni (Ed.). *Talmud Babilonese – Trattato Berakhòt* (pp. xxvii–xxx). Firenze: Giuntina.

Choueka, Y. (1980). Computerized full-text retrieval systems and research in the humanities: the Responsa project. *Computers and the Humanities*, *14*, 153–169.

Corradini, M.S. (2014). Lessico e tassonomia nell'organizzazione del "Dictionnaire des Termes Médico-botaniques de l'Ancien Occitan (DiTMAO). *Revue de Linguistique Romane*, *78*, 87–132.

Corradini, M.S. (1997). Ricettari medico-farmaceutici medievali nella Francia meridionale. Firenze: Olschki.

Corradini, M.S. (2016). La realizzazione del Dictionnaire des Termes Médico-botaniques de l'Ancien Occitan (DiTMAO): problemi di organizzazione della conoscenza medico-farmaceutica attestata nei manoscritti in occitano antico. In D. Trotter, A. Bozzi, & C. Fairon (Eds.). *Actes du XXVII Congrès international de linguistique et de philologie romanes (Nancy, 15–20 juillet 2013). Section 16: Projets en cours; ressources et outils nouveaux*. Nancy: Atilf. Retrieved from *http://www.atilf.fr/cilpr2013/actes/section-16.html*.

Freedman, H. (2014). *The Talmud. A Biography. Banned, Censored and Burned. The Book They Couldn't Suppress*. London: Bloomsbury (it.: Torino: Bollati Boringhieri, 2016).

Gaio, S., Borgo, S., Masolo, C., Oltramari, A., & Guarino, A. (2010). Un'introduzione all'ontologia DOLCE. *AIDA Informazioni*, *1–2*, 107–125. Retrieved from *https://www.academia.edu/11752175*.

Giovannetti, E., Albanesi, A., Bellandi, A., & Benotto, G. (2017). Traduco: a collaborative web-based CAT environment for the interpretation and translation of texts. *Digital Scholarship in the Humanities*, *32*, Issue suppl. 1, April 2017, 47–62. Retrieved from *https://doi.org/10.1093/llc/fqw054*.

Giovannetti, E., Bellandi, A., Del Grosso, A., Marchi, S., Pecchioli, A., & Piccini, S. (to appear). La terminologia del Talmud Babilonese: estrazione, rappresentazione e uso nel contesto della Linguistica Computazionale. *Materia Giudaica* (Atti del XXXIII Congresso Internazionale dell'AISG-Associazione Italiana per lo Studio del Giudaismo, 2–4 settembre 2019, Ravenna).

Jori, M., & Sartor, G. (2016). *Informatica giuridica*. Torino: G. Giappichelli.

Mel′čuk, I.A., Clas, A., & Polguère, A. (1995). *Introduction à la lexicologie explicative et combinatoire*. Bruxelles: Duculot.

Dell'Orletta, F., Venturi, G., Cimino, A., & Montemagni, S. (2014). T2K: A System to Automatically Extracting and Organizing Knowledge from Texts. In N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, & S. Piperidis (Eds.), *Proceedings of 9th Edition of International Conference on Language Resources and Evaluation (LREC 2014).* (pp. 2062–2070). Reykjavik: European Language Resources Association.

Talmud Babilonese (2016). *Talmud Babilonese. Trattato Rosh haShanà (Capodanno)*, ed. by R. Di Segni. Firenze: Giuntina.

Zitouni, I. (Ed.). (2014). *Natural Language Processing of Semitic Languages*. Heidelberg: Springer.

Riassunto: La linguistica computazionale affronta da molti anni e da parte di molti enti pubblici e privati i problemi posti dalla traduzione automatica che ha importanti ricadute applicative ed economiche. Molto più rari sono invece i casi in cui comunità scientifiche e/o soggetti industriali a livello internazionale investano risorse per rendere più semplice e veloce il lavoro di chi affronta opere antiche, spesso caratterizzate da grandi difficoltà interpretative. Il sistema TRADUCO, progettato e realizzato presso l'Istituto di Linguistica Computazionale del CNR di Pisa, consente a un gruppo di talmudisti di rendere in italiano corrente i trattati del *Talmud* babilonese, redatti in aramaico ed ebraico biblico. La lingua e la struttura del testo hanno reso improponibile la progettazione di algoritmi di *Machine Translation*, mentre ottimi risultati si sono ottenuti grazie a tecniche di *Translation Memory* e di *Edit Distance*. Queste, ben armonizzate fra loro, consentono al sistema di proporre agli specialisti una sempre più alta percentuale di traduzioni corrette, inserite in un ambiente di lavoro intuitivo. Il risultato è esportabile in file xml predisposti per le operazioni finali di stampa. Ciò ha consentito di pubblicare già 5 trattati in volumi cartacei che offrono testo tradotto, annotazioni, indici tematici, e altre informazioni. Molti volumi sono già stati tradotti e attualmente in fase di controllo editoriale. Varie prospettive si aprono, infine, per la fruizione del *Talmud* digitale in italiano. Fra esse, una fra le più interessanti riguarda la possibilità di associare, anche mediante tecniche di *Machine Learning* e *Named Entity Recognition*, valori semantici o concettuali (*Talmud Ontological Framework*) a porzioni di testo che riferiscono o discutono tematiche simili. Ciò consentirà di navigare su base semantica un archivio testuale tanto vasto ed eterogeneo. La strategia adottata risulta modulabile anche per altri progetti di carattere lessicografico come, per esempio, il DiTMAO (*Dictionnaire des Termes Médico-botaniques de l'Ancien Occitan*). Esso offrirà percorsi di navigazione semanticamente orientati nell'àmbito di un vasto corpus di testi digitalizzati di argomento medico-farmaceutico e botanico in occitanico medievale. Per tali ragioni TRADUCO e DiTMAO si configurano come istanze di un'infrastruttura tecnologica di linguistica e filologia computazionali fra le più innovative nel settore delle *Digital Humanities.*

Parole chiave: talmud, linguistica computazionale, filologia digitale, TM (Translation Memory). occitano medievale, ontologie