Alessandra Pecchioli, Davide Albanesi, Andrea Bellandi,
Emiliano Giovannetti, Simone Marchi

ANNOTAZIONE LINGUISTICA AUTOMATICA DELL'EBRAICO MIŠNAICO: ESPERIMENTI SUL TALMUD BABILONESE

1. *Introduzione* di Emiliano Giovannetti

Il Progetto Traduzione Talmud Babilonese (PTTB)[1] ha come obiettivo la traduzione in lingua italiana del Talmud Babilonese. Il progetto, nato nel 2010, coinvolge un team di traduzione formato da circa 70 studiosi, fra traduttori esperti, traduttori in formazione, istruttori, revisori di contenuto e revisori editoriali, affiancati da un team di informatici, linguisti computazionali e da uno staff amministrativo. Nel 2016 è uscito in forma cartacea il primo volume tradotto: il trattato Roš Hašanah. L'anno successivo è stato pubblicato il trattato Berakot in doppio volume. La traduzione commentata, con testo originale a fronte, è realizzata con strumenti avanzati di trattamento automatico del testo e della lingua integrati in un'applicazione, Traduco, creata *ad hoc* dall'Istituto di Linguistica Computazionale "Antonio Zampolli" del CNR. Se è vero che esistono già da tempo strumenti che supportano la traduzione di testi è altrettanto vero che nessuno di questi è stato progettato per la traduzione di opere con una storia linguistica e redazionale complessa come quella del Talmud.

Traduco, dal punto di vista tecnico, è uno strumento per la traduzione assistita da calcolatore (in inglese, *Computer-Assisted Translation*). Si tratta di un'applicazione web (che non richiede di installare software sul proprio computer ma soltanto di accedere al sistema tramite un browser) concepita per la traduzione collaborativa, in questo caso del Talmud babilonese in italiano. Dal 2012 a oggi Traduco si è evoluto e adeguato alle necessità imposte dall'opera in traduzione, attraverso un'architettura flessibile e adattabile ai contesti. L'impianto tecnologico che lo sostiene, infatti, può essere (in modo relativamente semplice) riutilizzato nel trattamento di altri testi, che possono essere tradotti, annotati, stampati e, più in generale, divenire oggetto di studio e ricerca.

Traduco, oltre a fornire gli strumenti di base per il supporto alla traduzione, costituisce l'infrastruttura tecnologica per l'intero progetto, esponendo funzioni per la gestione degli utenti, la supervisione del lavoro, la creazione di note, annotazioni, glossari e l'impaginazione del testo. Inoltre, nel sistema sono stati integrati algoritmi di linguistica computazionale per il trattamento del testo e della lingua, finalizzati a rendere più rapido e puntuale il lavoro di traduzione, revisione e curatela.

Oltre al lavoro descritto nel presente articolo, incentrato al trattamento automatico della lingua ebraica, sono in corso altre ricerche, tra le quali si cita la strutturazione delle voci di glossario attualmente presenti nel sistema in una risorsa digitale terminologica e ontologica. Attraverso tale risorsa, che integrerà un repertorio terminologico multilingue con una strutturazione concettuale del dominio talmudico (attraverso delle ontologie), sarà possibile condurre studi e ricerche su base semantica, interrogando il sistema in modo da visualizzare, ad esempio, tutti i passi che citano un particolare rabbino di una specifica generazione oppure quelli in cui si discute un tema particolare (rituale, legale, medico, botanico, ecc.).

Per quanto riguarda il testo oggetto di studio possiamo dire, semplificando, che il Talmud babilonese è composto essenzialmente da due lingue che, a loro volta, corrispondono a due testi distinti: Mišnah e Gemarah. Il primo dei testi è quello più antico, messo per iscritto tra il primo e il terzo secolo dell'era volgare ad opera dei *tanna'im*, in una lingua successiva a

[1] https://www.talmud.it/.

quella biblica: l'ebraico rabbinico antico.[2] Per questa lingua è attestata anche una fase più tarda (ebraico rabbinico tardo), idioma usato dagli *'amora'im*, nei midrašim aggadici, nel Talmud babilonese e in quello di Gerusalemme, fino al vi sec. dell'era volgare.

Per condurre un esperimento valido era necessario addestrare il calcolatore su una lingua quanto più omogenea e coerente possibile, ragione per cui si è optato per l'analisi dell'ebraico rabbinico antico, individuabile, all'interno del Talmud babilonese, soprattutto nella Mišnah (ma presente anche in altre opere, quali la Tosefta, i midrašim halakici e il Seder 'olam Rabbah).

Come sarà specificato nella sezione successiva lo scopo principale dell'annotazione della lingua nell'ambito di un progetto di traduzione assistita è quello di migliorare i suggerimenti forniti dal sistema attraverso la cosiddetta "memoria di traduzione". Inoltre, su un testo linguisticamente annotato è possibile effettuare ricerche su base linguistica, utili sia per lo studioso (in questo caso di Talmud), sia, in fase di traduzione, per il revisore e il curatore, che hanno la possibilità di verificare con maggiore precisione e velocità, quanti e quali interventi sono necessari affinché il testo tradotto raggiunga la sua forma finale.

Il presente articolo è strutturato come di seguito. Nella sezione 2 viene introdotto il problema dell'annotazione linguistica, evidenziandone finalità e criticità. La sezione 3 descrive la natura dell'annotazione linguistica automatica, con particolare riguardo allo stato dell'arte sul trattamento automatico dell'ebraico. La costruzione di un *corpus* mišnaico annotato linguisticamente è descritta nella sezione 4. Il processo di addestramento e valutazione degli annotatori automatici adottati per la sperimentazione è descritto nella sezione 5. Infine, nella sezione 6 sono delineati i passi successivi della ricerca.

2. *L'annotazione linguistica* di Alessandra Pecchioli

Un testo linguisticamente annotato è un testo nel quale vengono esplicitate le caratteristiche linguistiche di ciascuna parola di cui è composto. I livelli di descrizione della lingua possono essere molteplici: morfologico, sintattico, semantico, pragmatico. Solitamente, ma soprattutto a fini computazionali (si veda più avanti), partendo dal livello morfologico, l'analisi è condotta rispettando uno specifico ordine, in modo tale che i risultati ottenuti a ogni livello di esame possano essere utilizzati per quello successivo. Prima ancora di procedere all'analisi di un testo, esso va suddiviso in frasi, quindi, in unità ortografiche di base, dette *token*. In seguito si provvede ad assegnare ogni *token* (e relative porzioni, o *sub-token*, se la lingua lo richiede) a un lemma e alla sua categoria grammaticale (in inglese, *Part-Of-Speech*, abbreviato in POS).

Quest'ultimo punto, particolarmente delicato, costituisce uno dei maggiori problemi dell'analisi morfologica. Non sempre, infatti, è così agevole decidere quale POS assegnare. Essa, infatti, deriva da ciò che il *token* è per definizione e, allo stesso tempo, dalla relazione che esso intrattiene con il contesto in cui si trova, vale a dire dal legame che instaura con le parole adiacenti e a esso collegate (un aggettivo per esempio, in un determinato contesto, può avere funzione di nome).

Come specificato in seguito, attraverso un testo così annotato si può rendere un calcolatore capace di analizzare, in modo automatico, testi nuovi, a patto che utilizzino lo stesso tipo di lingua sulla quale è stato istruito.

2.*1 Finalità dell'annotazione*

Nel contesto della traduzione del Talmud in italiano lo scopo dell'annotazione linguistica è

[2] L'ebraico rabbinico rappresenta una fase linguistica abbastanza vicina al tardo ebraico biblico. Tra le differenze più importanti si citano: 1. la fusione della finale *mem* con *nun*, ragione per cui i maschili plurali finiscono in *nun*; 2. la particella relativa še- invece di *'ašer*; 3. la particella šel che sostituisce quasi del tutto lo stato costrutto; 4. l'uso molto frequente di *hayah* seguito da participio; 5. la scomparsa del *waw* inversivo/consecutivo; 6. la perdita dell'inf. assoluto, del coortativo e dello iussivo. Per approfondire la storia e lo studio della lingua mishnaica suggeriamo di leggere il paragrafo corrispondente sotto la voce "Hebrew Language" in *Encyclopaedia Judaica*, vol. 8, Keter Publishing House, Detroit 2007², pp. 639-650 e relativa bibliografia.

duplice. L'obiettivo principale è quello di dotare il sistema Traduco di un suggeritore automatico per la traduzione più efficiente, integrando nel testo l'annotazione linguistica e facendo in modo che il sistema possa sfruttarla per fornire suggerimenti più puntali su base lemmatica. Per fare un esempio, si considerino le seguenti frasi tratte dal Talmud: *Berakot*, 42b, *brk 'l hprprt* e *Berakot*, 44a, *mbrk 'l hmlyḥ*. Nell'attuale versione del sistema Traduco (priva di annotazione linguistica), la prima parte di tali enunciati differisce, ortograficamente, nelle due parole *brk* e *mbrk*; perciò la "misura di distanza" tra queste proposizioni, calcolata dal sistema per fornire i suggerimenti di traduzione (sulla base delle stringhe già tradotte conservate nella memoria di traduzione) sarà valutata a partire da questa differenza.[3] Invece, una volta lemmatizzate tutte le parole, i due termini in questione, condividendo il lemma, saranno considerati linguisticamente affini e la misura di distanza sarà ridotta di conseguenza, consentendo al sistema di fornire un più ampio e accurato ventaglio di suggerimenti.

Il secondo fine dell'annotazione linguistica è quello di permettere a un qualsiasi utente di interrogare il testo non solo per "parola chiave", ma anche su base linguistica, ad esempio cercando tutte le forme flesse di un particolare lemma o tutte le parole che derivano da una certa radice e che posseggono specifici tratti morfologici.

Un vantaggio importante apportato dall'uso dell'annotazione linguistica su un testo è quello di acquisire, in modo pressoché istantaneo, una visione d'insieme delle caratteristiche linguistiche e stilistiche di opere molto ampie, in questo caso del Talmud. A colpo d'occhio è infatti possibile avere contezza, per esempio, della distribuzione delle categorie lessicali (percentuale di verbi, nomi, aggettivi, ecc.) ed è pure immediatamente individuabile un particolare modello stilistico che si riproduce in più brani anche molto distanti tra loro. Questo tipo di lettura "a distanza"[4] si contrappone a una modalità di analisi più "ravvicinata", incentrata su un'indagine dei fenomeni linguistici più locale (ad esempio, la ricerca e lo studio della distribuzione di un dato lemma o di una struttura in un dato trattato). Ovviamente la lettura "a distanza", di cui una caratteristica principale è il brevissimo lasso di tempo in cui si svolge, è impossibile senza l'ausilio di un calcolatore.

Nel contesto del processo di annotazione linguistica qui descritto sono state introdotte due modalità distinte per l'annotazione delle parole: per "categoria" e per "funzione". Nella prima il lessema è esaminato per ciò che è per definizione, nella seconda, invece, lo stesso lessema è analizzato sulla base della relazione che esso sviluppa col contesto in cui si trova.

2.2 *Criticità e scelte*

Possiamo, ora, a esaminare quali sono i principali problemi che si devono affrontare nell'annotazione linguistica di una lingua semitica. Pur avendo trattato, in questo lavoro, l'analisi dell'ebraico, molte delle criticità da tenere in considerazione sono comuni al trattamento di altre lingue della stessa famiglia. Il primo problema riguarda l'accesso alle risorse linguistiche e agli strumenti di analisi esistenti che, nel caso dell'ebraico, sono disponibili quasi esclusivamente per l'ebraico moderno (per un'analisi dell'esistente si veda la sezione 3.1).

Una delle maggiori sfide che l'analisi morfologica delle lingue semitiche pone è quella della disambiguazione ortografica delle parole. Com'è noto, dal momento che la scrittura è quasi esclusivamente consonantica, ogni parola può avere molteplici letture. Di conseguenza, tornando al problema dell'annotazione automatica, si tratta di far capire al calcolatore quale sia la lettura giusta da scegliere. Il problema dell'ambiguità ortografica, cruciale in tutti gli studi su grandi *corpora* (tipicamente in ebraico e arabo moderno), non si rivela tuttavia così arduo laddove il testo in esame sia vocalizzato. L'edizione del testo in lingua originale del Talmud utilizzata nel progetto è effettivamente vocalizzata e il testo, di conseguenza, risulta ortograficamente poco ambiguo.

[3] Per una spiegazione dettagliata di come funziona il suggerimento automatico si faccia riferimento a: E. Giovannetti, D. Albanesi, A. Bellandi, G. Benotto, *Traduco: A collaborative web-based CAT environment for the interpretation and translation of texts*, «Digital Scholarship in the Humanities» 32 (suppl_1) (2017), pp. 47-62, Oxford Press University.

[4] F. Moretti, *Distant Reading*, Verso Books, London 2013.

Un'ulteriore criticità è rappresentata dalla definizione delle categorie grammaticali, raccolte in un insieme che, tecnicamente, è definito *tagset*. La maggior parte degli studi computazionali sull'analisi della lingua sono stati condotti su lingue indo-europee (in particolar modo sulla lingua inglese). Di conseguenza, può risultare difficile riutilizzare (anche in parte) i *tagset* creati per tali lingue. Le parti fondamentali della frase (verbo, nome e aggettivo) compongono le nozioni linguistiche di base. Nonostante ciò, in una data lingua emergono molte altre tipologie di unità lessicali che non è sempre facile racchiudere in una categoria ben precisa. Non a caso, vi sono ancora molte discussioni su come è meglio catalogare alcune parti del discorso e ogni lingua ha la sua parte di discussione. Anche per l'ebraico non esiste una lista completa e universalmente riconosciuta di POS, e vi è disaccordo tra grammatiche, dizionari e strumenti automatici disponibili. Ogni *tagset*, in definitiva, deve essere creato alla luce delle specificità della lingua che ci si appresta ad annotare.

Una volta definito il *tagset* resta da decidere quale sia la categoria grammaticale più adatta da associare a ogni *token*. Si possono raccogliere essenzialmente due tipi di informazione, il problema è come e se mantenerle entrambe, in particolare: i) la definizione del *token* in senso sintagmatico (i.e. che cosa rappresenta il *token* nel contesto) e ii) l'informazione lessicale che il *token* dà di per sé.

Per fare alcuni esempi i principali punti di disaccordo identificati sono:

- Participio/aggettivo: פְּסוּלָה → è un participio passivo o un aggettivo?

- Verbo/sostantivo: הַמַּדִּיר אֶת אִשְׁתּוֹ → "colui che fa un voto" oppure "votante"? (colui che consacra la moglie): si deve assegnare alla categoria verbo o sostantivo?

- Participio/indicativo: רַבִּי אֱלִיעֶזֶר בֶּן יַעֲקֹב אוֹמֵר → è un participio o un presente indicativo?

- Frase avverbiale o avverbio: שׁוֹאֵל מִפְּנֵי הַיִּרְאָה וּמֵשִׁיב מִפְּנֵי הַכָּבוֹד → va analizzato come *myn* + *pnym* o come *mpny*?

- Aggettivo/verbo: אִם יְכוֹלִין לְהַתְחִיל וְלִגְמוֹר עַד שֶׁלֹּא יַגִּיעוּ לַשּׁוּרָה — יַתְחִילוּ, *ykwlyn* è un aggettivo o un verbo?

Potremmo discutere molto su quale categoria sarebbe meglio loro assegnare e sul perché, ma recenti ricerche hanno dimostrato che mantenere più informazioni possibile è essenziale per avere un'analisi accurata. Con la creazione delle due pagine "categoria" e "funzione" siamo voluti andare incontro a questa esigenza.

Il *tagset* (seppur provvisorio) che qui suggeriamo per l'annotazione dell'ebraico della Mišnah è il seguente: aggettivo, avverbio, congiunzione, copula, esistenziale, interiezione, nome, numerale, prefisso, preposizione, pronome, nome proprio, verbo. Si potrebbe prevedere di aggiungere, inoltre: interrogativo, modale, negazione e quantificatore.[5]

Resta da definire esattamente che cosa si intenda con *token*. Lo studioso israeliano Shuly Wintner, nel suo capitolo sull'analisi morfologica delle lingue semitiche, afferma: "*Since most computational applications deal with written texts (as opposed to spoken language), the most useful notion is that of an orthographic word. This is a string of characters, from a well-defined alphabet of letters, delimited by spaces, or other delimiters, such as punctuation. A text typically consists of sequences of orthographic words, delimited by spaces or punctuation; orthographic words in a text are often referred to as tokens.*"[6]

La parola ortografica, o *token*, può essere, a sua volta, divisa in unità più piccole, i morfemi. Le radici sono alcuni tipi particolari di morfemi. Anche noi abbiamo quindi adottato la scelta di definire *token* ciò che in un testo scritto è delimitato da spazi e punteggiatura.

A partire dalle nozioni qui introdotte, nella sezione che segue sono descritte, per sommi capi, le modalità di annotazione linguistica automatica di testi.

[5] Per una discussione su questo si veda: M. Adler, *Hebrew Morphological Disambiguation: An Unsupervised Stochastic Word-based Approach*, Doctoral dissertation, Ben-Gurion University of the Negev 2007; Y. Netzer, M. Adler *et al.*, *Can You Tag the Modal? You Should.*, «Proceedings of the ACL-2007 Workshop on Computational Approaches to Semitic Languages», Prague 2007; Y. Netzer, M. Elhadad, *Generating Determiners and Quantifiers in Hebrew*, «Proceedings of the Workshop on Computational Approaches to Semitic Languages», Association for Computational Linguistics, Montreal 1998, pp. 89-96.

[6] S. Wintner, *Morphological Processing of Semitic Languages*, in I. Zitouni (ed.), «Natural Lan-

3. *L'annotazione linguistica automatica* di Andrea Bellandi

L'elaborazione del linguaggio naturale (o trattamento automatico della lingua, abbreviato in TAL, in inglese *Natural Language Processing*) è un'area che coinvolge l'informatica e l'intelligenza artificiale che si occupa dello studio di modelli e tecniche per l'analisi automatica delle lingue naturali. Tale area può essere ricondotta alla Linguistica Computazionale, un campo interdisciplinare che, più in generale, studia la lingua da un punto di vista teorico e applicativo in una prospettiva digitale. Tra i compiti fondamentali della linguistica computazionale e, in particolare, del TAL, vi è quello di definire modelli e metodi per l'annotazione linguistica automatica di testi tramite strumenti informatici.[7]

I modelli per il TAL possono essere di due tipi: a regole oppure statistici. I modelli statistici, a loro volta, possono essere di tipo supervisionato (*supervised*) o non supervisionato (*unsupervised*).[8] Attraverso i modelli supervisionati si addestra un calcolatore (in una procedura detta *training*) a svolgere il compito di annotazione linguistica automatica a partire da un insieme, detto *training set*, di testi già correttamente annotati.

Sebbene l'annotazione linguistica automatica includa varie fasi di analisi incrementali (di solito: la suddivisione di un testo in frasi, la suddivisione di frasi in *token*, l'analisi morfologica, la lemmatizzazione e l'analisi sintattica) in questo articolo ci focalizzeremo sull'analisi morfologica effettuata attraverso un approccio statistico supervisionato.

Il processo statistico di individuazione automatica delle "parti del discorso" (in inglese *Part-of-Speech* o POS) all'interno di un testo, così come schematizzato in Fig. 1, fa uso di algoritmi di apprendimento automatico (in inglese *machine learning*) supervisionato. Una volta compiuto l'addestramento dell'analizzatore questo verrà utilizzato per determinare la POS delle parole che compongono i testi da annotare linguisticamente. In breve, i passi da compiere per l'annotazione di un testo sono: i) creazione di un *corpus* di addestramento annotato manualmente, ii) addestramento di un analizzatore statistico mediante il *corpus* creato in precedenza, iii) analisi di nuovi testi.

A queste tre fasi iniziali se ne aggiunge generalmente una quarta per la valutazione delle prestazioni dell'analizzatore.

Innanzitutto, valutare le prestazioni di un analizzatore con lo stesso insieme di dati utilizzato per l'addestramento non sarebbe utile: il sistema, infatti, "ricordando" le analisi dei testi su cui è stato addestrato, se provato sui medesimi testi, li annoterebbe in modo perfetto. Si

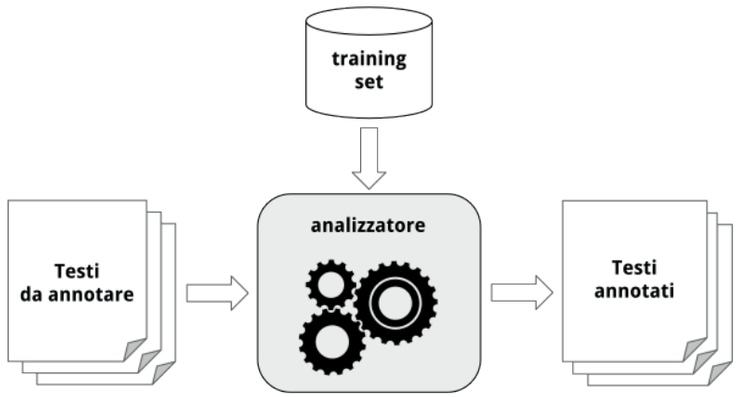


Fig. 1 - Schema sintetico del processo di addestramento di un analizzatore linguistico.

guage Processing of Semitic Languages», Springer, Heidelberg 2014, p. 44.

[7] A. Lenci, S. Montemagni, V. Pirrelli, *Testo e computer. Elementi di linguistica computazionale*, Carocci editore, Roma 2016.

[8] C.D. Manning, H. Schütze, *Foundations of*

tratta, quindi, di definire un insieme di dati correttamente annotati, diverso dal *training set*, da sottoporre all'analizzatore per verificarne l'accuratezza (Fig. 2).

Questo *test set* viene di solito ricavato da una porzione del *corpus* annotato di partenza che viene quindi suddiviso in due parti: il *training set* (per l'addestramento) e il *test set* (per la valutazione). C'è però un altro problema: il *test set* (solitamente di dimensioni molto inferiori rispetto al *training set*) potrebbe non essere rappresentativo delle caratteristiche linguisti-

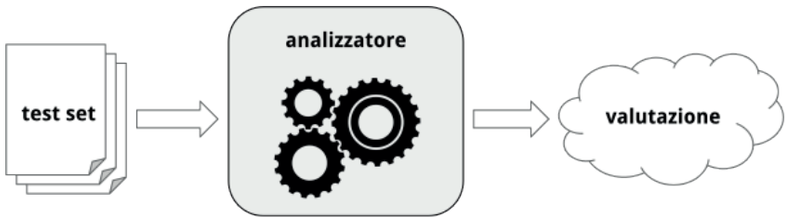


Fig. 2 - Schema sintetico di processo di valutazione di un analizzatore.

che del *corpus* dal quale è stato ricavato: se non viene selezionato nel modo opportuno potrebbe quindi portare a una valutazione errata delle prestazioni dell'annotatore. Per evitare questo problema esistono diverse metodologie, la più utilizzata è la cosiddetta *k-fold cross-validation*.[9] Questa metodologia, quando applicata alla valutazione di analizzatori linguistici, prevede che il *training set* e il *test set* siano estratti dal *corpus* annotato e applicati per l'addestramento e la validazione in una serie di iterazioni. Come rappresentato in Fig. 3, il *corpus* annotato viene partizionato in k porzioni (dette *fold*) di dimensione il più possibile equivalente. Successivamente, vengono effettuate k iterazioni di addestramento e di validazione. A ogni iterazione, una porzione viene utilizzata come *test set* (casella scura) e, le rimanenti k-1 porzioni (caselle bianche), come *training set*. L'annotatore viene quindi addestrato con il *training set* e, quindi, applicato all'annotazione linguistica del *test set* per valutarne le prestazioni. Una misura classica adottata a questo scopo è la cosiddetta *precision*, intesa come il rapporto tra il numero di annotazioni corrette e il numero di annotazioni totali. Una volta effettuate tutte le k iterazioni, il valore di *precision* finale è calcolato come media dei valori di *precision* ottenuti in ogni singola iterazione.

3.*1. Il problema dell'annotazione delle lingue del Talmud* di Alessandra Pecchioli

L'annotazione linguistica automatica del testo talmudico pone vari problemi. Infatti, al contenuto e alla complessità filologica del Talmud si aggiunge la ricchezza linguistica di cui si è parlato nell'introduzione e che, inevitabilmente, ne rende complicata l'annotazione automatica.[10]

Tra le risorse linguistiche disponibili per l'ebraico antico e l'aramaico, potenzialmente

Statistical Natural Language Processing, MIT Press, Cambridge MA 1999.

[9] R. Kohavi, *A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection*, «Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 2», Morgan Kaufmann Publishers, San Francisco 1995, pp. 1137-1143.

[10] A. Bellandi, A. Bellusci, E. Giovannetti, *Computer Assisted Translation of Ancient Texts:*

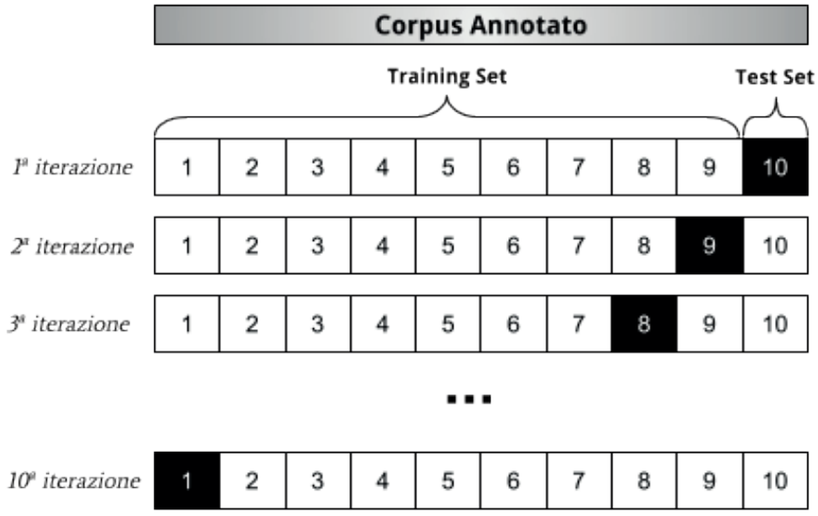


Fig. 3 - Il processo di valutazione con la tecnica iterativa del *k-fold cross-validation* con k pari a 10.

utilizzabili per l'addestramento di strumenti di annotazione automatica, citiamo qui: i) il database linguistico Hebrew Text Database[11] ETCBC accessibile attraverso SHEBANQ[12] un ambiente online per lo studio (prevalentemente della sintassi) della lingua ebraica di epoca biblica, sviluppato presso l'Eep Talstra Centre for Bible and Computer presso la Vrije Universiteit di Amsterdam, ii) il progetto Historical Dictionary[13] della Academy of the Hebrew Language di Israele, iii) il Comprehensive Aramaic Lexicon (CAL)[14] realizzato dallo Hebrew Union College di Cincinnati, e, infine, iv) il progetto Digital Mishna,[15] per la creazione di una edizione critica digitale della Mišnah condotto dal Maryland Institute of Technology in the Humanities.

Oltre alle citate risorse linguistiche non sono disponibili, a oggi, strumenti per l'annotazione automatica di lingue antiche semitiche nord-occidentali (mentre numerosi sono gli strumenti con annotazione manuale: Accordance,[16] Bibleworks,[17] Logos,[18] per citare solo i più conosciuti). Tra gli strumenti esistenti per l'analisi automatica dell'ebraico contemporaneo menzioniamo MILA,[19] specificatamente implementato per il trattamento dell'ebraico moderno, che di conseguenza, non è immediatamente utilizzabile per l'analisi delle lingue del Talmud. A tale proposito abbiamo condotto alcuni esperimenti di annotazione automatica con MILA su porzioni di testo redatte nelle principali fasi linguistiche del Talmud. Si riportano, di seguito, alcune con-

The Babylonian Talmud Case Study, «Proceedings of the 11th International Workshop on Natural Language Processing and Cognitive Science», B. Sharp, R. Delmonte, De Gruyter (edd.), München 2014, pp. 287-302.

[11] D. Roorda, W.T. van Peursen, C. Sikkel, *Hebrew Text Database ETCBC4b*, Data Archiving and Networked Services, Amsterdam 2015.

[12] https://shebanq.ancient-data.org/.

[13] http://maagarim.hebrew-academy.org.il.

[14] http://cal.huc.edu/.

[15] http://www.digitalmishnah.org/.

[16] https://www.accordancebible.com/.

[17] https://bibleworks.com/.

[18] https://www.logos.com/.

[19] A. Itai, S. Wintner, *Language Resources for Hebrew. Language Resources and Evaluation*, 42 (1), 2008: 75–98. https://doi.org/10.1007/s10579-007-9050-8.

siderazioni per le tre lingue analizzate:

ebraico biblico: la maggior parte del lessico dell'ebraico biblico si è conservato anche nell'ebraico moderno, pertanto, MILA ha riconosciuto la maggioranza dei vocaboli dal testo biblico. Molte costruzioni sintattiche tipiche dell'ebraico biblico sono, tuttavia, scomparse o mutate nell'ebraico moderno e, pertanto, MILA in questi casi non ha analizzato il testo secondo le regole morfo-sintattiche proprie dell'ebraico biblico, proponendo un'analisi errata.

ebraico mišnaico: l'ebraico moderno conserva la maggior parte del lessico e delle caratteristiche morfo-sintattiche dell'ebraico mišnaico, MILA, di conseguenza, ha analizzato correttamente la maggior parte delle parole. Tuttavia, le specificità dell'ebraico mišnaico, che in ebraico moderno risulterebbero errate, non vengono analizzate correttamente.

aramaico: l'ebraico e l'aramaico sono lingue semitiche ben distinte, il cui patrimonio lessicale coincide solo in parte. In particolare, ci sono casi in cui la stessa radice assume un diverso significato semantico in ebraico e in aramaico. MILA non ha riconosciuto le radici aramaiche e le ha analizzate come nomi propri. In presenza di vocaboli frequenti in aramaico ma poco attestati in ebraico moderno, MILA ha restituito un alto numero di disambiguazioni, in quanto, non comprendendo la grammatica aramaica, ha tentato di ricercare tutte le possibili radici da cui la parola in esame potrebbe essersi generata. Inoltre MILA non ha riconosciuto le forme morfologiche e sintattiche tipiche dei dialetti aramaici.

In letteratura[20] esistono comunque molte altre iniziative legate all'analisi automatica dell'ebraico moderno a vari livelli linguistici: strumenti specifici per la tokenizzazione (HebTokenizer[21]), progetti orientati al rilascio di strumenti *open source* per l'analisi grammaticale e la lemmatizzazione (MorphTagger,[22] NLPH[23]) e, infine, strumenti per l'analisi sintattica (yap,[24] hebdepparser,[25] UD_Hebrew[26]).

Appurata la difficoltà nel riutilizzare quanto già disponibile tra le risorse linguistiche e i software sviluppati è sorta l'esigenza di creare, all'interno del progetto, una propria risorsa annotata e una propria metodologia per l'annotazione linguistica automatica.

4. *L'annotazione della Mišnah* di Davide Albanesi

Per definire, innanzitutto, il nostro *tagset* si segue il seguente procedimento: i) dato il nostro *corpus* letterario (la Mišnah del Talmud babilonese) per iniziare a creare il *tagset* si consultano dizionari e grammatiche (vd più avanti in particolare), provando a riutilizzare le linee guida di *tagset* anche di altre lingue e cercando di adattarle al nuovo contesto, se sembrano utili e appropriate. Questa fase è piuttosto delicata in quanto, optare per un tipo di *tagset* piuttosto che per un altro rivela la nostra prospettiva sull'intera lingua e, considerato che molte parti sono oscure, contemporaneamente ci spinge a fare scelte che potrebbero sembrare drastiche, ma che vanno incontro alle capacità di apprendimento della macchina; ii) si comincia quindi l'annotazione; iii) dopo un certo periodo di lavoro si valuta l'accordo e/o il disaccordo tra le prime annotazioni e le più recenti, iv) si identificano i principali elementi di disaccordo; v) si ridefinisce il *tagset* e i suoi criteri, ricominciando dal punto ii).

Parallelamente, si comincia ad addestrare la macchina sulla porzione di *corpus* annotato, testando le sue capacità di analisi e imparando

[20] Cfr. D. Kamir, S. Naama, N. Yoni, *A Comprehensive NLP System for Modern Standard Arabic and Modern Hebrew*, «Proceedings of the ACL-02 Workshop on Computational Approaches to Semitic Languages», Association for Computational Linguistics, Stroudsburg 2002; cfr. S.B. Cohen, N.A. Smith, *Joint Morphological and Syntactic Disambiguation*, «Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning», Association for Computational Linguistics, Prague 2007.

[21] https://www.cs.bgu.ac.il/~yoavg/software/hebtokenizer.

[22] R. Bar-Haim, K. Sima'an, Y. Winter, Part-of-speech tagging of Modern Hebrew text, «Natural Language Engineering» 14 (2), Cambridge University Press, Cambridge 2008, pp. 223-251, http://www.cs.technion.ac.il/~barhaim/MorphTagger/.

[23] https://github.com/NLPH/NLPH.

[24] https://github.com/habeanf/yap.

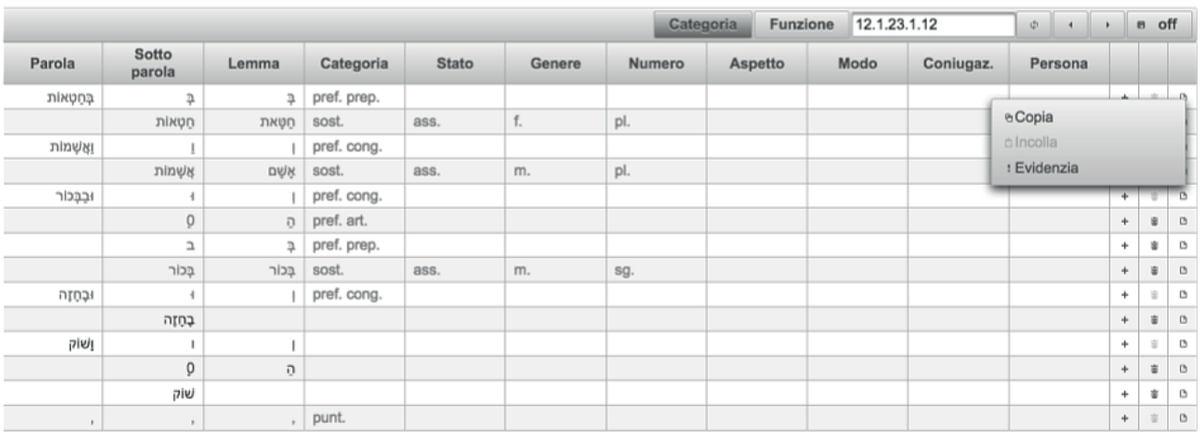
[25] https://www.cs.bgu.ac.il/~yoavg/software/hebparsers/hebdepparser/.

[26] https://github.com/UniversalDependencies/UD_Hebrew.

dagli errori. Scendendo più nei particolari, allo scopo di implementare un annotatore automatico dell'ebraico mišnaico, sono state estratte dal Talmud tutte le frasi della Mišnah ed è stata realizzata un'applicazione Web per l'annotazione grammaticale manuale (Fig. 4).

Tutte le informazioni finora registrate verranno in un secondo tempo utilizzate come base statistica per addestrare un sistema di annotazione automatica per l'analisi dell'intera Mišnah. Perché la precisione di questi software sia accettabile è necessario riuscire ad annotare manualmente il maggior numero di frasi possibile. Ad oggi sono state annotate linguisticamente 3356 frasi, per un totale di 10442 tokens.

Il software realizzato per l'annotazione presenta in forma tabellare le informazioni dell'analisi effettuata frase per frase. In alto a destra viene presentato l'indice della frase corrente in un campo di testo e due pulsanti per navigare avanti e indietro nella lista delle frasi. Il campo di testo è modificabile ed è possibile accedere rapidamente alla frase che si vuol visualizzare. Il sistema, una volta selezionata una frase da annotare, verifica se le parole (*tokens*) che la compongono siano già state analizzate e, in questo caso, ne calcola una probabile suddivisione in sottoparole (*subtokens*) attribuendovi le informazioni grammaticali già inserite in precedenza. Questo approccio basato sul riuso del



Parola	Sotto parola	Lemma	Categoria	Stato	Genere	Numero	Aspetto	Modo	Coniugaz.	Persona
בְּחִטְאוֹת	בְּ	בְּ	pref. prep.							
	חִטְאוֹת	חַטָּאת	sost.	ass.	f.	pl.				
וְאַשְׁמוֹת	וְ	וְ	pref. cong.							
	אַשְׁמוֹת	אָשָׁם	sost.	ass.	m.	pl.				
וּבְבָכוֹר	וּ	וְ	pref. cong.							
	בַּ	הַ	pref. art.							
	בְּ	בְּ	pref. prep.							
	בְּכוֹר	בְּכוֹר	sost.	ass.	m.	sg.				
וּבְחַזָּה	וּ	וְ	pref. cong.							
	בְּחַזָּה									
וְשׁוֹק	וְ	וְ								
	שׁוֹק	הַ								
,	,	,	punt.							

Fig. 4 - L'interfaccia per l'annotazione linguistica del corpus da utilizzarsi per l'addestramento dell'annotatore automatico.

lavoro pregresso, similmente a quanto avviene nella memoria di traduzione del sistema Traduco, consente di velocizzare le attività di annotazione evitando di ripetere le analisi già effettuate. Nei casi in cui il sistema trovi analisi multiple associate ad una specifica parola propone l'analisi più frequente.

In merito all'annotazione linguistica sono stati adottati la grammatica di Pérez Fernández[27] e, per la lemmatizzazione, il dizionario di Marcus Jastrow.[28] Il software così concepito permette di raccogliere il maggior numero di informazioni possibile su ogni lessema in quanto prevede una doppia descrizione: nella parte "categoria" si raccolgono i dati su ciò che ogni lemma rappresenta dal punto di vista grammaticale mentre, nella parte "funzione", ci si concentra sulla descrizione della funzione assunta da quel dato lemma quando questa non corrisponde alla "categoria".

[27] M. Pérez Fernández, J.F. Elwolde, *An Introductory Grammar of Rabbinic Hebrew*, Brill, Leiden 1999.

[28] M. Jastrow, *A Dictionary of the Targumim, the Talmud Babli and Yerushalmi, and the Midrashic literature*, Judaica Press, New York 1971.

5. *Addestramento e valutazione dell'annotatore* di Simone Marchi

Una volta costituito il *corpus* mišnaico annotato linguisticamente si è proceduto all'addestramento e alla valutazione dell'annotatore automatico (POS tagger). A questo fine sono stati scelti due tra gli algoritmi più utilizzati per il POS *tagging*: HunPos[29] (derivato dal TnT PoS Tagger)[30] e lo Stanford Log-linear Part-Of-Speech Tagger.[31] Entrambi gli algoritmi implementano modelli statistici supervisionati (cfr. sezione 3) e, di conseguenza, devono essere addestrati a partire da un *corpus* annotato manualmente. Parallelamente all'addestramento e alla concomitante valutazione degli annotatori scelti sono anche stati effettuati alcuni esperimenti volti a dimostrare quanto sia importante produrre corpora annotati di dimensioni rilevanti. In particolare, si è proceduto alla selezione di 4 distinti corpora di addestramento, corrispondenti a porzioni del 25%, 50%, 75% e 100% dell'intero *corpus* annotato. Per ognuno di questi *corpus* è stata adottata la strategia sopradescritta del *k-fold cross validation*, con k pari a 10, dividendo, cioè, il *corpus* in 10 partizioni.

In Fig. 5 sopra sono riportati i risultati dell'esperimento. Vi sono quattro coppie di colonne: quelle di colore grigio chiaro rappresentano la *precision* dello Stanford POS tagger mentre quelle più scure rappresentano la *precision* dell'HunPos Tagger. Dal grafico si evidenziano tre fatti: i) a parità di dimensione del *corpus* di addestramento l'algoritmo di Stanford fornisce prestazioni superiori rispetto a HunPos, ii) all'aumentare della dimensione del *corpus* aumenta la *precision* degli annotatori addestrati, iii) con un numero di *tokens* annotati di poco superiore a diecimila già si riescono a raggiungere buone performance, con una *precision* maggiore del 90% (in altre parole, l'annotatore sbaglia ad attribuire la POS una volta su dieci).

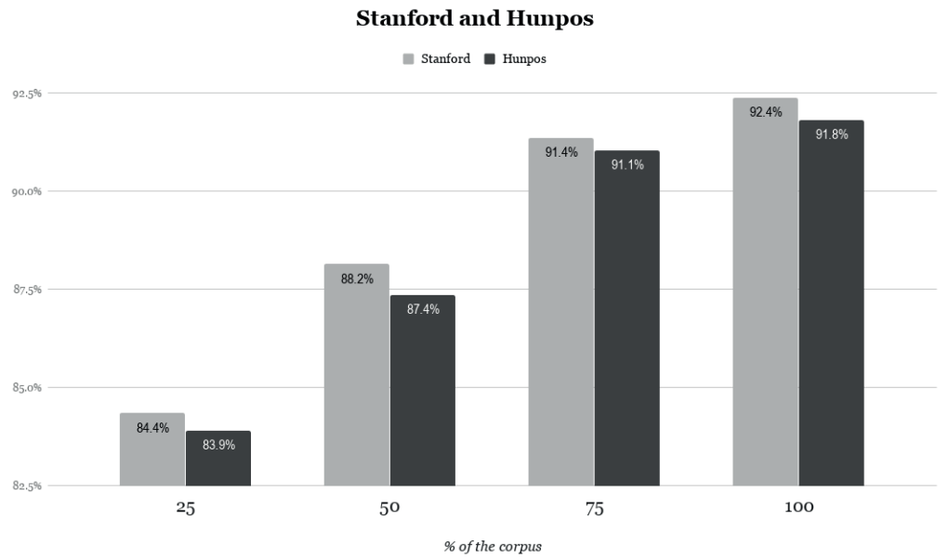


Fig. 5 - La *precision* dei due algoritmi scelti, calcolata con la tecnica del 10-fold cross-validation, migliora all'aumentare delle dimensioni del corpus annotato.

[29] P. Halácsy, A. Kornai, C. Oravecz, *HunPos: an open source trigram tagger*, «Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions (ACL '07)». Association for Computational Linguistics, Stroudsburg 2007, pp. 209-212.

[30] T. Brants, *TnT: A Statistical Part-of-Speech Tagger*, «Proceedings of the Sixth Conference on Applied Natural Language Processing (ANLC '00)». Association for Computational Linguistics, Stroudsburg 2000, pp. 224-231.

[31] K. Toutanova, D. Klein, C.D. Manning, Y. Singer, *Feature-Rich Part-of-Speech Tagging With a Cyclic Dependency Network*, «Proceedings of the

6. *Prossimi passi* di Emiliano Giovannetti

Al momento, l'annotatore automatico è stato addestrato per attribuire ai *token* del testo soltanto la categoria grammaticale. Il naturale sviluppo di questa ricerca prevede che, alle informazioni linguistiche da fornire in addestramento, si aggiunga anche il lemma. Un sistema così addestrato, che da "POS-*tagger*" diviene "lemmatizzatore", è in grado di attribuire automaticamente, a ogni parola analizzata, il relativo lemma, ovviamente con uno specifico grado di accuratezza da valutarsi secondo i medesimi criteri descritti nella sezione 3. Non appena la Mišnah del Talmud sarà stata lemmatizzata il software Traduco per la traduzione del Talmud in italiano, come descritto nella sezione 2.1, potrà sfruttare questa informazione aggiuntiva (al momento per la sola Mišnah) al fine di: i) sottoporre al traduttore suggerimenti di traduzione su base lemmatica e ii) consentire di effettuare ricerche per lemma. In una fase ulteriore di questa ricerca sarà considerata l'annotazione linguistica delle porzioni del Talmud scritte in altre lingue, a partire dall'aramaico babilonese, lingua nella quale è redatta la Gemarah.

7. *Riconoscimenti*

Il presente lavoro è stato sviluppato nell'ambito del progetto scientifico TALMUD e della collaborazione scientifica fra S.ca r.l. PTTB e ILC-CNR.

Alessandra Pecchioli
Università di Firenze
e-mail: alepec3@gmail.com

Davide Albanesi
ILC-CNR
e-mail: davide.albanesi@ilc.cnr.it

Andrea Bellandi
ILC-CNR
e-mail: andrea.bellandi@ilc.cnr.it

Emiliano Giovannetti
ILC-CNR
e-mail: emiliano.giovannetti@ilc.cnr.it

Simone Marchi
ILC-CNR
e-mail: simone.marchi@ilc.cnr.it

SUMMARY

The automatic linguistic analysis of ancient Hebrew represents a new research opportunity in the field of Jewish studies. In fact, very little has been produced, both in terms of linguistic resources and, above all, of tools for the analysis of ancient Hebrew. This article illustrates a work born within the Italian Translation of the Babylonian Talmud Project aimed at the construction of an automatic linguistic annotator of Mišnaic Hebrew.

KEYWORDS: Mišnaic Hebrew; Babylonian Talmud; Natural Language Processing.

2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology» Volume 1, Association for Computational Linguistics, Stroudsburg 2003, pp. 173-180.