IJCoL

Italian Journal of Computational Linguistics

Rivista Italiana di Linguistica Computazionale

Volume 2, Number 2
december 2016

Special Issue:
Digital Humanities and Computational Linguistics

Guest Editors:
John Nerbonne, Sara Tonelli

aAccademia university press

Registrazione presso il Tribunale di Trento n. 14/16 del 6 luglio 2016

Rivista Semestrale dell'Associazione Italiana di Linguistica Computazionale (AILC)

direttore responsabile
Michele Arnese

isbn 978-88-99982-26-3

Accademia University Press
via Carlo Alberto 55
I-10123 Torino
info@aAccademia.it
www.aAccademia.it/IJCoL_2_2

IJCoL

Volume 2, Number 2
december 2016

Special Issue:
Digital Humanities and Computational Linguistics

Guest Editors:
John Nerbonne, Sara Tonelli

CONTENTS

Il Sistema *Traduco* nel Progetto Traduzione del Talmud Babilonese

Andrea Bellandi*
Istituto di Linguistica Computazionale
"A. Zampolli"

Davide Albanesi*
Istituto di Linguistica Computazionale
"A. Zampolli"

Giulia Benotto*
Istituto di Linguistica Computazionale
"A. Zampolli"

Emiliano Giovannetti*
Istituto di Linguistica Computazionale
"A. Zampolli"

In the context of the Babylonian Talmud Translation Project, the Institute of Computational Linguistics of the CNR has developed Traduco, a collaborative web tool with some features that make it particularly suitable for translating texts that present interpretative problems. To date, Computer-Assisted Translation (CAT) tools are typically used for the translation of technical manuals, legislative texts, or websites and are mainly aimed at speeding up the translation process. Traduco incorporates most of the standard components of a traditional computer-assisted translation tool, but extends them with specific features necessary to support the interpretation and translation of complex texts that pose particular comprehension problems. In this article, we will present a specific case study related to a text with these characteristics: the babylonian Talmud. Traduco includes features for adding notes, bibliographic references, semantic annotations, and the creation of glossaries. Translators, revisors, editors, supervisors, and end-users accessing the system are supported throughout the translation process, from interpreting the original text to the editorial phase for the printing of translations, through the use of computer-assisted translation technologies, semantic annotation of the text, enrichment of translations with explanatory information, export of translations in XML and TEI, and integration of techniques for natural language processing. The design and development of Traduco required the adoption of a multidisciplinary approach that combines aspects of software engineering, computational linguistics, knowledge engineering, and digital publishing.

1. Il Progetto e il ruolo dell'Istituto di Linguistica Computazionale

Nato da un protocollo di intesa tra la Presidenza del Consiglio dei Ministri, il MIUR, il CNR, e l'UCEI - Collegio Rabbinico Italiano, il Progetto Traduzione del Talmud Babilonese[1] viene finanziato dal MIUR come progetto speciale del CNR. L'obiettivo del Progetto è la traduzione digitalizzata, in lingua italiana, del Talmud Babilonese (in breve, TB).

Il TB rappresenta uno dei testi fondamentali dell'Ebraismo. Esso costituisce la "registrazione" delle discussioni rabbiniche avvenute tra i Maestri di generazioni risalenti

* Istituto di Linguistica Computazionale "A. Zampolli", Via G. Moruzzi, 1 - CNR Pisa, Italy.
E-mail: name.surname@ilc.cnr.it

1 https://www.talmud.it/ (ultimo accesso: 28 novembre 2016)

al periodo compreso tra il III ed il V-VI secolo circa la comprensione del significato all'origine degli insegnamenti biblici, ma anche alla loro concreta applicabilità. Compilato nelle accademie talmudiche dell'antica Babilonia (l'attuale Iraq), il TB si basa su fonti provenienti da diverse epoche ed aree geografiche e si è mantenuto in una forma "fluida" almeno fino al VI secolo d.C. Il TB è diviso in sei "ordini" ("sedarim") corrispondenti a diverse categorie della legge ebraica, e discute 37 trattati, per un totale di 2711 folia nell'edizione stampata (Vilna, XIX secolo). Il Talmud si occupa di etica, giurisprudenza, liturgia, rituali, filosofia, commercio, medicina, astronomia, magia e molto altro ancora.

Il Progetto vede coinvolto, fin dal suo avvio nell'anno 2012, l'Istituto di Linguistica Computazionale "A. Zampolli" del CNR (in breve, ILC) come partner scientifico e tecnologico. Di fatto, l'adozione di un approccio interamente digitale alla traduzione collaborativa di un testo complesso come il Talmud Babilonese rende il Progetto unico nel suo genere.

Il lavoro di ricerca e sviluppo condotto presso l'ILC ha dato alla luce *Traduco*, un'applicazione web collaborativa progettata specificamente per la traduzione di testi che pongono particolari difficoltà interpretative.

I Sistemi di Traduzione Assistita (STA) si basano, essenzialmente, sul concetto di Memoria di Traduzione (MT), tramite la quale è possibile riutilizzare traduzioni precedenti e lavorare più velocemente. Questa tecnica permette principalmente di i) assicurare che il documento sia tradotto completamente, ii) garantire che i documenti tradotti siano coerenti nelle definizioni comuni, il fraseggio e la terminologia, iii) accelerare il processo complessivo di traduzione. Per una breve rassegna sulla tecnologia delle memorie di traduzione si faccia riferimento a (Somers 2003), (Reinke 2006), (Lagoudaki 2009).

Ad oggi esistono molti STA sia commerciali, come ad esempio Across[2], Déjà Vu[3], memoQ[4], SDL Trados[5], Similis[6], Transit NXT[7] e Wordfast[8], che non, come ad esempio OpenTM[9], OmegaT[10], Transolution[11], Matecat[12] e TinyTM[13]. Ognuno di questi strumenti supporta il traduttore nella propria attività sotto differenti aspetti. Alcuni integrano la memoria di traduzione con tecniche di Traduzione Automatica (TA), altri con analisi linguistico-semantiche e con memorie di traduzione multilingua aperte. Altri facilitano la gestione delle traduzioni tramite sistemi collaborativi che permettono la profilazione utente e il monitoraggio dell'avanzamento delle traduzioni stesse. Generalmente, tutti questi strumenti sono applicati alla traduzione di testi legislativi, manuali tecnici (ad esempio medicina o informatica), pagine Web, sottotitoli di film, ted talks, ecc... (vedi ad esempio i corpora paralleli in `http://opus.lingfil.uu.se/index.php`) e il loro obiettivo principale è, specialmente nel caso di progetti con documentazione di grandi

2 http://www.my-across.net/en/ (ultimo accesso: 28 novembre 2016)
3 http://www.atril.com/ (ultimo accesso: 28 novembre 2016)
4 https://www.memoq.com/ (ultimo accesso: 28 novembre 2016)
5 http://www.sdl.com/ (ultimo accesso: 28 novembre 2016)
6 http://www.similis.fr/ (ultimo accesso: 28 novembre 2016)
7 http://www.star-ts.com/ (ultimo accesso: 28 novembre 2016)
8 http://www.wordfast.com/ (ultimo accesso: 28 novembre 2016)
9 http://www.opentm2.org/ (ultimo accesso: 28 novembre 2016)
10 http://www.omegat.org/ (ultimo accesso: 28 novembre 2016)
11 https://bitbucket.org/fredrik_corneliusson/transolution/
12 https://www.matecat.com/ (ultimo accesso: 28 novembre 2016)
13 http://tinytm.sourceforge.net/en/index.html (ultimo accesso: 28 novembre 2016)

dimensioni, accelerare il processo complessivo di traduzione per risparmiare tempo e denaro.

L'approccio adottato per lo sviluppo di *Traduco*, invece, è stato basato, fin dal principio, sulle esigenze di traduttori impegnati in testi la cui traduzione richiede particolari difficoltà interpretative. La traduzione di tali testi, inoltre, può richiedere competenze molto specifiche nella lingua di origine che nel contenuto. Interpretare e tradurre una frase, o spesso anche una singola parola, può richiedere molto tempo e un traduttore ha spesso la necessità di condividere e confrontare una propria interpretazione con quella di altri traduttori. Da questo punto di vista un sistema collaborativo, che permette di accedere, in tempo reale, anche a traduzioni effettuate da altri utenti, diventa una necessità. Per una corretta comprensione da parte del lettore, poi, è necessario prevedere meccanismi di arricchimento del testo tramite note, commenti e glossari di varia natura. Sebbene i più moderni STA integrino anche tecniche di Traduzione Automatica, spesso non sono disponibili risorse linguistiche per il trattamento automatico di lingue antiche oppure grandi corpora paralleli tra la lingua sorgente e quella di destinazione, rendendo, di fatto, impossibile l'utilizzo di tecniche per la TA. E' il caso del Talmud Babilonese, le cui fasi linguistiche dell'ebraico antico e dell'aramaico, limitano la possibilità di un'analisi linguistica, come mostrato nella Sezione 4.

Il resto dell'articolo è strutturato come segue. Nella Sezione 2 verranno descritte le caratteristiche generali di *Traduco*, mentre nella Sezione 3 verranno illustrati i singoli componenti. Nella Sezione 4 saranno discussi aspetti di integrazione di tecnologie per il trattamento automatico della lingua. Nella sezione 5, infine, saranno fornite alcune considerazioni conclusive.

2. Il Sistema *Traduco*

Traduco è formato da vari componenti, ognuno dei quali implementa specifiche funzionalità rivolte principalmente a determinati profili di utenza, come descritto in Figura 1(a). Il Sistema presenta caratteristiche che vanno oltre la semplice traduzione e la relativa stampa. Gli utenti, siano essi traduttori, revisori, redattori o supervisori, possono utilizzare componenti dedicati che consentono di: i) agevolare il processo di traduzione mediante l'utilizzo di una memoria di traduzione per recuperare e confrontare informazioni in maniera efficiente (vedi Sezione 3.1); ii) inserire note, commenti, annotazioni e referenze bibliografiche (vedi Sezione 3.2); iii) supervisionare l'intero progetto, dalla traduzione al *publishing* (vedi Sezione 3.3); iv) esportare traduzioni e note in formati standard per i software di editoria digitale (vedi Sezione 3.4); v) sfruttare algoritmi per il trattamento della lingua come ausilio nel processo di suggerimento della traduzione (vedi Sezione 4). Per la fine del Progetto, Traduco consentirà di accedere alla edizione digitale dei vari trattati che costituiscono il TB e, inoltre, supporterà nella produzione delle relative edizioni a stampa.

Le successive due sottosezioni descrivono sia l'architettura del Sistema che le soluzioni tecnologiche adottate per la sua implementazione.

2.1 Caratteristiche di *Traduco*

Per il design dell'architettura di *Traduco* si è tenuto conto sia delle linee guida per la creazione di modelli e strumenti per l'editoria digitale - come suggerito dalla co-

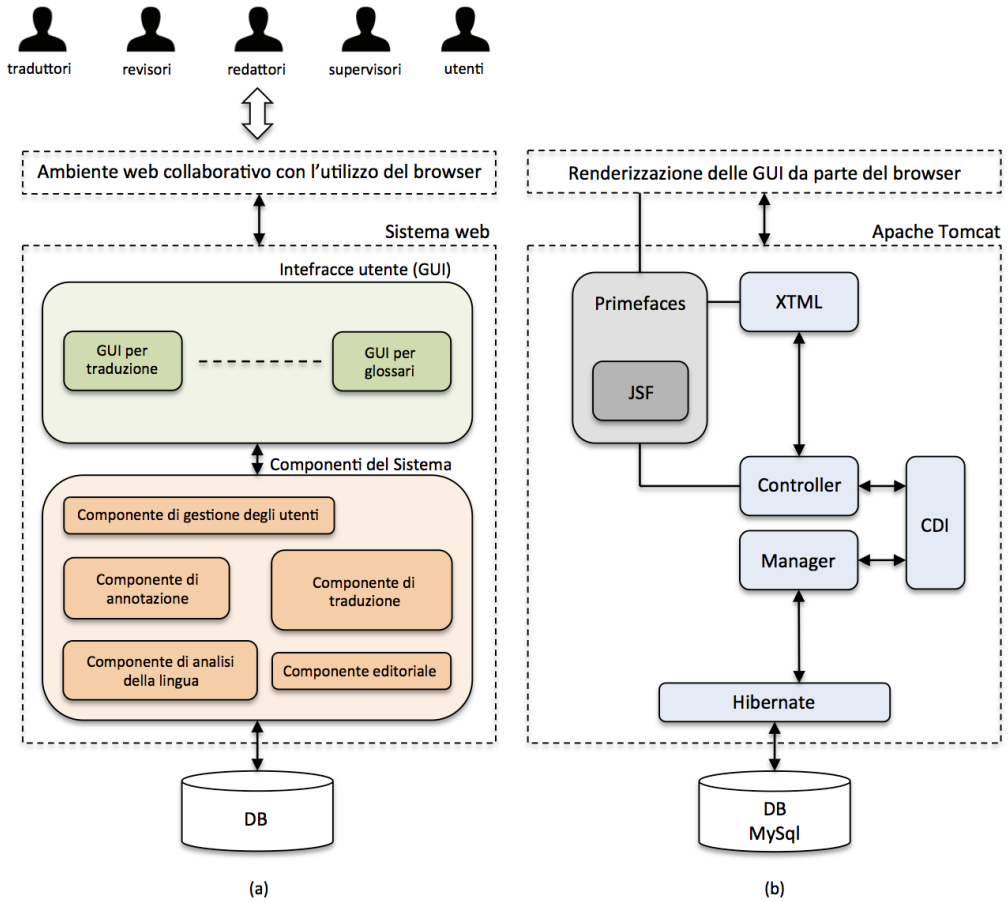


Figura 1
(a) I componenti del Sistema. (b) Architettura software del Sistema.

munità scientifica (ad esempio Interedition[14]) - che dei requisiti utente elaborati in accordo alle esigenze degli utenti del Sistema. Di fatto, nessuno dei sistemi disponibili commercialmente né di quelli liberamente distribuiti in ambito accademico è in grado di soddisfare, allo stesso tempo, tutti i requisiti necessari per tradurre testi di questa natura. Nella tabella di seguito riportata, per riga, riportiamo i diversi STA citati nella sezione precedente e, per colonna, le funzionalità specifiche che si ritengono necessarie per la traduzione di testi simili al Talmud. Oltre alla prima colonna che identifica lo strumento, abbiamo: **Web** - l'interfaccia del sistema è fruibile interamente attraverso un browser e non richiede alcuna installazione locale; **collaborativo** - il sistema consente a un team di utenti di lavorare contemporaneamente alla traduzione di uno stesso testo; **multi-ruolo** - possono accedere al sistema utenti con ruoli e autorizzazioni diverse; **annotazioni e glossari tipizzati** - è possibile associare a porzioni di testo annotazioni di varia natura e creare glossari tematici; **editoriale** - il sistema integra funzionalità di visualizzazione ed esportazione finalizzate alla redazione e alla stampa di volumi.

14 `http://www.interedition.eu/` (ultimo accesso: 28 novembre 2016)

Tabella 1
Comparazione dei principali tools esistenti di supporto alla traduzione.

Strumento	Web	Collaborativo	Multi-ruolo	Annotazioni e glossari tematici	Editoriale
Across	no	no	no	no	no
Déjà Vu	no[15]	parziale	no	no	no
memoQ	sì[16]	parziale[17]	sì	no	no
SDL Trados	no	parziale[18]	sì	no	sì
Similis	no	no	no	no	no
Transit NXT	parziale[19]	no	sì	no	no
Wordfast	sì	parziale[20]	no	no	no
OpenTM	no	parziale[21]	no	no	no
OmegaT	no	no	no	no	no
Transolution	no	no	no	no	no
Matecat	sì	sì	sì	no	no
TinyTM	no	no	no	no	no
Traduco	si	si	si	si	si

Traduco presenta le seguenti caratteristiche generali:

- è stato progettato e realizzato con un'architettura basata a componenti indipendenti connessi attraverso interfacce. La strutturazione architetturale a componenti è agevolata dalle tecnologie utilizzate, imperniate sul linguaggio a oggetti Java (vedi Sezione 2.2);
- è utilizzabile via web. Il Web costituisce l'ambiente di lavoro ideale per attività editoriali; a differenza delle applicazioni desktop, che richiedono installazioni di specifici programmi client sui computer di lavoro, le applicazioni cosiddette Web-based richiedono il solo utilizzo di un browser (e.g. Firefox, Safari, Chrome, ecc.) attraverso il quale l'utente può collegarsi al sistema in esecuzione su una macchina server remota;
- è collaborativo. L'ambiente Web, unitamente alla robustezza dell'impianto tecnologico adottato, consente a un team di utenti (traduttori, revisori, redattori, supervisori, esperti di dominio, ecc.) di poter lavorare sugli stessi dati in modo collaborativo. Il Sistema tiene traccia delle informazioni contenute nella MT e impedisce che gli stessi frammenti di testo siano tradotti da più di una persona. Inoltre, gli utenti supervisori possono

15 Limitatamente alla condivisione di MT (Déjà Vu TEAMserver - http://www.atril.com/node/8437).
16 Tramite MemoQ WebTrans.
17 Limitatamente alla condivisione di risorse e per la revisione (tramite memoQ cloud server).
18 Limitatamente alla condivisione di MT e termbases tramite GroupShare.
19 Limitatamente alla revisione tramite WebCheck.
20 Limitatamente alla condivisione di MT e glossari tramite Wordfast Server.
21 Limitatamente alla condivisione della MT e del dizionario tramite WEB-based Shared Memory.

tenere traccia in tempo reale dello stato di avanzamento del lavoro assegnato ad ogni traduttore;

- è basato su tecnologie open source. L'utilizzo di tali tecnologie è incoraggiato dalla comunità scientifica e permette agli sviluppatori di implementare estensioni e personalizzazioni al codice. In questo caso il Sistema è stato sviluppato utilizzando l'insieme di tecnologie open source del Java 2 Standard Edition (J2SE) framework. Tale framework è la piattaforma tecnologica più stabile, testata e documentata per l'integrazione di sistemi che richiedono gestione della persistenza dei dati, accessi distribuiti, transazionalità delle sessioni e librerie di componenti grafici per interfacce utente;
- è stato dotato di strumenti per l'annotazione del testo e predisposto per l'integrazione di tecniche di elaborazione della lingua: quando possibile, tali strumenti possono essere applicati sia all'originale che al testo tradotto per compiti di annotazione semantica, analisi linguistica (tipicamente tagging morfo-sintattico, stemming o lemmatizzazione), estrazione di terminologia, estrazione di entità nominate, ecc.; un testo annotato può essere poi interrogato su base linguistica, lessicale o semantica, e utilizzato per migliorare le performance della MT;
- è adattabile a diverse lingue: la tecnologia inclusa in *Traduco*, per esempio, si basa su UTF-8 per la codifica dei caratteri (coprendo così la stragrande maggioranza degli alfabeti) e su modelli statistici supervisionati per l'analisi linguistica che possono essere ri-addestrati per elaborare altre lingue (se sono disponibili corpora annotati nelle specifiche lingue).

Traduco, tuttavia, non è ancora dotato di alcune delle funzionalità più classiche dei sistemi di STA, quali l'importazione e l'esportazione di MT, o la gestione di dizionari bilingui.

2.2 Soluzioni Tecniche

Dal punto di vista tecnico, *Traduco* si basa sul design pattern conosciuto come "architettura a tre livelli", e sfrutta Apache Tomcat v7.0 come web server. L'architettura a componenti è stata implementata utilizzando lo Standard Edition framework Java 2 (J2SE), arricchito con le annotazioni CDI (Context Dependency Injection) utilizzando l'implementazione Weld v2.2.4. I servizi di persistenza e di interrogazioni relazionali sono gestiti da Hibernate v4.3.7, che è responsabile della mappatura delle classi Java alle relative tabelle di un database Mysql v5.0. Per ragioni di ottimizzazione di performance nel processo di recupero e confronto delle traduzioni, è stata utilizzata una struttura dati con indice invertito (Patil et al. 2011). Infine, il livello di presentazione è stato realizzato con la tecnologia JavaServer Faces (JSF), un framework per la creazione di interfacce utente basate su componenti per le applicazioni web fondato su Mojarra Oracle v2.2.9, utilizzando la libreria Primefaces v5.1. La Figura 1(b) riassume quanto detto.

3. I Componenti di *Traduco*

Traduco estende la maggior parte dei componenti standard di uno STA tradizionale con caratteristiche specifiche necessarie per supportare la traduzione di testi particolarmen-

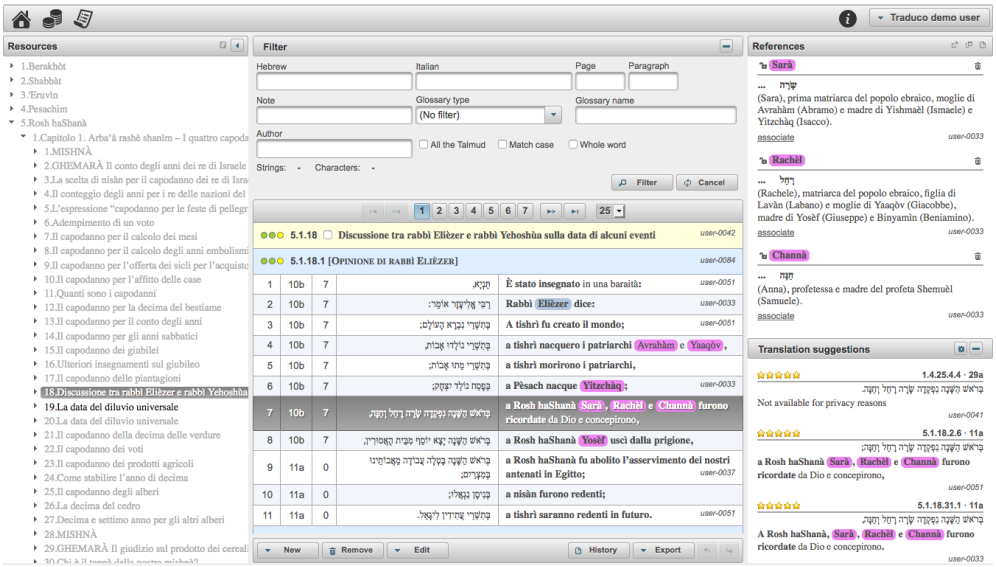


Figura 2
Interfaccia principale di *Traduco*. (a) struttura gerarchica del testo tradotto; (b) tabella delle traduzioni e filtro; (c) risorse della traduzione: note, glossari, suggerimenti alla traduzione.

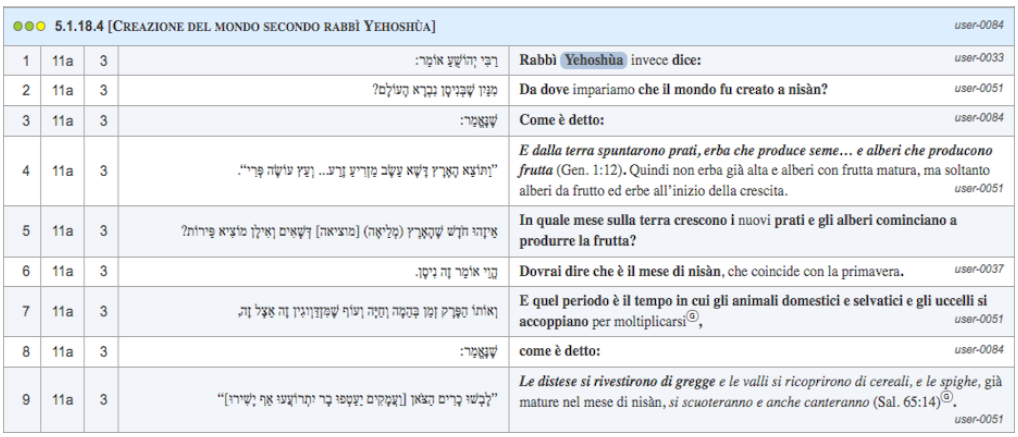
te complessi, come il TB. Il funzionamento di base di *Traduco* non è diverso da quello di altri STA. Come mostrato in Figura 2, un traduttore può visualizzare la struttura gerarchica (a sinistra) del testo tradotto, organizzato, nel caso del TB, in trattati, capitoli, blocchi, unità logiche e stringhe (le unità minime di testo che vengono tradotte).

Nella parte centrale dell'interfaccia, un traduttore può inserire le nuove stringhe affiancate dalle rispettive traduzioni, sia manualmente una dopo l'altra, oppure segmentando parte del testo originale creando più stringhe in una volta per poi tradurle separatamente. La segmentazione del testo è effettuata manualmente da parte del traduttore: la natura del testo, infatti, non consente l'applicazione di tecniche di segmentazione automatica. Appena sopra la tabella di traduzione, una sezione a scomparsa, chiamata "filtro", consente agli utenti di eseguire ricerche, sia sul testo sorgente che sul testo tradotto. Infine, sul lato destro dell'interfaccia sono riportate le note, le voci di glossario e i suggerimenti di traduzione relativi alla stringa selezionata.

In questo modo, il Sistema supporta la traduzione di testi complessi che spesso richiedono riformulazioni specifiche in modo da essere correttamente compresi anche da lettori privi di particolari competenze del testo o del dominio trattato. La Figura 3 mostra come la redazione di ciascuna stringa tradotta può essere eseguita differenziando tra le parti "letterali" (in grassetto) dalle aggiunte esplicative (in tondo), o "informazioni contestuali". Le stringhe che hanno la stessa parte letterale possono differire tra loro per le informazioni contestuali. Il trattato a cui la stringa tradotta appartiene è definito come "contesto".

3.1 Il Sistema di Suggerimento alla Traduzione

Una delle componenti fondamentali di uno STA è il Sistema della Memoria di Traduzione (SMT). Il SMT fa leva su una memoria di traduzione (MT), che consiste di un database



5.1.18.4 [Creazione del mondo secondo rabbì Yehoshùa]				user-0084
1	11a	3	רַבִּי יְהוֹשֻׁעַ אוֹמֵר:	**Rabbì Yehoshùa** invece **dice:** user-0033
2	11a	3	מִנַּיִן שֶׁבְּנִיסָן נִבְרָא הָעוֹלָם?	**Da dove** impariamo **che il mondo fu creato a nisàn?** user-0051
3	11a	3	שֶׁנֶּאֱמַר:	**Come è detto:** user-0084
4	11a	3	"וַתּוֹצֵא הָאָרֶץ דֶּשֶׁא עֵשֶׂב מַזְרִיעַ זָרַע... וַיְצַ עֵץ עֹשֶׂה פְּרִי".	***E dalla terra spuntarono prati, erba che produce seme… e alberi che producono frutta*** (Gen. 1:12)**.** Quindi non erba già alta e alberi con frutta matura, ma soltanto alberi da frutto ed erbe all'inizio della crescita. user-0051
5	11a	3	אֵיזֶהוּ חֹדֶשׁ שֶׁהָאָרֶץ (מְלֵיאָה) [מוציאה] דֶּשָׁאִים וְאֵילָן מוֹצִיא פֵּירוֹת?	**In quale mese sulla terra crescono i** nuovi **prati e gli alberi cominciano a produrre la frutta?**
6	11a	3	הֱוֵי אוֹמֵר זֶה נִיסָן.	**Dovrai dire che è il mese di nisàn,** che coincide con la primavera. user-0037
7	11a	3	וְאוֹתוֹ הַפֶּרֶק זְמַן בְּהֵמָה וְחַיָּה וְעוֹף שֶׁמִּזְדַּוְּגִין זֶה אֵצֶל זֶה.	**E quel periodo è il tempo in cui gli animali domestici e selvatici e gli uccelli si accoppiano** per moltiplicarsi[G]**,** user-0051
8	11a	3	שֶׁנֶּאֱמַר:	**come è detto:** user-0084
9	11a	3	"לָבְשׁוּ כָרִים הַצֹּאן [וְעֲמָקִים יַעַטְפוּ בָר יָרִיעוּ אַף יָשִׁירוּ]"	***Le distese si rivestirono di gregge*** *e le valli si ricoprirono di cereali, e le spighe*, già mature nel mese di nisàn, *si scuoteranno e anche canteranno* (Sal. 65:14)[G]**.** user-0051

Figura 3
Esempio di traduzione letterale (in grassetto) e informazione contestuale alla traduzione (in tondo) inserita al fine di rendere più comprensibile la traduzione italiana.

di coppie di stringhe (stringa source, cioè la porzione di testo originale e stringa target, cioè la relativa traduzione) che memorizza automaticamente tutte le stringhe di testo tradotte insieme al testo di origine durante il processo di traduzione (Reinke 2013). In pratica, lo scopo principale di un SMT è quello di consentire ai traduttori di riutilizzare le traduzioni già fatte. Un SMT è generalmente costituito da:

- un database MT, che contiene coppie di stringhe (s, t), dove s è il segmento in lingua originale del testo e t è la sua traduzione nella lingua di arrivo;
- una funzione di similarità Sim;
- una soglia σ.

Data una stringa s_q da tradurre, il SMT restituisce una traduzione t_q ricercando in MT la migliore corrispondenza, cioè una coppia (s, t_q) la cui somiglianza $Sim(s, s_q) \geq \sigma$, se esiste, è massima (Sikes 2007). La funzione Sim misura la somiglianza tra due stringhe nella lingua source. Essa produce un valore percentuale, dove 100% significa "stringhe identiche" (corrispondenza esatta). I valori percentuali intermedi sono chiamati corrispondenze fuzzy. Il SMT classifica le stringhe in base alla loro percentuale di similarità e presenta al traduttore le relative traduzioni nello stesso ordine. In genere, quando non c'è una corrispondenza esatta, il traduttore può comunque scegliere una delle traduzioni proposte e modificarla come meglio crede: la nuova traduzione prodotta andrà quindi ad aggiungersi alle altre nella MT.

3.1.1 Memoria di Traduzione (MT) e recupero dei suggerimenti

La MT è organizzata al livello di stringa, che da qui in poi chiameremo segmento, che costituisce una porzione di testo da tradurre di lunghezza arbitraria. La MT è definita come un insieme di quadruple $\{s_i, T_i, A_i, c_i\}$ con $i = 1..n$ dove ogni quadrupla è definita come:

- s_i, il segmento in lingua originale;

- $T_i = \{t_i^1, ..., t_i^k\}$, l'insieme di traduzioni di s_i con $k \geq 1$, dove ogni t_i^j include una parte letterale che corrisponde al segmento source e una parte esplicativa chiamata informazione contestuale;
- $A_i = \{a_i^1, ..., a_i^k\}$, l'insieme degli identificatori dei traduttori di ogni segmento tradotto s_i in T_i con $k \geq 1$;
- c_i , il contesto di s_i che si riferisce al trattato al quale s_i appartiene.

La maggior parte dei SMT si basa su varianti della distanza di Levenshtein normalizzata sulla lunghezza del segmento di interrogazione, ossia il numero minimo di operazioni di modifica necessarie per trasformare un segmento in un altro. Nel caso del TB, non sono stati considerati gli aspetti linguistici dei segmenti, dal momento che nessuno degli strumenti di trattamento automatico del linguaggio disponibili risulta adatto per l'analisi di lingue semitiche antiche, come ad esempio i diversi idiomi ebraico e aramaico attestati nel TB (vedere Sezione 4). E' stato scelto di adottare una misura di similarità Sim basata sull'edit distance, $ED(s_1, s_2)$, considerando due segmenti tanto più simili quando gli stessi termini tendono a comparire nello stesso ordine. Dato un segmento s_q da tradurre, la seguente formula misura la somiglianza, in percentuale, tra due segmenti s_q e s[22] :

$$Sim(s, s_q) = (1 - min(1, [ED(s, s_q) \ / \ |s_q|])) * 100 \quad (1)$$

L'algoritmo per il calcolo di $ED(s_1, s_2)$ è basato sulla programmazione dinamica, e l'implementazione adottata fa riferimento a (Navarro 2001). In breve, viene creata una matrice $M(0..|s_1|, 0..|s_2|)$, dove ogni elemento $m_{i,j}$ rappresenta il numero minimo di cambiamenti di parole necessari per trasformare $s_1(1..i)$ in $s_2(1..j)$. Il processo di calcolo è il seguente:

$$m_{i,j} = \begin{cases} i \text{ if } j = 0 \\ j \text{ if } i = 0 \\ m_{(i-1,j-1)} \text{ if } s_1(i) = s_2(j) \\ 1 + \mu \text{ if } s_1(i) \neq s_2(j) \end{cases}$$

dove $\mu = min(m_{(i-1,j)}, m_{(i,j-1)}, m_{(i-1,j-1)})$, e il costo finale è rappresentato da $m_{(|s_1|,|s_2|)}$. La Figura 4 mostra un esempio di calcolo di $ED(s_1, s_2)$. Grazie alle soluzioni tecniche descritte nelle sezioni precedenti, il SMT può recuperare e presentare le proposte di traduzione in pochi millisecondi, tra centinaia di migliaia di traduzioni. L'interfaccia utente di *Traduco* presenta al traduttore ogni suggerimento accompagnato da un numero di stelle, come appare in Figura 5. Il numero viene assegnato sulla base del grado di similarità del match tra segmenti d'origine: suggerimenti a cinque stelle sono considerati perfetti (corrispondenza esatta, Sim = 100%); quattro stelle indicano che alcune correzioni sono probabilmente necessarie (85% $\leq Sim \leq$ 99%); tre stelle indicano, nella maggior parte dei casi, i suggerimenti accettabili (70% $\leq Sim \leq$ 84%). Un traduttore può scegliere di filtrare i suggerimenti proposti in base all'autore, se la traduzione è già stata revisionata, o se il contesto della traduzione è lo stesso. Natu-

22 In accordo con $Sim(s, s_q) \geq \sigma$, presentata nella sezione precedente, abbiamo definito $\sigma = 0.7$ in maniera sperimentale con il gruppo di traduttori.

ralmente, ogni nuova traduzione viene aggiunta alla MT, aumentando così il *pool* di traduzioni disponibili.

		מקבלין	המכירין	התקינו	שלא	יהו	מקבלין	אלא	מן	המכירין	אותה
	0	1	2	3	4	5	6	7	8	9	10
באותה	1	**1**	2	3	4	5	6	7	8	9	10
שעה	2	2	**2**	3	4	5	6	7	8	9	10
התקינו	3	3	3	**2**	3	4	5	6	7	8	9
שלא	4	4	4	3	**2**	3	4	5	6	7	8
יהו	5	5	5	4	3	**2**	3	4	5	6	7
מקבלין	6	6	6	5	4	3	**2**	3	4	5	6
אלא	7	7	7	6	5	4	3	**2**	3	4	5
מן	8	8	8	7	6	5	4	3	**2**	3	4
המכירין	9	9	9	8	7	6	5	4	3	**2**	**3**

Figura 4
Esempio di calcolo di $ED(s_1, s_2)$, con $m(|s_1|, |s_2|) = 3$.

Una delle caratteristiche interessanti che sono state introdotte è il modo in cui sono classificati i suggerimenti. In particolare, vengono considerati i) gli autori delle traduzioni, cioè gli insiemi A_i e ii) i contesti c_i di riferimento di ogni segmento. Tali informazioni possono rivelarsi utili sia per i traduttori che per i revisori. Da un lato, i traduttori possono valutare l'affidabilità delle traduzioni suggerite sulla base dell'autorevolezza e delle competenze dei relativi traduttori. D'altra parte, i revisori possono sfruttare entrambi i tipi di informazioni per garantire una traduzione più omogenea e scorrevole.

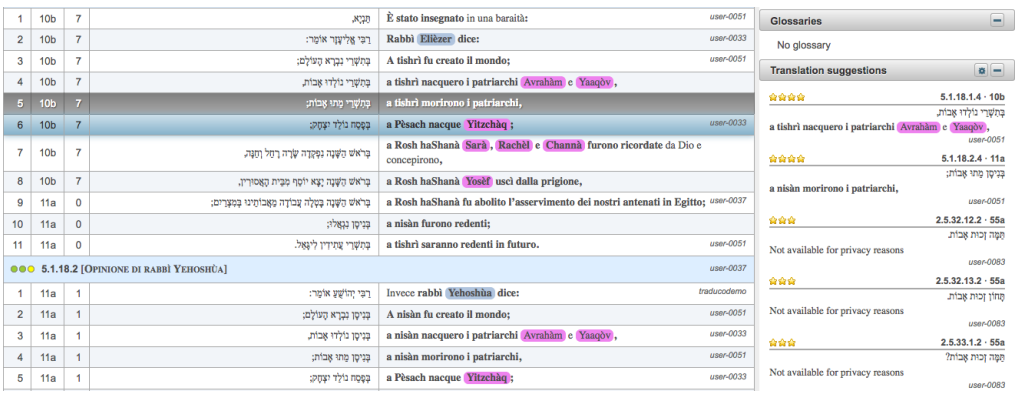


Figura 5
Classificazione dei suggerimenti alla traduzione.

3.1.2 Analisi della ridondanza della MT

La valutazione delle prestazioni di un sistema come *Traduco* non è un compito banale. A differenza di un tipico STA, l'obiettivo di *Traduco* non è solamente quello di incrementare il ritmo di traduzione. Esso, infatti, ha il compito di sostenere l'intero processo interpretativo-traduttivo, offrendo un ambiente collaborativo in cui gli utenti possono tradurre le proprie porzioni di testo sfruttando le traduzioni di segmenti simili, che,

d'altra parte, potrebbero variare molto in termini di informazioni contestuali, note esplicative e riferimenti bibliografici. Nella sezione 4 sono descritte, in maggiore dettaglio, alcune delle criticità riscontrate nel valutare le prestazioni di Traduco. E' stata analizzata

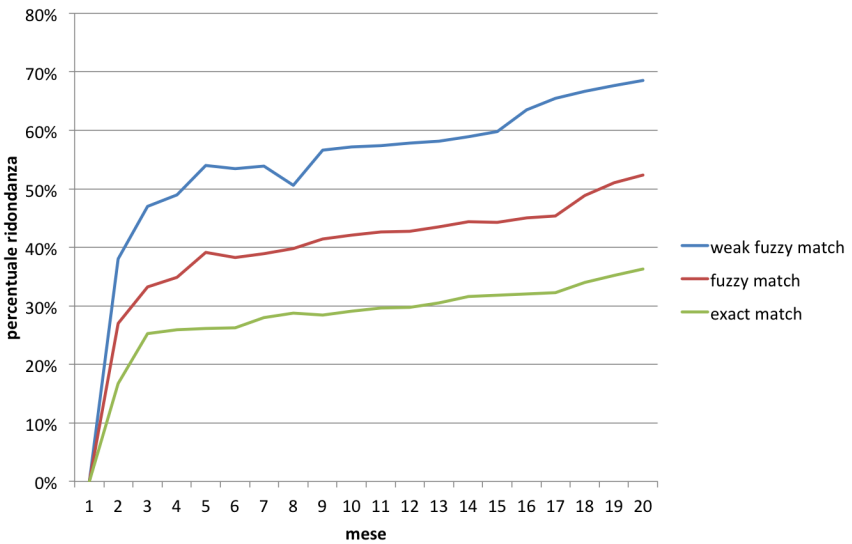


Figura 6
Ridondanza della MT, calcolata nei primi 20 mesi del PTTB.

la ridondanza della MT considerando i segmenti simili tramite una tecnica statistica di ricampionamento (Wu 1986). La MT è stata partizionata in gruppi di 1000 segmenti, ognuno dei quali è stato poi utilizzato come insieme di test per la traduzione dei propri segmenti con il resto della MT. Intuitivamente, questo metodo di ricampionamento utilizza la variabilità negli insiemi di test per trarre conclusioni circa la significatività statistica.

La Figura 6 mostra il tasso di ridondanza della MT nei primi 20 mesi di lavoro nel PTTB. Le curve di ridondanza sono disegnate considerando i 3 range di score della funzione di similarità presentati nella sezione precedente. La percentuale di segmenti trovati nella MT cresce logaritmicamente con il tempo (e di conseguenza con la dimensione della memoria): maggiore è la dimensione della memoria, migliore sarà la performance del SMT. Ad esempio, supponendo che le corrispondenze esatte forniscano almeno un suggerimento perfetto, al netto delle informazioni contestuali ogni traduttore risparmierebbe quasi il 40% del proprio tempo di lavoro.

3.2 Note e Annotazioni

Come anticipato nella sezione 3, una traduzione che vuole essere comprensibile ai non esperti può richiedere l'inserimento di integrazioni esplicative (che costituiscono interpretazioni esse stesse). Inoltre, il testo tradotto può richiedere l'aggiunta di annotazioni a vari livelli. *Traduco* permette ai traduttori di i) inserire note di diverso tipo (note generiche, note di revisione, note editoriali, ecc.) come mostrato in Figura 7(a), ii) annotare semanticamente porzioni arbitrarie del testo (Figura 7(b)), e iii) inserire riferimenti bibliografici parzialmente precompilati (ad esempio per le citazioni bibliche da completare con i numeri di capitolo e versetto) e i nomi dei rabbini. Più in dettaglio,

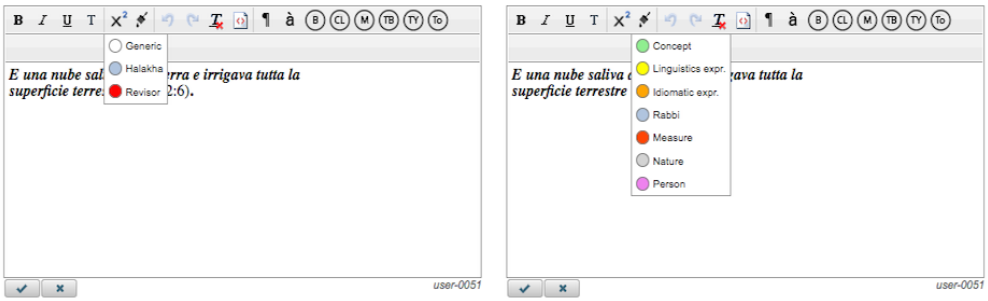


Figura 7
L'editor per l'inserimento delle traduzioni. (a) esempio di inserimento di note; (b) esempio di annotazione semantica.

per la traduzione del TB *Traduco* permette ai traduttori di annotare il testo tradotto (italiano) sulla base di alcune classi semantiche predefinite, tra cui figurano: nomi di persona, nomi di rabbini, entità legate alla natura (piante, animali, ecc.), unità di misura, espressioni idiomatiche talmudiche (ad esempio, "*tanya*", "è stato insegnato") e concetti talmudici (come "*Shemà*"). Ad oggi, il sistema non consente l'annotazione di porzioni di testo discontinue. Ogni annotazione semantica può essere accompagnata da una descrizione libera testuale (vedere il pannello glossari a destra della Figura 2(c)), una traslitterazione opzionale e una forma originale in ebraico. Una nuova annotazione può essere associata ad una forma canonica facendo riferimento a una voce di glossario pre-esistente. Un pannello dedicato del sistema (Figura 8) permette il *browsing* e l'interrogazione di tutte le voci, ognuna delle quali è indicata con la descrizione (glossa) e tutte le occorrenze nel testo. La disponibilità di un numero sufficientemente elevato di annotazioni apre la strada alla automatizzazione del processo di annotazione. Sfruttando l'analisi linguistica automatica del testo tradotto sarà in futuro implementato un algoritmo stocastico supervisionato, addestrato sulle annotazioni a disposizione, per suggerire nuove annotazioni che gli utenti potranno accettare o rifiutare. Inoltre, come descritto nella sezione 5, il repertorio terminologico costituito dalle varie voci del glossario potrà essere sfruttato per la costruzione di una base di conoscenza utilizzando un linguaggio formale di rappresentazione. Ad oggi[23], sono state inserite 3431 voci di glossario per un totale di 58188 annotazioni nel testo, 28338 delle quali già associate ad una specifica voce e, le altre, ancora da associare da utenti revisori preposti a tale compito. Sono altresì presenti 14263 note generiche e 2534 note di *Halakha* (Legge ebraica).

3.3 Supervisione del Processo di Traduzione

Traduco permette ai supervisori di coordinare diversi tipi di utenti riuniti in un team, che può essere composto da traduttori, revisori, redattori, gli stessi supervisori e altri profili utente. Ogni utente può avere uno o più ruoli e diritti specifici sulle funzionalità offerte dal Sistema. Inoltre, ogni utente può essere associato a una o più risorse, intese come specifiche porzioni del testo da tradurre (ad esempio, uno specifico capitolo). Ogni

23 28 novembre 2016

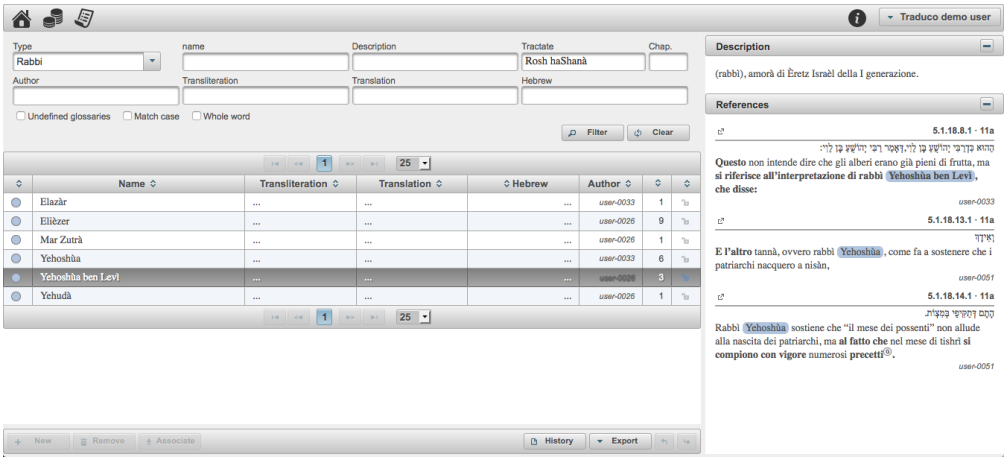


Figura 8
Il pannello dei glossari.

risorsa (nel caso del TB, un trattato, un capitolo, un blocco o una unità logica) è dotata di uno stato di avanzamento rappresentato come una sorta di semaforo (Figura 9): il rosso indica che la risorsa non è stata ancora assegnata, il giallo che è in fase di lavorazione e il verde che è completata per la specifica fase.

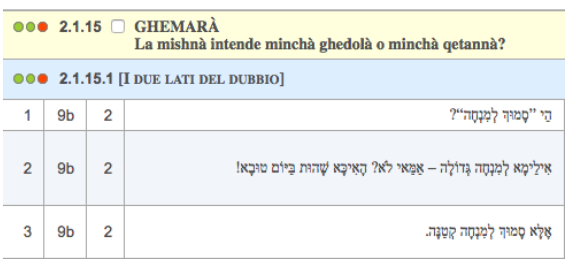


Figura 9
Indicatori del progresso del processo di traduzione; in questo esempio, l'unità logica con id 2.1.15.1 è stata tradotta, revisionata e redatta per l'edizione a stampa, mentre per il blocco a cui essa appartiene (2.1.15) non è ancora conclusa la fase di redazione (il relativo cerchietto indicatore è giallo).

Da sinistra a destra, i tre cerchietti rappresentano lo stato della traduzione, della revisione e della redazione. Il cambiamento di stato è consentito agli utenti in base ai loro diritti. I traduttori, per esempio, possono modificare soltanto lo stato della fase di traduzione (il cerchietto più a sinistra) e limitatamente alle risorse a cui sono stati assegnati. L'intero flusso della traduzione può essere controllato attraverso un pannello dedicato (Figura 10). La Figura 10(a), per esempio, mostra lo stato di avanzamento di una risorsa monitorando la produttività di tutti i tipi di utente. Per mezzo di una rappresentazione grafica è possibile esaminare il progresso generale (traduzione, revisione e redazione) della risorsa selezionata (Figura 10(b)).

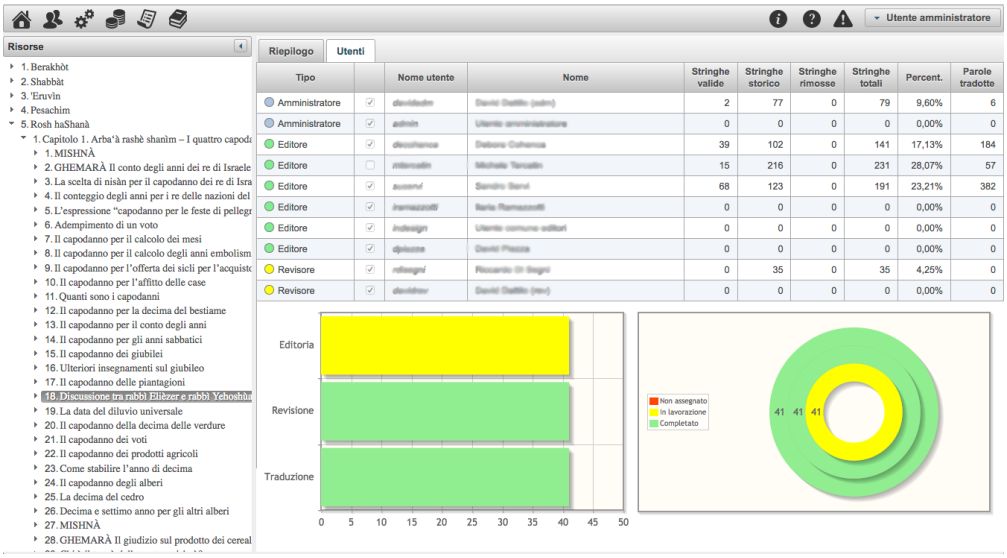


Figura 10
Il pannello di supervisione (i nomi utente sono oscurati per ragioni di privacy). (a) dati analitici per ogni utente. (b) sintesi di dati in forma grafica.

3.4 Aspetti di Editoria

In aggiunta alle funzionalità per la traduzione, l'annotazione, la creazione di glossari e la gestione complessiva del flusso di lavoro, *Traduco* comprende una serie di strumenti progettati per esportare i testi tradotti in diversi formati. Per la traduzione del TB, il sistema consente agli utenti di esportare una risorsa selezionata (da una singola unità logica a un intero trattato) in formato pdf e in quattro layout differenti: i) con note e voci di glossario in calce ad ogni pagina nella quale compaiono, ii) con tutte le note e voci di glossario raccolte alla fine di ogni folio, iii) con le stringhe raggruppate per paragrafo (per fini editoriali) e, infine, iv) dove le stringhe sono omesse ma compaiono soltanto le voci di glossario presenti nella risorsa. Inoltre, per supportare la produzione dell'edizione a stampa, consente di esportare la traduzione, le note e i glossari in Adobe InDesign XML. E' attualmente in fase di sviluppo anche un esportatore in formato TEI. Per *Traduco*, abbiamo scelto di adottare Teibook[24], uno schema conforme alle linee guida TEI P5[25] volto a rappresentare la struttura dei libri stampati. Teibook è stato progettato in riferimento ad HTML 5 (per garantire la conservazione della semantica HTML) ed è in grado di codificare tutte le informazioni generate dall'elaborazione di testi allo scopo di produrre editoria elettronica di qualità (ePub TOC, sito on-line), consentendo, nel contempo, successivi arricchimenti semantici.

24 `http://obvil.paris-sorbonne.fr/developpements/teibook` (ultimo accesso: 28 novembre 2016)
25 TEI Consortium, eds. Guidelines for Electronic Text Encoding and Interchange. `http://www.tei-c.org/P5/` (ultimo accesso: 28 novembre 2016)

4. Integrazione di Tecnologie del Linguaggio

In letteratura sono documentati due tipi di approcci volti a massimizzare la funzione Sim: quelli che si basano sulle misure di confidenza della Machine Translation (He et al. 2010b), (He et al. 2010a), (Smith and Clark 2009), (Koehn and Senellart 2010), (Kun, Chengqing, and Keh-Yih 2014), (Dong et al. 2014) e quelle che vanno al di là del semplice confronto delle forme superficiali introducendo informazioni linguistiche e parafrasi (Planas and Furuse 1999), (Pekar and Mitkov 2007), (Utiyama et al. 2011), (Gupta and Orasan 2014), (Ganitkevitch, Van Durme, and Callison-Burch 2013), (Hodasz and Pohl 2005).

Ad ogni modo, come anticipato, esistono situazioni in cui non sono disponibili né strumenti né risorse linguistiche per il TAL. Nel caso del TB, ad esempio, se si escludono alcuni strumenti per l'analisi dell'ebraico moderno, come MILA (Itai 2006) e HebMorph[26], non esiste alcuno strumento che possa essere di supporto per l'analisi dell'Ebraico Antico e dell'Aramaico. Per ovviare (almeno in parte) a questa mancanza, abbiamo ideato una tecnica basata sulla semantica distribuzionale (SD) allo scopo di migliorare la qualità dei suggerimenti: tutte le parole che condividono contesti simili nel testo da tradurre vengono recuperate e utilizzate per arricchire la MT. Il peso dell'operazione di sostituzione viene quindi modificato nel calcolo della funzione $ED(s_1, s_2)$: quanto più i vettori che rappresentano le parole da esaminare sono vicini nello spazio distribuzionale, minore è il peso che deve essere assegnato alla sostituzione. Abbiamo sperimentato questa tecnica su alcuni trattati del TB. Il primo esperimento che abbiamo condotto è costituito da sei passaggi:

1. è stato applicato un algoritmo di Semantica Distribuzionale (Baroni and Lenci 2010) su una porzione del TB in modo da costruire la matrice parola–contesto;
2. è stata creata una lista di coppie composte da parole correlate, RW_{Talmud};
3. sono stati raccolti tutti i diversi segmenti source tradotti con la stessa frase italiana (ottenendo 313 corrispondenze);
4. sono stati selezionati i segmenti source aventi la stessa lunghezza[27], per facilitare il processo di valutazione manuale;
5. per ogni corrispondenza, sono state identificate le coppie di parole $(w_i; w_j)$ per le quali i segmenti source sono differenti;
6. per ognuna delle coppie *c* in $(w_i; w_j)$, abbiamo cercato il rispettivo valore di coseno in RW_{Talmud};
7. i tipi di relazione che coinvolgono coppie di parole recuperate dall'algoritmo sono state riportate in Tabella 2.

Prima di condurre una valutazione empirica dell'approccio, è possibile osservare come la misura incrementata attraverso la SD tenda naturalmente a ridurre il costo della funzione ED() (migliorando la funzione Sim) tra due segmenti source, in proporzione

26 Un analizzatore e disambiguatore morfologico per l'Ebraico. `http://code972.com/hebmorph` (ultimo accesso: 28 novembre 2016)

27 Notare che tale selezione riduce drasticamente i casi possibili di confronto calcolati dalla funzione ED().

Tabella 2
Tipi di relazioni recuperate dall'algoritmo di SD.

Tipo di Relazione	#Occ.	Esempi
Interlinguistica	23	Parole che hanno lo stesso significato ma sono scritte in lingue diverse, tipicamente Ebraico e Aramaico, ad esempio la parola "*adatti*"(כְּשֵׁרִין e כְּשֵׁרִים)
Ortografica	5	Parole che si presentano con grafie diverse, ad esempio la parola "*campanelli*" (בַּזְוִגִין e בַּזְוִגִין)
Morfologica	21	Parole che hanno la stessa radice ma flessioni diverse, ad esempio il verbo "*uscire*", coniugato come "*esce*" (יוֹצֵא) e come "*uscirà*" (יֵצֵא)
Sintattica	62	Parole che presentano o meno articoli e congiunzioni, data la natura delle lingue semitiche coinvolte, ad esempio "*posto*" (מָקוֹם) e "*il posto*" (הַמָּקוֹם) e parole che compaiono in frasi distinte con diverso ordine, ad esempio רַב פָּפָּא אָמַר e אָמַר רַב פָּפָּא che significano rispettivamente "*Rav Yochanan disse*" e "*disse Rav Yochanan*"
Lessico-Semantica	23	Parole che presentano una relazione di sinonimia e quasi-sinonimia, quali "*mettere*" (הִנִּיחַ) e "*porre*" (נָתַן) o "*per conto di*" (מִשּׁוּם) e "*a nome di*" (מִשְּׁמֵיהּ)

alla similarità delle parole diverse. Quindi, la percentuale di corrispondenze non esatte può essere incrementata, e, conseguentemente: i) alcuni suggerimenti alla traduzione possono salire nel ranking della SMT e ii) possono venire proposti nuovi suggerimenti che altrimenti non sarebbero stati recuperati.

Questa tecnica eredita i limiti degli approcci basati sulla Semantica Distribuzionale, tra i quali spicca l'esigenza di avere un corpus sufficientemente grande affinché le relazioni tra le parole che condividono contesti simili possano essere considerate statisticamente significative. Nel caso dell'esperimento qui descritto il corpus selezionato risulta sufficientemente grande per fornire risultati soddisfacenti, per quanto sia costituito da una porzione relativamente piccola del TB. Ad ogni modo, dato che l'approccio in esame è non supervisionato e linguisticamente agnostico, può essere applicato ad altri linguaggi source e ad altri testi. Conseguentemente, può rivelarsi utile in quelle traduzioni che coinvolgono lingue non supportate da risorse per il TAL, dato che, come riportato in Tabella 2, permette anche di recuperare segmenti che differiscono per varianti morfologiche o sintattiche.

La valutazione dei miglioramenti ottenuti attraverso l'introduzione delle misure incrementate utilizzando la SD può risultare molto complicata in questo contesto. In primo luogo, i corpora tipicamente utilizzati in letteratura per valutare le performance di un SMT, sono molto diversi dal BT (specialmente in termini di varietà lessicale). Le valutazioni basate sul postediting e sul numero di battute non sembrano essere appropriate per il nostro contesto. Infatti, la traduzione di un nuovo segmento che è stato già (anche in parte) tradotto, può essere significativamente diversa e richiedere, ad esempio, un numero molto maggiore di parole nel segmento target per spiegare la frase e renderla chiara al lettore contemporaneo, al quale vanno fornite le informazioni di contesto necessarie alla comprensione della frase stessa. Il TB mostra inoltre un grado elevato di variabilità nella traduzione di segmenti source identici, specie rispetto

ad altri corpora. Segmenti simili possono infatti richiedere tempi di traduzione molto diversi, ad esempio nel caso in cui un segmento contenga parole insolite che devono essere appropriatamente illustrate utilizzando un'annotazione. Per ulteriori dettagli relativamente all'esperimento qui descritto si veda (Bellandi et al. 2016).

5. Conclusioni

In questo articolo abbiamo introdotto *Traduco*, un ambiente web collaborativo sviluppato per supportare la traduzione di testi che presentano significativi problemi interpretativi. Rispetto a STA esistenti, *Traduco*, ad oggi, presenta alcune peculiarità che consentono di: i) creare un team di utenti, con diversi ruoli, e consentire loro di lavorare, contemporaneamente, allo stesso testo; ii) personalizzare i suggerimenti proposti, scegliendo di visualizzare solo le proprie traduzioni, o le traduzioni appartenenti a specifici contesti; iii) arricchire la traduzione con informazione esplicativa contestuale, distinguendola dalla traduzione più letterale; iv) annotare semanticamente porzioni di testo per creare glossari tematici; v) produrre viste ed esportazioni volte ad agevolare i processi editoriali. Da un punto di vista più generale, considerando le caratteristiche di cui sopra, *Traduco* si propone come strumento per la traduzione di testi umanistici che per loro stessa natura sono intrinsecamente diversi da testi legislativi o manuali tecnici, essendo più complessi e richiedendo processi ermeneutici particolarmente impegnativi.

Nelle prossime fasi del progetto saranno introdotte tecniche per la gestione del contenuto del testo. L'analisi dei risultati provenienti dall'annotazione semantica del TB, infatti, ha suggerito l'idea di fornire una strutturazione formale dei concetti relativi ai termini e alle entità annotati. Una formalizzazione del contenuto del testo (per esempio modellata con un'ontologia) consentirà di interrogare lo stesso sia su base semantica che su base linguistica, superando i limiti imposti dalle semplici ricerche effettuate per parola chiave. Alcuni esperimenti sono già stati documentati in (Bellandi et al. 2014).

Sebbene il Sistema sia stato implementato in un contesto di ricerca, *Traduco* è solido e flessibile abbastanza da poter essere facilmente personalizzato per la traduzione di altri testi e in altre lingue. *Traduco* è attualmente utilizzato dalla comunità del PTTB composta da oltre 80 utenti (tra traduttori, revisori, redattori e supervisori), che ne testano quotidianamente le funzionalità, l'efficienza e le performance. Una versione demo di Traduco (corredata da una guida utente nella quale sono descritti casi d'uso esemplificativi) può essere testata al seguente url: `http://talmud-dev.ilc.cnr.it:8082/talmud/`, nome utente e password: *traducodemo*.

Acknowledgments

Questo lavoro è stato condotto nel contesto del progetto di ricerca TALMUD tramite la partnership scientifica tra S.c.a r.l. "Progetto Traduzione del Talmud Babilonese" (PTTB) e ILC-CNR sulla base del "Protocollo d'Intesa" tra la Presidenza Italiana del Consiglio dei Ministri, il Ministero dell'Educazione, dell'Università e della Ricerca, l'Unione delle Comunità Ebraiche Italiane, Il Collegio Rabbinico Italiano e il Consiglio Nazionale delle Ricerche (21/01/2011).

Bibliografia

Baroni, Marco and Alessandro Lenci. 2010. Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics*, 36(4):673–721.

Bellandi, Andrea, Alessia Bellusci, Enrico Carniani, and Emiliano Giovannetti. 2014. Content elicitation: Towards a new paradigm for the analysis and interpretation of texts. In *Proceedings of the 13th IASTED International Conference on Software Engineering*, pages 17–19, Innsbruck.

Bellandi, Andrea, Giulia Benotto, Gianfranco Di Segni, and Emiliano Giovannetti. 2016. Investigating the application and evaluation of distributional semantics in the translation of

humanistic texts: a case study. In *Proceedings of the 2nd Workshop on Natural Language Processing for Translation Memories (NLP4TM 2016)*, pages 6–11, Portorose (Slovenia).
Dong, Meiping, Yong Cheng, Yang Liu, Jia Xu, Maosong Sun, Tatsuya Izuha, and Jie Hao. 2014. Query lattice for translation retrieval. In *Proceedings of the 25th International Conference on Computational Linguistics (COLING 2014)*, pages 2031–2041.
Ganitkevitch, Juri, Benjamin Van Durme, and Chris Callison-Burch. 2013. Ppdb: The paraphrase database. In *Proceedings of the NAACL-HLT*, pages 758–764, Atlanta, Georgia.
Gupta, Rohit and Constantin Orasan. 2014. Incorporating paraphrasing in translation memory matching and retrieval. In *Proceedings of the 17th Annual Conference of European Association for Machine Translation*, pages 3–10, Atlanta, Georgia.
He, Yifan, Yanjun Ma, Josef van Genabith, and Andy Way. 2010a. Bridging smt and tm with translation recommendation. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 622–630, Uppsala, Sweden.
He, Yifan, Yanjun Ma, Andy Way, and Josef van Genabith. 2010b. Integrating n-best smt outputs into a tm system. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, COLING '10, pages 374–382.
Hodasz, Gabor and Gabor Pohl. 2005. Metamorpho tm: A linguistically enriched translation memory. In *Proceedings of the International Workshop Modern Approaches in Translation Technologies*, pages 26–30, Borovets, Bulgaria.
Itai, Alon. 2006. Knowledge center for processing hebrew. In *Proceedings of the 5th International Conference on Language Resources and Evaluation - Workshop "Towards a Research Infrastructure for Language Resources"*, Genoa, Italy.
Koehn, Philipp and Jean Senellart. 2010. Convergence of translation memory and statistical machine translation. In *Proceedings of the AMTA Workshop on MT Research and the Translation Industry*, pages 21–31, Denver, Colorado.
Kun, Wang, Zong Chengqing, and Su Keh-Yih. 2014. Dynamically integrating cross-domain translation memory into phrase-based machine translation during decoding. In *Proceedings of the 25th International Conference on Computational Linguistics (COLING 2014)*, pages 398–408, Dublin, Ireland.
Lagoudaki, Elina. 2009. Translation editing environments. In *In MT Summit XII: Workshop on Beyond Translation Memories*.
Navarro, Gonzalo. 2001. A guided tour to approximate string matching. *ACM Comput. Surv.*, 33(1):31–88, March.
Patil, Manish, Sharma V. Thankachan, Rahul Shah, Wing-Kai Hon, Jeffrey Scott Vitter, and Sabrina Chandrasekaran. 2011. Inverted indexes for phrases and strings. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '11, pages 555–564, New York, NY, USA. ACM.
Pekar, Viktor and Ruslan Mitkov. 2007. New generation translation memory: Content-sensivite matching. In *Proceedings of the 40th anniversary congress of the swiss association of translators, terminologists and interpreters*.
Planas, Emmanuel and Osamu Furuse. 1999. Formalizing translation memories. In *Proceedings of Machine Translation Summit VII*, pages 331–339, Singapore.
Reinke, Uwe. 2006. Translation memories. *Encyclopedia of Language and Linguistics*, pages 61–65.
Reinke, Uwe. 2013. State of the art in translation memory technology. *Translation: Computation, Corpora, Cognition*, 3(1):27–48.
Sikes, Richard. 2007. Fuzzy matching in theory and practice. *Multilingual*, 18(6):39–44.
Smith, James and Stephen Clark. 2009. Ebmt for smt: a new ebmt-smt hybrid. In *Proceedings of the 3rd International Workshop on Example-Based Machine Translation (EBMT 2009)*, pages 3–10, Dublin, Ireland.
Somers, Harold L. 2003. Translation memory systems. *Computers and translation: A translator's guide*, 35:31–48.
Utiyama, Masao, Graham Neubig, Takashi Onishi, and Eiichiro Sumita. 2011. Searching translation memories for paraphrases. In *Proceedings of the Machine Translation Summit XIII*, pages 325–331, Xiamen, China.
Wu, C.F.J. 1986. Jackknife, bootstrap and other resampling methods in regression analysis. *The Annals of Statistics*, 14(4):1261–1295.