Live chat e natural language processing *in sinergia per il miglioramento dei servizi bibliotecari*

MARIA VITTORIA MUZZUPAPA
Ufficio Servizi bibliografici digitali
Università degli studi di Torino
mariavittoria.muzzupapa@unito.it

MARCO STEFANO TOMATIS
Ufficio Servizi bibliografici digitali
Università degli studi di Torino
marcostefano.tomatis@unito.it

FRANCO CARLO BUNGARO
Area Servizi bibliotecari di ateneo
Università degli studi di Torino
franco.bungaro@unito.it

DOI: 10.3302/2421-3810-201701-041-1

Virtual Reference: il contesto operativo

Lo studio che presentiamo si inserisce in un contesto di ricerca, quello sul reference virtuale, relativamente giovane poiché la sua storia comincia circa trent'anni fa.
Per dare un riferimento più preciso, segnaliamo che nel 2006 B. Sloan pubblica *Twenty years of virtual reference* in cui, riferendosi al report relativo a un studio del 1986 sull'uso dell'e-mail dell'Health Sciences Center come strumento di reference[1], ricorda quello che egli stesso definisce essere una pietra miliare: «[…] what I believe may be the first journal article devoted to virtual reference»[2]. A nostro avviso, l'altro passaggio storico significativo di questo ambito disciplinare si verifica intorno alla fine degli anni Novanta, quando il reference virtuale vive una svolta decidendo di usare le *live chat* come mezzo di comunicazione con i propri utenti.
Si amplia a questo punto il concetto di *virtual* (o *digital*) *reference*, che acquisisce due nuovi attributi: sincrono e asincrono. Taher descrive in maniera molto semplice e chiara la differenza tra le due tipologie di assistenza, usando come punto di vista preferenziale l'intervista all'utente:

> The reference interview in cyberspace can be defined as a real-time or non-real-time, Web-based-text, and sometimes voice and/or video-24 communication

Per tutti i siti web l'ultima consultazione è stata effettuata il 13 maggio 2017.

[1] In modo particolare Sloan cita il lavoro di Ellen H. Howard - Terry A. Jankowski, *Reference services via electronic mail*, «Bulletin of the Medical Library Association», 74 (1986), n. 1, p. 41-44, <http://pubmedcentralcanada.ca/pmcc/articles/PMC227777/pdf/mlab00053-0061.pdf>.
[2] Bernie Sloan, *Twenty years of virtual reference*, «Internet reference services quarterly», 11 (2006), n. 2, p. 91-95: 91.

> between the reference interviewer and interviewee. Real-time reference is synchronous interaction, or an instant transaction between the interviewer and the interviewee. Non-real-time reference is a delayed interaction, which is a follow-up of a reference inquiry received by e-mail or Web forms[3].

L'assistenza online e in tempo reale ha posto immediatamente i bibliotecari di fronte alle difficoltà di condurre un'intervista non in compresenza con l'utente. La comunità si è dovuta orientare tra nuove modalità di comunicazione e interazione. In modo particolare negli Stati Uniti, dove Lankes nel 2004 testimonia una fervente attività attorno al *digital reference*[4], si moltiplicano gli studi relativi alla gestione delle interviste all'utenza e le abilità necessarie ai bibliotecari per poter affrontare il nuovo mezzo di assistenza. Nel 2002 l'IFLA pubblica la prima edizione delle linee guida per il reference digitale[5].

Il reference digitale sincrono comincia a essere esplorato nelle nuove possibilità che offre: nel 2003 Marsteller e Mizzy studiano le trascrizioni delle conversazioni del primo anno completo di attività della chat della Carnegie Mellon University. In particolare analizzano le differenti tipologie delle richieste degli utenti, delle domande dei bibliotecari e delle risposte degli utenti. Si scontrano però con una carenza di strumenti disponibili per questo tipo di lavoro: la letteratura offre ancora pochi studi sui metodi di analisi, soprattutto quelli elaborati non risultano standardizzati:

> As indicated in the literary review, the authors explored using schema from existing reference interaction research, but they found no standardization. The authors therefore decided to craft their own schema, or set of categories, that attempted to capture the nuances of digital reference[6].

Quindi i due autori sviluppano uno schema che categorizzi le diverse tipologie di domande e risposte ottenute dalle trascrizioni. I dati ricavati dalla loro analisi vengono registrati in un foglio di calcolo e analizzati con l'aiuto degli statistici.

Nello stesso anno, sempre alla Carnegie Mellon, Gerlich e Berard[7] lanciano la sperimentazione della *READ Scale*, una scala divisa in sei livelli di competenza richiesta al bibliotecario per fornire il servizio di reference appropriato alla domanda ricevuta.

Un anno più tardi Hirko e Ross propongono un accurato manuale sulla formazione per i bibliotecari di reference virtuale, che fornisce appendici con esercizi, esempi di transazioni tra bibliotecari e utenti, strumenti per migliorare l'utilizzo di hardware, software e glossari[8]. Sempre nel 2004 la Reference User and Service Association (RUSA) divulga le *Guidelines for implementing and maintaining virtual reference services*[9].

Per diversi anni gli studi sul reference digitale sincrono si concentrano sulle competenze da sviluppare da parte dei bibliotecari e per comprendere i comportamenti degli utenti. Ogni ricerca elabora una propria procedura di analisi, ma i procedimenti sono difficilmente ripetibili perché basati essenzialmente su un'iniziale lettura ed etichettatura delle conversazioni pressoché manuali, per poi riportare i dati rilevati in software che li rielaborino e li valutino da un punto di vista statistico.

Presso la Biblioteca di Giurisprudenza dell'Università di Georgetown, Morais e Sampson producono un'analisi dei contenuti delle conversazioni avvenute tramite chat: osservano i comportamenti tenuti dagli utenti e dai bibliotecari per comprendere le tipologie di domande dei primi e quindi classificare i loro bisogni informativi. In questo caso gli autori hanno stampato le conversazioni e uno staff selezionato si è occupato di classificare le chat per tipologie secondo criteri prestabiliti[10].

Il reference digitale, dove le *live chat* acquisiscono un peso sempre maggiore, costituisce un bacino estre-

[3] Mohamed Taher, *The Reference Interview Through Asynchronous E-Mail and Synchronous Interactive Reference: Does It Save the Time of the Interviewee?*, «Internet reference services quarterly», 7 (2002), n. 3, p. 23-34: 24-25.

[4] Cfr. R. David Lankes, *The digital reference research agenda*, «Journal of the American Society for Information Science and Technology», 55 (2004), n. 4, p. 301-311, <https://davidlankes.org/rdlankes/Publications/Journals/jasisdr.pdf>.

[5] Cfr. IFLA, *Digital Reference Guidelines*, 2002, <https://www.ifla.org/publications/ifla-digital-reference-guidelines>.

[6] Matthew r. Marsteller - Denianna Mizzy, *Exploring the Synchronous Digital Reference Interaction for Query Types, Question Negotiation,and Patron Response*, «Internet reference services quarterly», 8 (2003), n. 1-2, p. 149-165: 153-154.

[7] Cfr. Bella K. Gerlich - G. Lynn Berard, *Introducing the READ Scale: qualitative statistics for academic reference services*, «Georgia Library quarterly», 43 (2007), n. 4, p. 7-13, <http://digitalcommons.kennesaw.edu/glq/vol43/iss4/4>.

[8] Cfr. Buff Hirko - Mary B. Ross, *Virtual reference training: the complete guide to providing anytime. anywhere answers*, Chicago, American Library Association, 2004, p. 160.

[9] Cfr. RUSA, *Guidelines for Implementing and Maintaining Virtual Reference Services*, 2004, <http://www.ala.org/rusa/resources/guidelines/virtrefguidelines>.

[10] Cfr. Yasmin Morais - Sara Sampson, *A content analysis of chat trancripts in the Gerogetown Law Library*, «Legal reference services quarterly», 29 (2010), n. 3, p. 165-178: 169.

mamente ricco di informazioni e di dati cui ricorrere per interpretare, sviluppare e migliorare i servizi di assistenza all'utenza. Per esempio forniscono dati per migliorare l'usabilità dei siti delle biblioteche[11], ma sempre attraverso una preliminare analisi manuale, con successiva codifica della tipologia dei contenuti; infine i dati ottenuti sono trattati con diversi software per fare un conteggio statistico e poi frequenziale.
Fino a questo punto abbiamo potuto osservare come, principalmente in ambito statunitense, le trascrizioni delle chat siano state analizzate da un punto prevalentemente manuale e statistico, per poter comprendere i bisogni informativi dell'utenza o per meglio definire l'impiego o la formazione del personale bibliotecario addetto al reference digitale.
Una voce fuori dal coro è Mungin, che in un recente studio concentra le sue attenzioni non su un obiettivo prestabilito, ma utilizza un'analisi dei dati prettamente qualitativa al fine di individuare le criticità del servizio:

> Rather than forming hypotheses about the chat reference data before analyses, this library seeks to dig deeply into the large dataset available and then form and test hypotheses from there, which will inform future research on this data set and will inform policy within the organization[12].

In Europa l'introduzione di sistemi di *virtual reference* basati su chat è avvenuta solo successivamente. Nel 2002 Bakker presenta un progetto olandese di implementazione di *virtual reference* attraverso il software QuestionPoint[13] cui intende partecipare la Royal Library[14]. Rösch nel 2003 illustra una panoramica sull'attivazione del servizio di *virtual reference* in alcune biblioteche universitarie tedesche[15]. In Spagna, presso la biblioteca dell'Universidad de Sevilla, nel 2009 è stato avviato un progetto di implementazione del *virtual reference* utilizzando LibraryH3lp[16]. In Inghilterra una pubblicazione di Haerkoenen, Blackmore e Beadle descrive l'attivazione di un servizio di *live chat* alla Cardiff University[17].
La scena universitaria italiana registra un numero ancora esiguo di progetti legati all'utilizzo di strumenti di assistenza sincrona operati online. Al momento i sistemi bibliotecari di ateneo delle università di Parma e di Torino sono esempi di realtà che offrono ai loro utenti un servizio di *live reference* dedicato.
Per quanto riguarda l'aspetto relativo al monitoraggio dei servizi di *virtual reference* legati alla chat sincrona, ci sembra doveroso precisare che a differenza della fervida attività statunitense, in Europa vi è scarsa traccia di progetti analoghi. Uno dei più significativi, oggetto di un articolo del 2014[18], consiste nell'analisi del servizio di reference online offerto dall'Universidad de Sevilla in cui, basandosi sulla scala READ, viene esaminata la diversa tipologia delle richieste degli utenti.
A differenza dei progetti di analisi già citati, tutti orientati in primo luogo al miglioramento della qualità della comunicazione tra utente e operatore, la ricerca oggetto del presente articolo si colloca all'interno di una tipologia prettamente legata all'analisi di tipo quantitativo e statistico volta alla creazione automatica di FAQ (*frequently asked questions*). In letteratura, lo studio che per finalità si inserisce in una posizione maggiormente affine alla nostra è quello sviluppato nel 2009 dalla University of Notre Dame (USA) da Jones, Kayongo e Scofield. Nel progetto in questione, l'analisi delle conversazioni di reference tramite chat è stata utilizzata al fine di produrre una lista di FAQ che, a differenza della semplice esposizione su una pagina web in formato elenco, consenta di consultare le FAQ

[11] Cfr. Suhua C. Fan - Jennifer M. Welch, *Content analysis of virtual reference data: reshaping library website design*, «Medical reference services quarterly», 35 (2016), n. 3, p. 294-304.
[12] Michael Mungin, *Sats Don't Tell the Whole Story: Using Qualitative Data Analysis of Chat Reference Transcripts to Assess and Improve Services*, «Journal of library & information services in distance learning», 11 (2017), n. 1-2, p. 25-36: 27.
[13] Per maggiori informazioni su QuestionPoint cfr. <www.questionpoint.org>.
[14] Cfr. Trix Bakker, *Virtual Reference Services: Connecting Users with Experts and Supporting the Development of Skills*, «LIBER quarterly», 12 (2002), n. 2-3, p. 124-137, <http://doi.org/10.18352/lq.7676>.
[15] Cfr. Hermann Rösch, *Digital Reference in Deutschland:* Überblick *und spezifischer Kompetenzbedarf*, in *World Library and Information Congress: 69th IFLA General Conference and Council*, Berlin, 1-9 August 2003, <https://archive.ifla.org/archive/IV/ifla69/papers/107g-Roesch.pdf>.
[16] Cfr. Nieves González-Fernández-Villavicencio - Juan-Antonio Barrera-Gómez - María-José Gómez-Fernández - Miriam Moscoso-Castillo - Victoria Santos-Flores - Marta Suárez-Samaniego, *Referencia virtual en la Biblioteca de la Universidad de Sevilla: una experiencia colectiva*, «El profesional de la información», 18 (2009), n. 6, p. 633-641, <http://eprints.rclis.org/17328>.
[17] Cfr. Sonja Haerkoenen - Andrew Blackmore - Robert Beadle, *Creating a successful chat library service: "Ask a librarian live" at Cardiff University*, «SCONUL focus», 56 (2002), p. 45-50, <http://orca.cf.ac.uk/id/eprint/43117>.
[18] Cfr. Nieves González-Fernández-Villavicencio - Encarnación Cánovas-Álvarez - Consuelo Arahal-Junco, *Evaluación del servicio de referencia de una biblioteca universitaria: Biblioteca de la Universidad de Sevilla*, «Revista española de documentación científica», 37 (2014), n. 2, p. e045, <http://dx.doi.org/10.3989/redc.2014.2.1072>.

stesse attraverso un motore di ricerca[19].
Il Sistema bibliotecario dell'Università degli studi di Torino, da quando ha inaugurato il servizio di reference digitale sincrono, ha potuto raccogliere moltissimi dati grazie alle caratteristiche del software adottato che salva le conversazioni in *log*. I dati cui ci si riferisce nel lavoro oggetto del presente articolo sono i contenuti stessi delle conversazioni.
La combinazione di due professioni, il bibliotecario di reference e il linguista computazionale, ha permesso un'inversione delle azioni di indagine finora descritte nella letteratura citata. Il punto di partenza sono state l'analisi e l'estrazione automatica di informazioni. Solo successivamente è stato necessario un intervento manuale che all'interno del nostro progetto ha portato alla produzione di FAQ.

Premesse

Il Sistema bibliotecario di ateneo dell'Università di Torino ha lanciato il servizio di reference sincrono[20] il 20 gennaio 2014 nell'ambito dell'uscita in produzione di TUTTO[21], nome con cui si è ribattezzato il *discovery tool*[22] Primo di Ex Libris. La chat si serve del software statunitense SnapEngage[23], semplice e intuitivo da installare e usare: a differenza degli *instant messenger* (per esempio Skype) non richiede necessariamente la creazione da parte dell'utente di un *account* né di una comunità di contatti per poter comunicare con gli altri. TUTTO, certamente molto intuitivo nelle sue funzioni, era una novità per gli utenti dell'Università di Torino (da ora in avanti UniTO). In quanto punto di accesso unificato alle risorse bibliografiche di UniTO, per la prima volta presentava agli utenti una ricerca integrata: con un'unica interfaccia infatti diventava possibile ricercare i dati bibliografici provenienti sia dal catalogo, sia dalle piattaforme editoriali. Questo richiedeva però alcune abilità nel reperimento dei documenti, per esempio saper distinguere tra risorse elettroniche e cartacee, capire come accedere al documento selezionato, sapere applicare filtri ecc.
Per un bibliotecario, queste sono le consuete difficoltà di un utente, ma nel caso di TUTTO è importante considerare che ci trovavamo in un ambiente differente da quello della biblioteca fisica: il *discovery* è del tutto simile a un motore di ricerca con un singolo *box* dove inserire i termini e con la possibilità di cercare e di collegarsi a documenti di diverse nature.
Ciò comporta alcuni presupposti quali la capacità dell'utente di sapersi orientare in un ambiente progettato con un proprio linguaggio e con particolari modalità di accesso agli oggetti in esso contenuti. Per questo come è importante imparare a orientarsi nella biblioteca fisica, interpretandone il linguaggio, le scritte esplicative e le modalità di accesso ai documenti in essa disponibili, così la chat in TUTTO era necessaria per aiutare gli utenti a capire le indicazioni bibliografiche e in che "direzione" andare per reperire i contenuti. Inoltre qualcuno doveva fornire le "chiavi" per aprire le porte della biblioteca virtuale che gli utenti avevano deciso di consultare.
Quindi inizialmente la chat integrata in TUTTO aveva lo scopo di guidare nell'uso del *discovery*, per dare istruzioni e valorizzare il servizio; ma presto si è rivelata anche un termometro di grande importanza per cogliere le esigenze degli utenti. Man mano si sono evidenziate delle vere e proprie casistiche di problematiche che hanno portato l'Ufficio Servizi bibliografici digitali di UniTO a investire nell'analisi delle conversazioni avvenute in chat. Il frutto di questo lavoro è stata la produzione di una lista di *frequently asked question* (da qui in poi FAQ).

Le dinamiche delle chat[24]

In gran parte delle occasioni, l'utente apre la chat perché ha bisogno di aiuto per avviare la propria ricerca e non conosce i servizi bibliografici a sua disposizione. Oppure perché non ha gli strumenti per comprende-

[19] Cfr. Sherri Jones - Jessica Kayongo - John Scofield, *Ask Us Anytime: Creating a Searchable FAQ Using Email and Chat Reference Transcripts*, «Internet reference services quarterly», 14 (2009), n. 3-4, p. 67-81 <http://dx.doi.org/10.1080/10875300903256555>.
[20] Con reference sincrono si intende un servizio di assistenza in tempo reale, nello specifico ci riferiamo a una *live chat*. Diverso è il servizio di reference asincrono, in cui domanda e risposta si articolano in fasi temporali separate, come per esempio un'assistenza fornita via e-mail. Nel caso del servizio di reference online fornito dal Sistema bibliotecario d'ateneo dell'Università di Torino, l'applicativo utilizzato passa automaticamente in modalità e-mail, quindi asincrona, quando l'operatore non risulta attivo.
[21] Cfr. <www.tutto.unito.it>.
[22] I *discovery tool* appartengono all'ultima generazione di strumenti per la ricerca bibliografica. Si tratta di prodotti in grado di raccogliere e organizzare in un unico indice metadati bibliografici provenienti da diversi database. Il principale è quello centrale, alimentato con dati forniti dagli editori. Gli altri database sono quelli locali dell'istituzione che acquisisce il *discovery*, per esempio catalogo e collezioni digitali.
[23] Per maggiori informazioni su SnapEngage, cfr. <www.snapengage.com>.
[24] Dal 2 febbraio 2016 è iniziata la sperimentazione del servizio di *live chat* anche sui siti di alcune biblioteche dell'ateneo, nell'ambito del progetto "Integrazione di servizi e attività delle biblioteche accademiche di UniTo con altre realtà del territorio: biblioteca digitale e percorsi di information

re il mancato accesso ai contenuti di un documento: “Sto facendo una ricerca per la tesi ma non trovo niente! - Devo iniziare la tesi e ho bisogno di cercare materiale, da dove parto? - Ho trovato un libro ma non capisco in quale biblioteca devo andare - Perché ho trovato un articolo, ma non riesco a scaricarlo?”.
Nell’esperienza maturata in questi quasi tre anni di lavoro in chat, abbiamo potuto cogliere che i nostri utenti prima di contattarci partono dai motori di ricerca o dai siti istituzionali del Sistema bibliotecario[25] e delle biblioteche UniTO.
Spesso Google è il mezzo con cui iniziano la vera e propria ricerca bibliografica, limitandosi a inserire nel *box* il titolo dell’articolo, o una stringa o il nome della piattaforma editoriale che sanno essere di loro riferimento. Certamente in risposta si ottengono moltissimi risultati, ma quando arriva il momento di aprire un *full text* e si è *off-campus*[26], si verifica il primo “stop” della ricerca. Oppure, partendo dai siti delle istituzioni bibliotecarie dell’ateneo, il grosso scoglio è capire quale sia il servizio più efficace per selezionare il materiale bibliografico o la fonte più appropriata.
L’insieme di queste motivazioni, riflessioni e domande ci ha portati ad aggiungere un accesso alla chat di assistenza sul sito del Sistema bibliotecario di ateneo. A questo punto è stato necessario ricalibrare l’azione del bibliotecario-operatore di chat: dalle istruzioni sull’uso del *discovery*, abbiamo ampliato la nostra assistenza anche all’uso di altri servizi[27] forniti in ateneo per la ricerca bibliografica.
Le chat sono state affrontate con le tradizionali tecniche di intervista di reference, considerando che una delle parti più complesse di questo aspetto è la negoziazione della domanda.
Spesso il richiedente non ha ancora chiaro il suo bisogno e non sempre l’oggetto della sua richiesta d’esordio corrisponde alla sua vera esigenza. Per esempio: “Cerco il libro “X” e non trovo niente!” il più delle volte significa non che l’utente ha cercato il libro con TUTTO o col catalogo e non l’ha trovato, ma che lo ha cercato attraverso il web (magari anche navigando nel nostro sito), senza individuare lo strumento specifico con cui eseguire la ricerca[28].
Il reference digitale sincrono è spesso definito in letteratura *quick* o *ready reference* perché ritenuto un tipo di assistenza rapido, dove difficilmente è possibile approfondire il problema posto o curare la qualità di una risposta[29]. Fatte salve le situazioni in cui l’affluenza di utenti sia molto alta (più richieste in simultanea per ciascun operatore), non è detto che si debbano avere per forza scambi sintetici. Molto dipende dalla dinamica che si stabilisce tra utente e bibliotecario. La cosa più importante che un operatore di chat deve fare è interpretare rapidamente la predisposizione o meno di chi lo contatta a ricevere istruzioni dettagliate[30]. Quindi è fondamentale avviare una conversazione in modo da trattenere l’utente in chat per fornirgli il maggior numero possibile di informazioni utili a soddisfare la sua domanda, senza però arrivare ad annoiarlo. Come suggerito da Bridgewater e Cole:

> Let’s not assume that there is a direct correlation between the amount of time we spend with a patron and their satisfaction with the encounter - in reference, either face to face or virtual, quantity does not equal quality[31].

Non solo. Gli utenti che usano una *live chat* stanno

literacy”. Le biblioteche coinvolte sono la “Norberto Bobbio” (Dip. di Culture, politiche e società), la “Arturo Graf” (Biblioteca storica gestita dal Sistema bibliotecario d’ateneo), la Biblioteca di Economia e management e la Biblioteca federata di Medicina “Ferdinando Rossi”. Le chat sono gestite da operatori locali dedicati al servizio. Il lavoro di analisi riportato in questo articolo fa riferimento solo ed esclusivamente ai *log* generati nelle istanze presenti sui siti <www.tutto.unito.it> e <www.sba.unito.it>.

[25] Cfr. <www.sba.unito.it>.

[26] Per *off-campus* si intende un accesso online eseguito con una rete che non sia quella dell’ateneo, per esempio quella di casa, e quindi non autorizzata dall’editore ad aprire o scaricare i contenuti delle pubblicazioni.

[27] Primo tra tutti il *Proxy* UniTO, un sistema che, modificando le impostazioni del *browser*, consente all’utente di accedere ai servizi degli editori come se navigasse all’interno della rete di ateneo. Da fine 2016, l’Ufficio Servizi bibliografici digitali ha lanciato anche il nuovo servizio Bibliopass, basato su software EZProxy. La finalità di Bibliopass è analoga a quella del *proxy*, ma ha una modalità di utilizzo più semplice per l’utente, che inserendo le sole credenziali istituzionali può navigare tra le risorse sottoscritte da UniTO anche se accede da reti esterne. Gli operatori della chat collaborano con gli amministratori di Bibliopass nell’assistenza agli utenti.

[28] «One person tries to describe for another person not something he knows, but rather something he does not know», Robert Taylor, *Question-Negotiation and Information Seeking in Libraries*, Bethlehem, Center for the Information Sciences Lehigh University, 1967, p. 5.

[29] Cfr. Rachel Bridgewater - Meryl B. Cole, *Instant Messaging Reference: A practical guide*, Oxford, Chandos Publishing, 2009, p. 110-117.

[30] Chiaramente questa caratteristica non è solo del reference online sincrono. La ritroviamo anche nel reference offerto al bancone di una biblioteca.

[31] R. Bridgewater - M. B. Cole, *Instant Messaging Reference* cit., p. 84.

agendo online, cercano le informazioni bibliografiche o l'accesso a queste ultime possibilmente dallo stesso ambiente, ovvero il web. In caso di interruzioni della ricerca, di impossibilità di capire come accedere a una fonte, il reference sincrono offre la possibilità di chiedere e risolvere nell'immediato il problema, cioè risulta utile[32]: l'utente non avrà bisogno di fermarsi nella sua attività, magari per avviare uno scambio di e-mail con la biblioteca, che per quanto rapido, richiederà nella migliore delle ipotesi qualche ora, o addirittura non dovrà lasciare la sua postazione di lavoro per recarsi fisicamente in biblioteca per la soluzione di una questione magari di tipo tecnico.

In questa sede tralasceremo la metodologia con cui sono state trattate in questi tre anni le richieste per le ricerche bibliografiche (ci riferiamo alla costruzione di stringhe, uso di *thesauri*, ricerche di documenti ecc.); ci concentreremo invece sulle richieste legate ai servizi offerti dal Sistema bibliotecario di UniTO.

Abbiamo già accennato alla possibile difficoltà degli utenti di selezionare i servizi per la ricerca e di accedere alle risorse elettroniche quando sono *off-campus*. Sono proprio queste le richieste che ricorrono stabilmente nelle nostre *live chat* da quando abbiamo inaugurato il servizio. La costante frequenza con cui alcune domande vengono poste è diventata motivo di riflessione dell'ufficio, facendo sorgere la necessità di creazione di un modello di FAQ mirato e calibrato sui bisogni emersi.

Le chat già avvenute costituivano il bacino da cui estrapolare le domande e le risposte su misura per i nostri utenti.

L'analisi dei contenuti delle chat[33] e l'estrazione delle FAQ

SnapEngage raccoglie in *log* e mette a disposizione in file di testo le conversazioni avvenute tra utenti e operatori. In tre anni è stata collezionata una quantità di materiale molto significativa e difficile da analizzare manualmente, per un totale di 2.636 scambi, sia sincroni che asincroni. Quindi è stato fissato il primo obiettivo: cercare in maniera sistematica nelle chat domande e risposte già formulate e già elaborate, così da non doverle produrre *ex novo*. In caso di risposta positiva, avremmo avuto l'individuazione di problematiche reali già trattate e risolte, meccanizzando l'analisi.

La prima operazione è stata quella di selezionare delle parole chiave (*keyword*) che facessero emergere i bisogni informativi degli utenti.

Vista la quantità di scambi avvenuti, non era possibile effettuare manualmente un'estrazione dei termini[34]; si è quindi praticata una "tokenizzazione" del file testuale che ci ha permesso il conteggio delle occorrenze di ogni parola.

Essendo lo scopo del nostro lavoro la stesura di FAQ sui servizi offerti dal Sistema bibliotecario di ateneo, è stata operata una cernita sul risultato ottenuto da tale conteggio. Infatti, per il momento, sono stati scartati i termini strettamente legati alle ricerche bibliografiche (titoli di documenti), alle discipline, alle banche dati e tutte quelle parole che nelle loro funzioni sono paragonabili a *stopword*. Questa selezione è avvenuta in base a due criteri: il primo basato sulle competenze di reference dei bibliotecari-agenti della chat e sulla loro esperienza maturata nel corso della gestione del servizio. Il secondo è stato fornito dai dati numerici ricavati dal conteggio, che ha in effetti confermato l'incidenza di alcuni termini, compresi quelli selezionati dai bibliotecari.

È importante chiarire che le conversazioni sono state analizzate facendo la ricerca delle stesse *keyword* prima negli interventi degli utenti e poi in quelli degli agenti. Date le due diverse matrici, gli output hanno restituito liste di termini con valori delle occorrenze molto diversi[35]. Abbiamo ritenuto quindi significativi i risultati dei conteggi fatti nelle battute degli utenti. Queste certamente rappresentavano le richieste genuinamente formulate da chi è il destinatario del servizio in elaborazione.

Il passo successivo è stato ricercare la lista di *keyword* nelle richieste degli utenti; quindi di associare a ogni loro intervento la successiva risposta dell'agente.

Il risultato di questo procedimento ha permesso di ottenere in maniera automatica coppie di scambi, molti dei quali con un significato compiuto e in larga parte

[32] Lynn S. Connaway - Marie L. Radford, *Seeking Synchronicity: Revelations and Recommendations for Virtual Reference*, Dublin (OH), OCLC Research, 2011, p. 59, <http://www.oclc.org/en/reports/synchronicity.html>.

[33] Tale analisi dei contenuti è avvenuta secondo la norma che regolamenta il trattamento dei dati, ai sensi dell'art. 13 del Codice della *privacy* (d.lgs. n. 196/2003). Quando l'utente inizia la conversazione in chat, un messaggio automatico lo informa del trattamento dei dati e lo invia a una pagina di dettaglio sulle modalità di gestione delle eventuali informazioni personali (<http://www.serviziweb.unito.it/primo/help/privacy_policy_snapengage.html>).

[34] Tra utenti e operatori-bibliotecari avremmo dovuto visionare un totale di 63.876 righe.

[35] Queste differenze sono giustificabili per via delle diverse priorità dei due soggetti considerati e per il diverso uso del linguaggio. Per il bibliotecario l'uso dei termini avrà prevalentemente una giustificazione tecnica, piuttosto che spontanea.

in grado di esprimere una domanda con la relativa risposta.
Ciò ha comportato due risultati: il primo è stata un'analisi automatica delle conversazioni utente/bibliotecario, con un notevole risparmio di tempo rispetto al lavoro manuale. Il secondo, ancora più importante, è stato l'ottenimento di richieste reali e con i termini consueti degli utenti; successivamente, quello di risposte già articolate e soprattutto provate come utili perché già verificate durante l'assistenza sincrona. Nel corso di quest'ultima, infatti, è possibile accompagnare l'utente finché non riesce nel suo intento, superando il rischio di frustrazione tanto diffuso in chi approccia sistemi non familiari.

Un esempio

Sono numerose le problematiche emerse durante le nostre assistenze, ma quella che è ricorsa e ricorre più frequentemente è la configurazione del *Proxy* UniTO per l'accesso ai contenuti digitali *off-campus*. Per esemplificare il lavoro di analisi dei contenuti fin qui descritto, la useremo come oggetto di osservazione.
La richiesta di assistenza per l'accesso "da casa" rivela una lacuna nelle informazioni possedute dall'utente. Google è uno strumento fondamentale, garantisce in maniera quasi certa una risposta a qualunque richiesta gli sia posta (non entreremo in questo caso nel merito della valutazione qualitativa dal punto di vista accademico delle fonti da cui provengono le risposte del motore di ricerca). Ma Google indicizza pagine web, che sono sempre, o quasi, accessibili. In definitiva, se Google mi dice che esiste una risposta alla mia domanda, allora penso di poterla anche leggere.
Come è ben noto, per accedere ai contenuti editoriali nella maggior parte dei casi è necessario un abbonamento. Quando i nostri utenti UniTO fanno ricerca nel web o con i nostri servizi, lo fanno con un approccio ispirato a Google: se trovo la descrizione di un articolo mentre studio a casa perché non riesco anche a leggerne il *full text*?
Abbiamo poi utenti che hanno già consultato riviste elettroniche tramite le postazioni di una biblioteca o la rete d'ateneo, perciò sanno che UniTO vi è abbonata. Quando poi fanno la ricerca da casa, non avendo una sottoscrizione, ricevono una risposta negativa al tentativo d'accesso. La loro richiesta pertanto è se possono usufruire dei contenuti editoriali tramite l'ateneo collegandosi da casa.
Infine, le richieste d'assistenza sono arrivate anche da utenti perfettamente consapevoli dell'esistenza del servizio *proxy*, ma che in quel momento avevano difficoltà nella gestione della configurazione.
Sulla base di queste premesse e sostenuti dal conteggio delle occorrenze operato attraverso la "tokenizzazione", sono state selezionate come *keyword* sull'argomento i termini "*proxy*" (574 occorrenze), "accedere" (398 occorrenze) e "accesso" (199 occorrenze), "riesco" (445 occorrenze)[36].
Nel dettaglio: "*proxy*" ha permesso di estrarre quegli scambi che vedevano nella domanda dell'utente una richiesta di aiuto nella configurazione del servizio, segnalazioni di eventuali malfunzionamenti, di anomalie, di impossibilità di accesso ai contenuti o al servizio *proxy* stesso ecc.
Fanno eccezione quelle conversazioni in cui la *keyword* compare nel testo dell'utente perché ripetizione di un'istruzione ricevuta precedentemente da un operatore.
"Accedere" e "accesso", invece, sono state *keyword* utili a individuare le conversazioni in cui l'utente aveva formalmente capito la necessità di una procedura o di una via istituzionale per accedere a un servizio, compreso quello della consultazione *off-campus* delle risorse online. I termini sono stati scelti perché esprimono una certa competenza di linguaggio. Sottintendono la consapevolezza che per compiere alcune operazioni sono necessarie delle condizioni. Alle volte gli utenti le usano nel giusto contesto, altre volte no, ma sta a noi bibliotecari attribuire i giusti significati. Certamente, nell'ambito del nostro progetto, "accedere" e "accesso" ci hanno permesso di estrarre conversazioni legate all'uso dei servizi e quindi anche del *Proxy* UniTO.
Il terzo termine "riesco" esprime il desiderio di raggiungere un obiettivo, di compiere un'azione: nel nostro caso è stato la chiave per individuare richieste di assistenza ad ampio spettro. Tra esse quelle di aiuto per modificare le impostazioni di navigazione.

Il risultato

Qui di seguito mostriamo un esempio di scambio con domanda e risposta estratto tramite la procedura automatica cui sopra abbiamo accennato:

- PRIMA
<keyword=ricerca>ciao , cosa offre in più TUTTO rispetto a siti di ricerca bibliografica come world of knowledgeo o Science Direct ?

[36] Per avere un termine di paragone, diciamo che, escludendo i conteggi di articoli, verbi ausiliari, segni d'interpunzione, preposizioni, la parola "grazie" ricorre 3.867 volte e "biblioteca" 774.

<answer from Maria Vittoria>ti permette di consultare contemporaneamente le piattaforme e le banche dati | cui UniTO è abbonata | interroghi anche il catalogo | e ricerchi tra gli abbonamenti alle riviste elettroniche | in sintesi : | è un unico accesso alle risorse possedute dall' università | anche se non contiene letteralmente tutto | per esempio | se fai una ricerca | nell' elenco dei risultati ti restituisce sia quelli provenienti da w of knowledge e science direct |[37]
- DOPO
Cosa offre in più TUTTO rispetto a siti di ricerca bibliografica come Web of Knowledge o Science Direct?
Ti permette di consultare contemporaneamente le piattaforme e le banche dati cui UniTO è abbonata, comprese Web of Knowledge e Scopus. Interroghi anche il Catalogo e ricerchi tra gli abbonamenti alle riviste elettroniche. In sintesi: è un unico accesso alle risorse possedute dall'Università, anche se non contiene letteralmente tutto.

Questo caso mostra come il lavoro successivo alla selezione della coppia utile alla produzione di FAQ sia consistito semplicemente (e nella maggior parte dei casi) in un'opera di rieditazione e non di produzione di contenuti. La domanda infatti risulta già chiara e composta. La risposta invece ha necessitato di piccole rielaborazioni, prevalentemente formali.

L'approccio *natural language processing*

Premesse metodologiche

Nonostante la realizzazione di una pagina di FAQ sia un'operazione concettualmente semplice che consiste essenzialmente nella selezione mirata dei più significativi scambi comunicativi finalizzati alla risoluzione di un determinato problema, se svolta manualmente può rivelarsi particolarmente onerosa, richiedendo un quantitativo di tempo dedicato alle operazioni di analisi e selezione proporzionale alle dimensioni della base di dati testuale su cui si intende agire. Proprio in virtù di questa valutazione, risulta evidente la necessità di orientarsi in direzione di una manipolazione automatica dei dati di testo, cosa possibile mediante l'utilizzo di metodologie e procedure tipiche della linguistica computazionale, in particolare di quella branca denominata linguistica dei *corpora*. Proprio l'applicazione di metodologie ormai consolidate nell'ambiente dell'elaborazione delle lingue naturali a un settore relativamente di nicchia quale il reference bibliotecario rappresenta un importante elemento di novità sulla scena italiana, nonché un fattore di innovazione all'interno del più generale contesto biblioteconomico. La scelta di utilizzare una tecnologia di elaborazione dei dati mutuata da una tradizione più propriamente legata all'analisi elettronica dell'informazione linguistica è data dal fatto che l'insieme delle registrazioni della chat in nostro possesso costituisce una raccolta di dati che, opportunamente trattata, può essere considerata a tutti gli effetti un *corpus* linguistico secondo la seguente definizione:

> Raccolta di testi (scritti, orali o multimediali) o parti di essi in numero finito in formato elettronico trattati in modo uniforme (ossia tokenizzati ed addizionati di markup adeguato) così da essere gestibili ed interrogabili informaticamente; se (come spesso) le finalità sono linguistiche (descrizione di lingue naturali o loro varietà), i testi sono perlopiù scelti in modo da essere autentici e rappresentativi[38].

Partendo da tale assunto, risulta evidente che la prima operazione svolta sia stata la riunione dei numerosi singoli file di *log* in un unico documento; operazione che ha richiesto la massima attenzione affinché anche nel nuovo documento si potesse garantire il mantenimento della corretta sequenzialità temporale dei diversi scambi comunicativi. Il passo immediatamente successivo alla ricostituzione dei *log* in un file unitario ha riguardato le necessarie operazioni di pulitura dei dati stessi. In particolare, si è visto innanzitutto necessario procedere a un intervento di ricodifica di alcuni caratteri dal formato tipicamente utilizzato nelle pagine HTML allo standard UTF8 (ad esempio: "(" = "(" , "à" = "à"). A ciò ha fatto seguito la ricostruzione su una singola linea di quegli interventi degli utenti frazionati su più righe successive (Fig. 1), unitamente all'eliminazione di tutti i dati non testuali o non rilevanti per gli scopi del progetto.

```
Visitor: 14:05:26: Tipologia :  03C-Nota a Sentenza
Autori:  DEZZANI, Flavio
Autori :  F. Dezzani, L. Dezzani
Titolo :  Documento Oic n. 4 - La fusione inversa
Data:  2010
Titolo della rivista :  IL FISCO
Volume :  22
Pagina iniziale :  1-3415
Pagina finale :  1-3421
È visualizzato nelle collezioni: Nota a Sentenza
```

Fig. 1: Informazioni organizzate in righe multiple.

[37] Il carattere "|" (*pipe*) è stato inserito per indicare che nella fonte originale la risposta è suddivisa su più righe. Invece, nella procedura automatica di estrazione delle FAQ queste righe sono state riunite in un unico blocco di testo.

[38] *Corpora e linguistica in rete*, a cura di Emanuele Barbera, Elisa Corino, Cristina Onesti, Perugia, Guerra edizioni, 2007, p. 70.

Struttura dei documenti originali e prime elaborazioni

Al fine di comprendere meglio la natura degli interventi svolti, è necessario in primo luogo esaminare le caratteristiche strutturali del documento originale, descrivendo quindi le modifiche apportate. Il primo aspetto su cui è necessario concentrare l'attenzione riguarda il formato del documento da elaborare. Nel nostro caso il file complessivo consiste in una sequenza di 2.636 interscambi formattati da SnapEngage stesso secondo lo schema CSV (*comma separated values*). Per quanto concettualmente paragonabile a una tabella di database, questa tipologia di file rappresenta un documento di testo in cui ogni singola riga (*record*) si compone di una serie prestabilita di campi separati da virgola. Rispetto a quanto si può evincere dalla Fig. 2, che a prima vista sembra non rispettare i termini appena indicati, è necessario precisare che la presenza di caratteri di invio (LF, *line feed*) all'interno dello scambio comunicativo tra i due interlocutori, dà inevitabilmente luogo alla frammentazione dell'intero *record* su righe multiple. Per quanto concerne le operazioni di filtraggio, è doveroso puntualizzare che questa elaborazione viene svolta in un secondo momento e può riguardare tanto campi specifici quali la durata della chat o l'indirizzo IP dell'utente, quanto righe intere caratterizzate da contenuti non rilevanti ai fini di questo progetto quali l'apertura di specifiche pagine web da parte dell'utente ("Visitor page reloaded", Fig. 2). Sebbene a prima vista possano essere considerate superflue o scarsamente rilevanti, le attività appena menzionate rappresentano al contrario un fattore di primaria importanza: solo così facendo, infatti, è possibile rendere l'intero documento anonimo, eliminando eventuali riferimenti ai dati personali che possano portare all'identificazione dell'utente. In aggiunta a quanto descritto, si è deciso di facilitare l'identificazione di ogni singola sessione di chat e, di conseguenza, rendere più agevoli le successive elaborazioni, mediante l'inserimento di un elemento di annotazione meta-testuale (comunemente noto con il termine *markup*) rappresentato dalla coppia di etichette compatibili con il formato XML <chat>…</chat> (Fig. 2). Come ultima considerazione, è necessario segnalare che ogni chat si apre sempre con un messaggio automatico che comunica all'utente la disponibilità di fornire assistenza e la relativa *policy* sulla *privacy*[39]. Trattandosi di un messaggio ripetuto e sempre identico nel tempo, abbiamo deciso di eliminare tale *incipit* dal file pre-elaborato onde evitare operazioni su dati falsati che porterebbero a risultati fuorvianti.

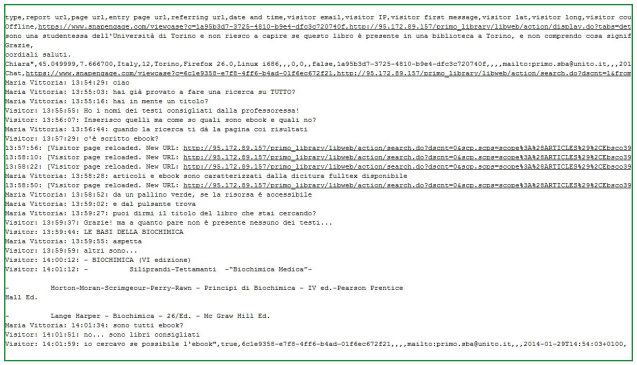

Fig. 2: Frammento di file di *log* originale.

La "tokenizzazione"

Per quanto riguarda il testo della chat, è fondamentale chiarire che in questa seconda fase di preparazione anch'esso ha dovuto affrontare un primo livello di manipolazione, subendo un processo automatico che prende il nome di "tokenizzazione". A questo proposito:

> Per tokenizzazione si intende grossomodo l'operazione di individuazione (in genere tramite un *blank* a destra ed a sinistra) dei token, ossia delle unità minime che il PC tratterà. Siccome queste, peraltro, non corrispondono sempre alle parole grafiche di un testo "tipograficamente composto", basta già questa operazione da sola a distinguere i due oggetti[40].

Da quanto sopra, risulta pertanto evidente che questa elaborazione, indispensabile per garantire la correttezza dei risultati prodotti, consiste essenzialmente nell'individuazione di ogni singolo elemento testuale dotato di valore semantico autonomo (*token*) per poi procedere, laddove necessario, alla sua successiva separazione dalle forme testuali attigue mediante un carattere di spazio (Fig. 2). È interessante segnalare che all'interno di questa fase, per quanto non direttamente utile ai fini della produzione del documento di FAQ, è stata prestata notevole attenzione anche alla corretta identificazione dei vari *emoticon* che gli inter-

[39] Il messaggio automatico recita: «Ciao, consulta l'informativa sulla privacy <http://www.serviziweb.unito.it/primo/help/privacy_policy_snapengage.html> dove spieghiamo in modo semplice come tratteremo i dati che tu eventualmente potresti fornirci durante la chat. Come posso aiutarti?»

[40] *Corpora e linguistica in rete* cit., p. 27.

locutori possono utilizzare all'interno del loro scambio comunicativo. Per motivi di completezza è opportuno segnalare che le conversazioni tra utente e operatore possono contenere forme invariabili complesse che da un punto di vista strettamente linguistico rappresentano voci lessicali autonome. Tali elementi sono stati oggetto di accurata selezione affinché fosse possibile gestirli come *token* unitari: queste forme, infatti, sono del tutto riconducibili a espressioni poliematiche, ossia quei costrutti particolari il cui significato è indipendente e non desumibile dagli elementi lessicali che lo compongono (ad esempio "ferro da stiro" o "forza lavoro"). Pertanto, sulla base di queste valutazioni, sono stati predisposti specifici criteri di raggruppamento finalizzati alla corretta gestione di:

- riferimenti bibliografici completi o parziali («A Head for an Eye: Revenge in the Cambodian Genocide Hinton, Alexander Laban American Ethnologist, 1998 Aug, Vol.25(3), pp.352-77»);
- nomi di case editrici o piattaforme editoriali (Franco Angeli; Science Direct);
- informazioni bibliografiche di natura amministrativa («Inventario INP 4362 Collocazione P*126 B 07 Note 1 v.»);
- messaggi o avvisi di sistema (No Proxy Detected; Documento ammesso al prestito);
- luoghi o istituzioni (Palazzo Nuovo; Servizio bibliotecario d'ateneo);
- nomi propri o marchi registrati (Isaiah Berlin; Macbook Air).

In ultima istanza, bisogna considerare che all'interno dei file di *log* originali non sono registrati esclusivamente i contenuti delle comunicazioni svolte in modalità sincrona, ma anche le richieste che giungono nei periodi in cui il servizio di reference online in modalità sincrona non è attivo[41] (162 occorrenze). Oltre a ciò, è anche presente un totale di 148 casi in cui l'utente abbandona il servizio senza stabilire un effettivo contatto con l'operatore. Questo comportamento dell'utenza si può verificare immediatamente dopo l'apertura del servizio, come risposta alla frase automatica di accoglienza[42] (113 occorrenze) oppure come repentina interruzione della comunicazione e conseguente impossibilità da parte dell'operatore di fornire risposta al quesito posto[43] (35 occorrenze). Al fine di preservare l'integrità dei dati originali, senza tuttavia inquinare i contenuti utili per il progetto, si è deciso di filtrare questi interventi particolari salvandoli in file dedicati. Una volta ottenuto un documento di testo caratterizzato da 2.326 blocchi omogenei di righe di comunicazione sincrona contenenti esclusivamente i dati relativi a ruolo dello scrivente (utente o operatore), orario e contenuto dello scambio intercorso (Fig. 3), i successivi passaggi rappresentano il cuore dell'elaborazione.

Zipf, frequenze e parole chiave

L'indagine basata sulla metodologia della linguistica dei *corpora* implica necessariamente l'utilizzo di strategie di analisi di natura quantitativa e qualitativa. Dal punto di vista quantitativo, la messa in atto di modelli statistici quali la Legge di Zipf[44] è divenuta ormai prassi consolidata in quanto capace di fornire informazioni importanti sulle caratteristiche di un testo in base

```
<chat>
Visitor : 13:54:08: Ciao ! Mi servirebbe un ebook di biochimica generale . Come posso trovarne uno ?
Maria Vittoria : 13:54:29: ciao
Maria Vittoria : 13:55:03: hai già provato a fare una ricerca su TUTTO ?
Maria Vittoria : 13:55:16: hai in mente un titolo ?
Visitor : 13:55:55: Ho i nomi dei testi consigliati dalla professoressa !
Visitor : 13:56:07: Inserisco quelli ma come so quali sono ebook e quali no ?
Maria Vittoria : 13:56:44: quando la ricerca ti dà la pagina coi risultati
Visitor : 13:57:29: c' è scritto ebook ?
Maria Vittoria : 13:58:28: articoli e ebook sono caratterizzati dalla dicitura fulltex disponibile
Maria Vittoria : 13:58:52: da un pallino verde , se la risorsa è accessibile
Maria Vittoria : 13:59:02: e dal pulsante trova
Maria Vittoria : 13:59:27: puoi dirmi il titolo del libro che stai cercando ?
Visitor : 13:59:37: Grazie ! ma a quanto pare non è presente nessuno dei testi ...
Visitor : 13:59:44: LE BASI DELLA BIOCHIMICA
Maria Vittoria : 13:59:55: aspetta
Visitor : 13:59:59: altri sono ...
Visitor : 14:00:12: - BIOCHIMICA ( VI edizione )
Visitor : 14:01:12: - Siliprandi-Tettamanti - " Biochimica Medica " - - Horton-Moran-Scrimgeour-Perry-Rawn - Principi di Biochimica - IV ed. - Pearson
Maria Vittoria : 14:01:34: sono tutti ebook ?
Visitor : 14:01:51: no ... sono libri consigliati
Visitor : 14:01:59: io cercavo se possibile l' ebook
</chat>
```

Fig. 3: Frammento di file di *log* elaborato.

[41] Ad esempio «Buongiorno, non riesco a visualizzare un articolo DUECENTO AND TRECENTO II (EXCLUDING DANTE)».

[42] Ad esempio «grazie mille! se ho bisogno chiederò sicuramente!».

[43] Ad esempio «buon giorno, sto cercando un libro e sto diventando matta, potete aiutarmi?».

[44] Cfr. Alessandro Lenci - Simonetta Montemagni - Vito Pirrelli, *Testo e computer: elementi di linguistica computazionale*, Roma, Carocci, 2016, p. 137.

al suo lessico. L'approccio in questione rappresenta per l'universo linguistico l'*alter ego* della distribuzione gaussiana di un campione. Tuttavia, a differenza della curva di Gauss, capace di rappresentare graficamente i valori di una certa variabile espressione di un fenomeno noto (ad esempio la statura di tutte le persone di una data età, i voti di profitto di una classe di studenti ecc.), nel caso dell'analisi statistica applicata all'ambito linguistico, l'unico dato a noi disponibile è rappresentato dalla modalità di utilizzo dei vari elementi lessicali (*token*) all'interno dell'opera di uno specifico autore. Pertanto, poiché il modello di Zipf si fonda sul presupposto empirico secondo cui, a livello lessicale, la comunicazione avvenga sfruttando il principio del "massimo risultato con il minimo sforzo", da un punto di vista operativo tale modello si traduce in una funzione di proporzionalità inversa tra il numero di occorrenze proprie di un *token* e la sua posizione (rango) all'interno di un elenco (lista di frequenze). Il rapporto tra i valori di rango e frequenza saranno quindi lo specchio della ricchezza lessicale propria del testo sottoposto ad analisi. In merito al caso specifico oggetto del presente articolo, è opportuno sottolineare che allo scopo di ottenere un'immagine più chiara del linguaggio utilizzato, si è resa necessaria la creazione di due liste di frequenza distinte tra utente e operatore.
Oltre a ciò, poiché il sistema riconosce e gestisce in maniera diversa i caratteri maiuscoli rispetto alla loro controparte minuscola, alle liste già indicate sono stati aggiunti altri due elenchi in cui l'intero lessico è ricondotto alla sola variante minuscola. Quest'ultimo espediente si è rivelato estremamente utile per consentire lo svolgimento di calcoli statistici sul lessico utilizzato nella chat partendo da presupposti meno restrittivi poiché capaci di articolarsi su modelli differenti. Di conseguenza, al fine di ottenere la lista di frequenze che ci ha permesso di estrarre con elevata facilità la serie di parole chiave più rilevanti utilizzate dagli utenti nelle loro svariate richieste a favore dei servizi disponibili, si è ritenuto necessario operare un filtraggio automatico di quegli elementi lessicali caratterizzati tanto da frequenze estremamente elevate, quanto estremamente basse. La successiva operazione, svolta sull'elenco di parole chiave risultante dai precedenti passaggi, ha riguardato la meticolosa selezione manuale di quei termini che, sulla base di criteri semantici, si sono rivelati maggiormente significativi. La lista ristretta così prodotta è stata quindi utilizzata come base per l'estrazione automatica dei diversi contesti d'uso a partire dal documento "tokenizzato". Questa elaborazione ha portato, come ultimo passaggio, alla creazione di un documento di sequenze omogenee di scambi costituiti dalle richieste di supporto da parte dell'utente a cui fanno seguito le relative risposte. Proprio tale documento rappresenta il risultato finale che può essere utilizzato per fornire un punto di riferimento primario per tutti i fruitori dei servizi bibliotecari dell'Università di Torino.

Tipologia	Token (nr.)	Type[45] (nr.)
Utente	116.802	9.905
Operatore	207.887	9.246

Tab. 1: Dati statistici relativi al lessico del servizio di chat suddivisi tra utente e operatore senza distinzione tra carattere maiuscolo e minuscolo.

L'algoritmo di estrazione delle FAQ

Il sistema di estrazione dei contesti di frase utili per la produzione del documento di FAQ è gestito da un algoritmo che opera un confronto tra la lista di parole chiave ordinate in base alla loro frequenza e i singoli *token* che compongono le varie righe di quel file di *log* già filtrato e riformattato secondo quanto descritto nei precedenti paragrafi. Una volta che questo confronto abbia prodotto un esito positivo, il sistema interromperà l'attività in corso e memorizzerà il contenuto dell'intera riga, per poi spostarsi in quelle successive per verificare la presenza di eventuali ulteriori prosecuzioni dell'intervento dell'utente oppure l'esistenza della risposta da parte dell'operatore. Sulla base di quanto indicato, risulta chiaro, quindi, come il sistema di estrazione operi utilizzando tre cicli iterativi nidificati che agiscono su elementi distinti. Il primo di questi cicli ha la funzione di scandire il vettore in cui si trovano immagazzinate le parole chiave dopo essere state acquisite dal loro file di origine. Il secondo ciclo, che si attiva immediatamente dopo la selezione della parola chiave, si occupa, per contro, di effettuare una scansione di tutti i *token* che compongono la riga del file di *log* presa in esame.
Dal risultato positivo del confronto dei due elementi lessicali di cui sopra, svolto mediante l'utilizzo di criteri di *pattern matching* fra stringhe di caratteri, si attiva il terzo processo iterativo che, a differenza dei precedenti, opererà sulla struttura del file a livello di riga di testo. Proprio in questa fase, attraverso una verifica dell'autore dell'intervento comunicativo, distinto tra utente e operatore, il sistema sarà in grado di operare una corretta ricostruzione del blocco di domande e risposte. È bene chiarire che quanto finora descritto

[45] «Il type [...] sarebbe in prima approssimazione il descrittore della classe di tutti i token identici», *Corpora e linguistica in rete* cit., p. 35.

presuppone implicitamente l'esistenza di un ulteriore ciclo iterativo a monte avente compito di coordinare la lettura sequenziale delle singole righe del file di *log* originale con tutti i processi di elaborazione definiti nel corpo del programma. Questa funzione operativa, per quanto assolutamente fondamentale per il corretto funzionamento del sistema nel suo complesso, non è stata finora menzionata in quanto parte costitutiva e integrante del linguaggio stesso utilizzato per la realizzazione dei vari programmi di elaborazione. Il linguaggio in questione, adottato nella sua versione più recente, rappresenta la versione *open source* (certificata dalla licenza GNU[46]) di AWK[47], un prodotto appositamente progettato per facilitare la manipolazione di dati testuali e sviluppato originariamente in ambiente UNIX nel 1977 da Alfred Aho, Peter Weinberger e Brian Kernighan, il creatore del famoso linguaggio di programmazione "C".

Riassumendo, ai fini della produzione automatica del documento di FAQ sarà necessario eseguire la seguente catena di elaborazione del file di *log* originale:

- pre-elaborazione del file di *log* in formato *.csv finalizzato alla ristrutturazione su unica riga del medesimo atto comunicativo frazionato su righe successive;
- produzione di un file di *log* opportunamente filtrato e "tokenizzato", un file di richieste degli utenti inviate al di fuori del regolare orario di servizio e un file con richieste prive di risposta a seguito dell'abbandono della *chat* da parte dell'utente;
- creazione di due file di dati statistici distinti tra operatore e utente, basati sui principi della Legge di Zipf;
- selezione delle parole chiave (*keyword*) dai file statistici strutturati in ordine decrescente;
- creazione automatica di FAQ sulla base di un elenco di *keyword* ordinato alfabeticamente.

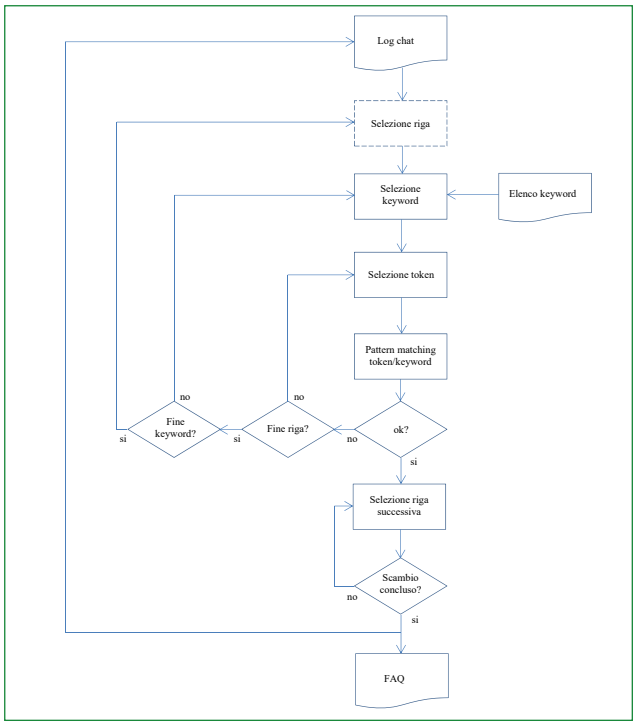

Fig. 4: Catena di elaborazione del file di *log* originale.

Conclusioni

Il progetto descritto ha raggiunto diversi obiettivi e ne pone di nuovi. L'elaborazione computazionale ha facilitato l'elaborazione statistica delle chat avvenute in questi tre anni e, di conseguenza, il lavoro di analisi sia dal punto di vista semantico sia da quello biblioteconomico, in particolare per quanto riguarda il reference. L'erogazione di servizi di un sistema bibliotecario necessita di una verifica accurata della riuscita degli intenti. Una richiesta è di per sé sintomo di un *gap* tra chi offre il servizio e chi lo usa. Cogliere questo messaggio e cercare di intervenire costituiscono il primo passo nella missione del bibliotecario.

La realizzazione di FAQ sulla base di richieste reali è stata la nostra iniziativa di partenza. L'implementazione della chat tra i nostri servizi prima e l'elaborazione delle conversazioni poi, infatti, ci hanno fornito numerosi spunti e hanno fatto emergere altrettante richieste e implicite esigenze da parte dei nostri utenti. Tutto ciò sarà oggetto delle nostre prossime attività di ricerca, sempre col supporto dell'approccio basato sulla linguistica computazionale. In particolare, il passo successivo riguarderà l'utilizzo dei dati già elaborati al fine di condurre analisi mirate non solo a evidenziare i tratti in comune e le differenze tra il linguaggio espresso dagli utenti e quello degli operatori, ma anche a evidenziare le reali necessità dell'utenza in un ottica di estensione e rafforzamento del servizio verso quelle realtà bibliotecarie non ancora direttamente coinvolte nel progetto. Anche in questo caso, l'unico scopo sarà quello di stabilire un dialogo, sia diretto che indiretto, con i nostri utenti per contribuire a rendere il mondo delle biblioteche un ambiente aperto, collaborativo e in continua crescita.

[46] Cfr. <https://www.gnu.org/licenses/licenses.it.html>.
[47] Cfr. <https://www.gnu.org/software/gawk/manual/gawk.html>.

ABSTRACT

Nel 2014 il Sistema bibliotecario dell'Università di Torino inaugura il servizio TUTTO realizzato mediante il discovery tool Primo e con esso il servizio di reference live chat, nato per fornire assistenza e orientamento alla ricerca bibliografica degli utenti.
Il presente articolo descrive la progettazione di un sistema in grado di redigere in maniera automatica una pagina di risposte alle domande più frequenti (FAQ, frequently asked questions) poste dagli utenti dei servizi bibliotecari dell'Università di Torino agli operatori del servizio di live chat. Il progetto è nato e si è sviluppato all'interno dell'Ufficio Servizi bibliografici digitali e innovazione tecnologica e si è dimostrato un metodo di analisi di grande efficacia e versatilità. La realizzazione di quanto presentato in questa sede è stata possibile attraverso l'analisi delle registrazioni (log in gergo tecnico) degli scambi avvenuti in modalità sincrona tra utenti e operatori nel periodo 2014-2016. Proprio l'uso di tecniche di estrazione automatica dei dati (text mining) ha permesso la produzione di coppie di domande e risposte sulla base di conversazioni realmente avvenute. Le metodologie applicate sono tipiche delle digital humanities, quell'ambito disciplinare che trae origine in egual misura dalle scienze linguistiche e dalle scienze dell'informazione, con particolare attenzione all'elaborazione automatica delle informazioni linguistiche (NLP, natural language processing).

IMPROVING LIBRARY SERVICES THROUGH LIVE CHAT AND NATURAL LANGUAGE PROCESSING

In 2014 the University of Turin Library system implemented the service TUTTO, which was based on the discovery tool Primo. At the same time, a live chat reference service was activated to provide users with support to bibliographic research. This paper aims at describing the design of a system which is able to automatically produce a list of frequently asked questions starting from the analysis of the communication exchanges between the Library service users and its librarians who provide online reference support. The project was developed by the University of Turin's Digital library service and technology innovation office to provide users with a set of ready-to-use information which may help solve the most basic issues before asking the virtual reference librarians for further, specific help. The automatic production of a FAQ list from real communication exchanges was possible after semi-automatically analyzing all chat log files from 2014 to 2016. To achieve this, a methodological approach derived from digital humanities was adopted. This particular academic field integrates data analysis strategies with programming skills by embracing a large number of subjects from linguistics to natural language processing. The results obtained confirm both the effectiveness of the work we have carried out and of the methodology adopted.