Implementazione di software per la gestione dei corpora LBC

Riccardo Billero

Introduzione

L'importanza del lessico dei beni culturali italiani va ben oltre i confini delle comunità linguistiche ed è oggetto di interesse per studiosi afferenti a diverse discipline (come storia dell'arte, letteratura, linguistica), per i professionisti (traduttori, guide turistiche, organizzatori di eventi) e per gli studenti di tali discipline. Attraverso l'utilizzo delle moderne tecnologie messe a disposizione dall'Informatica Umanistica, l'Unità di Ricerca "Lessico multilingue dei Beni Culturali" (LBC) si pone l'obiettivo di creare strumenti utili per i soggetti summenzionati, concentrandosi sull'arte di Firenze e della Toscana. Tra tali strumenti rivestono particolare importanza i sei corpora LBC, liberamente disponibili online, la cui creazione è descritta nel seguito.

1. Informatica Umanistica e Lessico dei Beni Culturali

L'Informatica Umanistica è un campo di studio in continua evoluzione, di cui esistono molteplici definizioni, alcune più orientate verso le tecnologie e le metodologie di indagine, altre verso l'innovazione dei contenuti e dei programmi di ricerca all'interno degli studi umanistici, mediante l'ausilio delle nuove tecnologie (Numerico, Vespignani 2003: 13-14). Per Svensson (2010) si tratta fondamentalmente dell'intersezione tra le scienze umanistiche e le tecnologie dell'informazione; ad oggi «si privilegia il concetto di *digital humanities* per riferirsi a una realtà più vasta che riguarda non solo le metodologie di costituzione

Riccardo Billero, University of Florence, Italy, riccardo.billero@unifi.it, 0000-0002-6099-222X

FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)

di corpora e di annotazione di testi, nonché l'elaborazione di programmi atti a leggere i dati raccolti in essi, ma anche la realizzazione di applicazioni pensate per poter funzionare su diversi dispositivi elettronici, utili sia alla visualizzazione, fruizione e condivisione di enormi quantità di dati, che all'interpretazione semantica degli stessi da parte dei computer, per menzionare solo alcune possibilità» (Zotti, Pano Alamán 2017: 8).

Qualunque sia la definizione che preferiamo adottare, possiamo comunque affermare che l'incontro tra l'informatica e le scienze umane offre alle nuove generazioni di ricercatori, docenti e studenti un modo nuovo di avvicinarsi alla cultura, mediante l'uso degli strumenti messi oggi a disposizione delle tecnologie contemporanee.

In questa ottica intendiamo qui descrivere il contributo che il progetto Lessico dei Beni Culturali ha ricevuto dall'Informatica Umanistica grazie al lavoro dell'autore.

Il progetto Lessico multilingue dei Beni Culturali (in breve LBC) è attivo da diversi anni presso l'omonima Unità di Ricerca con sede nel Dipartimento di Formazione, Lingue, Intercultura, Letterature e Psicologia (FORLILPSI) dell'Università degli Studi di Firenze (Billero, Nicolás Martínez 2017; Billero 2020) e si pone gli obiettivi di:

a. creare una piattaforma web di riferimento utile allo studio del lessico dei beni culturali;
b. creare dei corpora di riferimento per ognuna delle lingue coinvolte nel progetto[1];
c. fornire l'accesso a tali corpora e degli strumenti utili per la ricerca al loro interno, al fine di abilitare e promuovere studi in settori quali linguistica, lessicografia o ambiti socioculturali.
d. creare un dizionario multilingue dei termini artistici che utilizzi i corpora come riferimento.

Attualmente, le risorse principali del progetto LBC sono sei corpora contenenti una raccolta di testi in varie lingue (come di seguito descritto) e sei lemmari relativi al linguaggio dell'arte, estratti dai corpora per ciascuna lingua e successivamente elaborati dai membri dei gruppi linguistici coinvolti nel progetto. Nel seguito saranno descritti in particolare i corpora LBC ed il relativo processo di creazione.

2. I corpora LBC

I corpora LBC sono una delle principali risorse dell'Unità di Ricerca "Lessico multilingue dei Beni Culturali". Fin dall'inizio dei lavori (nel 2016), è stato previsto che i corpora LBC sarebbero stati utilizzati non solo dai membri dell'Unità

[1] Al momento le lingue coinvolte nel progetto sono: cinese, francese, inglese, italiano, portoghese, russo, spagnolo, tedesco, turco.

di Ricerca LBC[2] nella fase di implementazione del dizionario, ma anche da altri utenti, quali ad es. traduttori e specialisti in storia dell'arte italiana.

In considerazione di ciò, i corpora LBC sono stati progettati per avere un ampio uso specialistico, con i seguenti requisiti di base:

a. disporre di materiale linguistico utile per creare e documentare future voci del dizionario multilingue LBC;
b. disporre di materiale linguistico utile per svolgere studi linguistici, letterali o culturali;
c. includere testi che sensibilizzino il grande pubblico al patrimonio culturale di Firenze e della Toscana.

Per la creazione dei corpora LBC, abbiamo considerato testi scritti dal Rinascimento ai giorni nostri ed aventi per soggetto il patrimonio culturale e, in particolare, una visione di Firenze o della Toscana descritta da diversi punti di vista, sia oggettivi che personali. L'ambito geografico cambierà in futuro, per coprire l'intera Italia e altre culture. Tra i testi inseriti all'interno dei nostri corpora, possiamo enumerare:

a. biografie, es. Cellini, *La vita* o Dumas, *Une année à Florence*;
b. testi di arte, es. Vasari, *Le Vite* o Alarcón, *Vidas de los pintores y estatuarios*;
c. romanzi, es. Forster, *A Room with a View*;
d. saggi, es. Feuillet, *L'art italien* o Burckhardt, *Geschichte der Renaissance*;
e. guide turistiche, es. Routard, *GeoGuide*;
f. dizionari specialistici, es. Viollet Le Duc, *Dictionnaire de l'architecture*.

Per ognuno dei testi presenti nei corpora sono memorizzati sia l'anno di redazione che l'anno di pubblicazione, con quest'ultimo considerato secondario rispetto al primo. In effetti, l'anno di redazione sarà il dato di fondamentale importanza per l'estrazione delle informazioni, poiché è in qualche modo rappresentativo delle caratteristiche linguistiche del periodo in esame; infatti i testi sono stati inseriti in tutti i corpora rimanendo fedeli all'edizione utilizzata, senza produrre alcun tipo di modernizzazione.

Va sottolineato che all'interno dei nostri corpora sono stati inclusi sia libri interi che frammenti 'auto-conclusivi' (come un capitolo di un libro, un articolo di una rivista); questa scelta è stata fatta perché in molti casi i contenuti dei libri nella loro interezza non sempre coincidevano con gli interessi del progetto.

È stato svolto un lavoro preliminare per determinare le categorie da utilizzare per catalogare i testi, a seconda dei generi testuali inclusi nei corpora. Innanzitutto abbiamo differenziato i testi *tecnici* da quelli *informativi*, soprattutto in base al target di riferimento, ovvero se il destinatario è o meno uno specialista. La terza categoria individuata è quella dei *dizionari*, che sono da considerarsi testi tecnici nel nostro campo di interesse. Infine, è stato inserito il genere *lette-*

[2] Esempi di utilizzo possono essere trovati in Carpi 2017; Farina, Billero 2018; Billero, Carpi 2018; Lanini, Nicolás 2018; Garzaniti 2020; Farina, Frinz 2020.

rario in quanto si rivela utile per fornire una visione di Firenze e della Toscana, esprimendo come il suo patrimonio è percepito da un punto di vista esterno. In dettaglio, sono state individuate le seguenti categorie e sottocategorie:

1. Divulgativo – Blog
2. Divulgativo – Guida
3. Divulgativo – Ricettario
4. Divulgativo – Rivista
5. Tecnico – Architettura
6. Tecnico – Arte
7. Tecnico – Edilizia
8. Tecnico – Enogastronomia
9. Tecnico – Storia
10. Letterario – Biografico
11. Letterario – Fiction
12. Letterario – Saggistica
13. Dizionario – Monolingue
14. Dizionario – Multilingue

La suddetta classificazione si è rivelata finora efficace, tuttavia occorre sottolineare che è ancora provvisoria, come notano gli stessi membri dell'Unità di Ricerca, e potrebbe essere soggetta a revisioni nelle successive fasi di lavoro.

3. Il flusso di lavoro per la creazione dei corpora

È stato creato un flusso di lavoro adeguato alla generazione dei corpora LBC che ha permesso di eseguire tutti i passaggi necessari ad ottenere corpora consultabili online partendo da un semplice elenco di testi da utilizzare.

Le responsabilità per l'esecuzione di tale flusso di lavoro sono suddivise nel modo seguente:

1. Il processo di selezione, digitalizzazione e raccolta dei testi (con relativi metadati) da inserire nei corpora è curato dai responsabili dei gruppi linguistici;
2. La gestione IT dell'intero progetto LBC era invece sotto la nostra responsabilità.

Le due aree di responsabilità sono essenzialmente indipendenti l'una dall'altra, o più correttamente vengono svolte in momenti diversi. Naturalmente, va notato che la seconda richiede i risultati della prima come dati in input. Le due sezioni seguenti illustreranno le due diverse aree delineate sopra, concentrandosi in particolare sugli aspetti tecnici e informatici.

4. La raccolta dei testi

Il flusso di lavoro relativo al processo di selezione, digitalizzazione e raccolta dei testi (con relativi metadati) da inserire nei corpora è stato condotto dai diversi gruppi linguistici, uno per ogni lingua inclusa nel progetto (Farina 2016).

In primo luogo, ogni gruppo ha compilato una bibliografia di testi (scritti nella relativa lingua di riferimento) aventi come tema il patrimonio culturale di Firenze e della Toscana. In particolare occorre osservare che il gruppo di lavoro di lingua italiana ha condotto una ricerca al fine di creare una bibliografia dei testi più importanti che dovevano essere inclusi nei corpora, i cosiddetti 'testi fondatori'; tali testi sono considerati rilevanti o per la loro terminologia artistica (l'arte italiana è un punto di riferimento per la storia dell'arte) o perché sono classici dell'arte italiana (come Vasari o Michelangelo) che hanno come oggetto Firenze.

Successivamente, ogni gruppo ha individuato le traduzioni di questi testi e ha cercato di identificare altri testi, nella propria lingua, aventi come soggetto l'arte a Firenze o in Toscana, cioè libri di viaggiatori di passaggio a Firenze (es. *Geschichte der Renaissance* di Burckhardt).

Al fine di consentire l'accesso ai file a tutti i membri dell'Unità di Ricerca, si è deciso di utilizzare una soluzione di archiviazione dati su cloud, basata dapprima su cartelle Dropbox, una per ogni lingua coinvolta nel progetto. I responsabili dei diversi gruppi linguistici hanno avuto accesso ai propri dati in modalità di lettura e scrittura, accedendo invece ai dati delle altre lingue in modalità di sola lettura.

Dropbox è stato inizialmente scelto per i suoi vantaggi, come la facilità della configurazione iniziale (da parte del responsabile IT) e la facilità d'uso per tutti i membri del gruppo. In un secondo momento, sia perché questa soluzione si è rivelata troppo limitata in termini di spazio disponibile per le esigenze del gruppo, sia per gestire i nostri documenti 'privatamente', è stato noleggiato un VPS (Virtual Private Server) dedicato al progetto e opportunamente configurato con NextCloud[3], che ha sostituito in breve tempo Dropbox, seppur utilizzando lo stesso criterio di gestione dei file.

I responsabili dei vari gruppi linguistici sono stati incaricati di raccogliere – all'interno della propria cartella cloud – i file contenenti i vari testi utilizzati per costruire il corpus.

Per i testi da inserire nel corpus viene utilizzato Microsoft Word; questo software è stato scelto in considerazione della necessità dei membri dell'Unità di Ricerca di mantenere alcuni degli elementi di formattazione di base come grassetto, corsivo e note a piè di pagina. Word è stato preferito anche ad altre soluzioni (come LibreOffice Writer) a causa dell'esperienza dei membri dei gruppi

[3] Per quanto riguarda la scelta del software, sono stati preferiti strumenti *open source*, per vari motivi: il codice open source è di qualità superiore e rimane affidabile nel tempo; la trasparenza dell'open source e il controllo fornito da tutti i contributori garantisce una maggiore sicurezza; l'open source ha ampie comunità di utenti che consentono un miglioramento continuo delle prestazioni e un'ampia fonte di risposte a tutte le preoccupazioni; le piattaforme open source consentono una maggiore interoperabilità, contro la rigidità di un'unica soluzione che non può soddisfare tutte le esigenze. Tuttavia, tali considerazioni non erano necessarie per quanto riguarda Microsoft Word ed Excel, gli altri strumenti usati dai componenti dei gruppi linguistici, essendo strumenti di provata efficienza e già a disposizione dei ricercatori coinvolti nel progetto.

con Word e perché Dropbox (la soluzione cloud utilizzata inizialmente) ha una versione online del software. Inoltre, dato che il formato *.docx* (il formato di file utilizzato da Word 2007 in poi) è uno standard XML, è stato possibile implementare un software – descritto di seguito – per l'estrazione automatica dei testi dai file e il loro inserimento nei corpora.

Prima di inserire ogni file di testo, i responsabili dei vari gruppi linguistici devono salvare i relativi metadati compilando una riga in un file Microsoft Excel[4], contenuto nella cartella cloud della propria lingua; in questo modo ottengono il nome che deve essere assegnato al file Word (stabilito algoritmicamente dal suddetto file Excel).

L'utilizzo di un nome di file ottenuto algoritmicamente si è reso necessario per facilitare il lavoro dei membri dell'Unità di Ricerca LBC nella fase di consultazione dei corpora; questo nome deve essere 'intelligibile', cioè costituito da una serie di elementi che facilitano la ricerca del testo di interesse, distinguendolo facilmente dagli altri presenti nei corpora. In particolare, alcuni dei metadati inclusi nel foglio Excel contribuiscono a creare il nome univoco assegnato al file contenente il testo; questo nome viene stabilito utilizzando una apposita formula Excel e associando delle abbreviazioni per i vari metadati.

Le informazioni utilizzate per la creazione dei nomi dei file sono:

a. lingua originale del testo; ogni lingua è associata al proprio codice, secondo la codifica ISO 639-2 Alpha-2, che consente di specificare un paese o una regione[5]. Ad esempio, un libro in inglese britannico sarà identificato dal codice *en-GB*.
b. Categoria e sottocategoria del testo come sopra descritto. Ad esempio, un libro classificato come Letterario – Saggistica sarà identificato dall'acronimo: *LET_sag*.
c. Autore del testo. La cartella di lavoro di Excel include un foglio di lavoro degli autori che deve essere compilato dai responsabili dei gruppi linguistici con il nome dell'autore e un'abbreviazione univoca di 3-4 lettere che contribuirà a creare l'identificativo del file di testo utilizzato nei corpora. Ad esempio, tutti i libri scritti da Ruskin avrebbero *Rusk* nel loro identificativo. All'interno del foglio di lavoro dei metadati, le celle della colonna Autore mostrano una casella a discesa riempita automaticamente con i nomi degli autori presenti nel suddetto foglio di lavoro degli autori; in questo modo è possibile ridurre al minimo gli errori di battitura o errate attribuzioni degli acronimi.
d. Titolo del libro (o rivista o miscellanea). Come nel caso degli autori, all'interno della cartella di lavoro di Excel è presente un foglio di lavoro *Titoli* che

[4] La scelta di Microsoft Excel per l'archiviazione dei dati è stata fatta per gli stessi motivi per cui è stato scelto Word, ovvero la sua pervasività, la sua semplicità d'uso e il fatto che il formato *.xlsx* (il formato di file utilizzato da Excel 2007 in poi) è uno standard XML.

[5] In particolare: per il francese si può scegliere tra la variante della Francia e quella del Canada; per l'inglese tra le varianti del Regno Unito e degli USA; per il portoghese tra Portogallo e Brasile; per lo spagnolo tra Spagna e America Latina; per il tedesco tra Germania, Austria o Svizzera.

deve essere compilato dai responsabili dei gruppi linguistici, che inseriscono il titolo del libro e un'abbreviazione univoca di 3-4 lettere che contribuirà a creare l'identificativo del file di testo utilizzato nei corpora. Ad esempio, il libro *Mornings in Florence*, sarà associato all'abbreviazione *Morn*. All'interno del foglio dei metadati, le celle della colonna *Titolo* visualizzano una casella a discesa riempita automaticamente con i titoli presenti nel suddetto foglio di lavoro *Titoli*, come nel caso degli autori.

e. Titolo del frammento (o articolo su rivista o miscellanea). Come nel caso degli autori e dei titoli, all'interno della cartella di lavoro di Excel è presente un foglio di lavoro *Frammenti* che deve essere compilato dai responsabili dei gruppi linguistici, che inseriscono il nome del frammento e un'abbreviazione univoca di 3-4 lettere che contribuirà a creare l'identificativo del file di testo utilizzato nei corpora. Se il testo considerato è stato preso nella sua interezza, deve essere indicato utilizzando l'abbreviazione *integ*.
   Analogamente ai due casi precedenti, all'interno del foglio di lavoro dei metadati, la colonna *Frammenti* mostra una casella a discesa riempita automaticamente con i titoli presenti nel suddetto foglio di lavoro Frammenti, come nel caso degli autori e dei titoli.
f. Anno di redazione del testo. Ad esempio, un libro scritto nel 1875 sarà identificato dall'acronimo *1875r*, dove *r* significa redazione. Per la data di redazione è possibile indicare l'anno preciso (se noto), oppure un'ipotesi dell'anno o del secolo di riferimento.
g. Anno di pubblicazione del testo. Ad esempio, un'edizione del 1881 di un libro utilizzato sarà identificata dall'acronimo *1881p*, dove *p* sta per pubblicato. Per la data di pubblicazione è possibile specificare l'anno preciso (se noto), oppure un'ipotesi dell'anno o del secolo di riferimento.

Pertanto, il libro *Mornings in Florence*, scritto in inglese nel 1875 e pubblicato nell'edizione del 1881, classificato come Letterario – Saggistica, sarà identificato dal nome di file seguente:

en-GB_LET_sag_Rusk_Morn_integ_1875r_1881p

dove:
en-GB indica la lingua inglese nella variante del Regno Unito;
LET_sag indica la categoria;
Rusk è un'abbreviazione per l'autore Ruskin;
Morn è una abbreviazione del titolo del testo;
integ indica che il testo è stato considerato nella sua interezza;
1875r indica che il testo fu redatto nel 1875;
1881p indica che il testo fu pubblicato nel 1881.

Per rendere i testi riconoscibili rispetto a quelli in lingua originale vengono giustapposti alcuni elementi al nome che il testo ha nella versione in lingua originale. A tal fine i responsabili linguistici devono inserire – nella stessa cartel-

la di lavoro Excel fin qui considerata – informazioni sulla versione tradotta del testo; in particolare, i dati essenziali per la creazione del nome sono i seguenti:
a. lingua di traduzione; come per la lingua originale, è possibile specificare una lingua relativa a un paese o una regione;
b. anno di traduzione (come l'anno di redazione del testo in lingua originale);
c. anno di pubblicazione (come quello del testo in lingua originale).

Ad esempio, il caso precedente del testo inglese *Mornings in Florence*, scritto nel 1875 e considerato nell'edizione del 1881, classificato come Letterario – Saggistica e identificato nell'edizione inglese dal nome del file già visto nell'esempio:

en-GB_LET_sag_Rusk_Morn_integ_1875r_1881p

è identificato nella traduzione spagnola dal nome:

en-GB_LET_sag_Rusk_Morn_integ_1875r_1881p_es-ES_1910t_1910p

dove la parte giustapposta include:
es-ES a indicare la lingua spagnola (nella sua variante di Spagna);
1910t a indicare che il testo fu tradotto (*t*) nel 1910;
1910p a indicare che il testo è considerato nella versione pubblicata (*p*) nel 1910.

Oltre ai metadati richiesti per creare il nome del file, è anche possibile inserire le seguenti informazioni nel foglio di lavoro di Excel, quali metadati aggiuntivi per il testo:
- supporto per il quale il testo è stato originariamente creato (carta, CD/DVD, sito web);
- nome e cognome del curatore (se miscellanea);
- numero di volume della rivista o mese (se presente);
- indirizzo web (se presente) da cui proviene il libro o la pagina consultata;
- casa editrice;
- luogo di pubblicazione.

Se il testo è tradotto si possono inserire anche i seguenti metadati:
- titolo del libro o della rivista o miscellanea tradotta;
- titolo del frammento (o dell'articolo se è incluso in una rivista o miscellanea);
- luogo di pubblicazione;
- nome e cognome del traduttore;
- indirizzo web (se presente);
- casa editrice;
- commenti vari.

Come risultato di questo flusso di lavoro, all'interno della nostra installazione di NextCloud saranno presenti diverse cartelle, una per ogni lingua coinvolta nel progetto, contenenti tutti i testi da inserire nei corpora; tale operazione è gestita dalla seconda parte del flusso di lavoro.

5. La creazione dei corpora

La seconda parte del flusso di lavoro, riguardante tutti i processi digitali necessari per raccogliere i testi (con i relativi metadati) e inserirli all'interno di corpora strutturati per l'utilizzo da parte del software di gestione dei corpora, era invece di nostra responsabilità, come sopra indicato. In particolare, a tal fine, sono stati realizzati due appositi software (o, più correttamente, script), utilizzando i linguaggi di programmazione Bash e Python, automatizzando così l'intero flusso di lavoro.

Come accennato in precedenza, dopo una fase iniziale durante la quale Dropbox è stato utilizzato come soluzione cloud, si è deciso di utilizzare NextCloud sul VPS proprietario. Abbiamo noleggiato una macchina virtuale fornita da un servizio online, su cui era installata la distribuzione Debian GNU/Linux; tale VPS non solo ha permesso la creazione di un nostro cloud, ma ha anche consentito l'esecuzione del suddetto script e la possibilità di rendere disponibili i vari corpora tramite un servizio web.

La seconda parte del flusso di lavoro viene coordinata da uno script Bash[6], il quale richiama altri script o software in un ordine ben determinato, ognuno responsabile dell'esecuzione di uno dei tre passaggi che trasformano una raccolta di testi in formato Microsoft Word (e i loro metadati in formato Microsoft Excel) in un insieme di corpora, ovvero:

1. riconciliare ogni testo con i suoi metadati e raccoglierlo in un unico documento con tutti gli altri testi;
2. tokenizzare e lemmatizzare tutti i testi;
3. compilare il corpus.

Il primo passaggio viene eseguito da un apposito script Python[7], il quale legge il foglio Excel contenente l'elenco dei testi da inserire nel corpus e dei relativi metadati e per ogni riga di informazione in esso presente individua i vari metadati e il nome assegnato al relativo file Word (generato automaticamente da una apposita formula, come sopra indicato); con queste informazioni, lo script può quindi cercare nella cartella cloud il file Word e aprirlo.

A partire dal 2007, ogni documento Microsoft Word viene archiviato utilizzando il formato Office Open XML (standard internazionale ISO/IEC DIS 29500). Questo tipo di documento è costituito da un file compresso in cui sono presenti vari documenti XML che specificano i molteplici elementi del documen-

[6] Bash è una shell e un processore di comandi per i sistemi operativi GNU/Linux che è ampiamente utilizzata come shell di login predefinita per la maggior parte delle distribuzioni Linux e può leggere ed eseguire comandi da un file, il cosiddetto script di shell.

[7] Alla base della scelta di utilizzare Python come linguaggio di programmazione ci sono numerosi motivi: si tratta di un linguaggio completamente gratuito, utilizzabile su varie piattaforme (Linux, Windows ecc.), già installato di default in molte distribuzioni Linux (inclusa Debian), facile da usare, dotato di librerie per gestire agevolmente documenti in formato Microsoft Word ed Excel e più in generale in formato testo e XML, con risultati altamente performanti.

to Word, come il testo in esso contenuto, la sua formattazione, gli oggetti inseriti, ecc. Grazie alla conoscenza di questo standard, lo script Python apre il documento Word, ne legge il testo e lo copia come contenuto in un opportuno nodo di un documento XML, senza alcun elemento di formattazione; invece gli attributi di tale nodo vengono impostati utilizzando tutti i relativi metadati contenuti nel file Excel. Come risultato del processo, lo script Python genera in output – per ogni lingua – un documento XML contenente tutti i testi con i relativi metadati; quel documento verrà quindi utilizzato come input per il passo seguente.

Tuttavia, poiché la lettura di file Word ed Excel richiede una certa quantità di tempo per elaborare il contenuto che diventa eccessivo quando centinaia o migliaia di testi vengono inseriti nel corpus, è stata implementata una cache per accelerare la ricompilazione del corpus. Tale cache si basa sul principio che una volta che i testi e i relativi metadati vengono memorizzati, saranno modificati molto raramente e sostanzialmente solo per correggere errori materiali riscontrati successivamente all'inserimento del testo e dei relativi metadati. In questo modo diventa più veloce controllare – ogni volta che il corpus viene ricompilato – se sono state apportate modifiche; l'introduzione della cache ci ha permesso di ridurre il tempo di ricompilazione di un corpus da poche ore a circa dieci minuti, per un corpus di un milione di parole. Nonostante aggiunga complessità al flusso di lavoro, l'utilizzo della cache ne riduce i tempi di esecuzione in base al principio che nuovi testi e relativi metadati possono essere continuamente inseriti, mentre raramente accade che i testi esistenti vengano modificati (Billero 2020).

Il secondo passaggio dello script Bash consente la tokenizzazione e quindi la lemmatizzazione del testo. La procedura di tokenizzazione trasforma il file XML ottenuto come output del passaggio precedente nel cosiddetto file verticale, ovvero un file strutturato in modo che su ogni riga sia presente un solo token (una singola parola o un unico simbolo grammaticale). Questa modalità di memorizzazione dei dati è funzionale alla procedura di lemmatizzazione dove il cosiddetto Part-Of-Speech (o più brevemente POS) e il lemma vengono memorizzati all'interno del file verticale accanto a ciascun token.

Ad esempio, la frase «This book is easy to read» verrà memorizzata come segue:

Word	**POS**	**Lemma**
This	DT	this
book	NN	book
is	VBZ	be
easy	JJ	easy
to	TO	to
read	VB	read
.	SENT	.

Per questa seconda fase, abbiamo scelto di utilizzare TreeTagger, un tagger per parti del discorso indipendente dalla lingua (Schmid 1994, 1995), uno dei software più importanti della sua categoria. Pertanto, lo script Bash richiama

TreeTagger, utilizzando come input il file XML ottenuto come risultato del precedente passaggio contenente tutti i testi con i relativi metadati.

Infine, il terzo passaggio consiste nella compilazione del corpus, ovvero nel memorizzarlo in un formato facilmente accessibile agli utenti finali. Dato il numero diversificato di utenti, la consultazione delle informazioni dovrebbe essere quanto più ampia possibile; pertanto, dovrebbe essere utilizzato un buon software di analisi e gestione dei corpora. Esistono due diverse tipologie di tale software: quelli che possono essere utilizzati solo su un computer desktop e quelli disponibili come servizio web. Per soddisfare gli scopi del progetto, abbiamo scelto un software che potesse essere utilizzato come servizio web.

Una ricerca dello stato dell'arte ha identificato l'esistenza di tre principali software all'interno della famiglia sopra descritta: CWB (IMS Open Corpus WorkBench), INL BlackLab e SketchEngine. Va notato che mentre i primi due software sono disponibili come open source, SketchEngine è solitamente disponibile solo a pagamento, ma attualmente è gratuito per i membri degli istituti di ricerca in Europa. Tuttavia, è stata rilasciata anche una versione gratuita ridotta, chiamata NoSketchEngine, priva delle caratteristiche più interessanti per lo studio lessicografico incluse nella versione commerciale, come la funzione Word Sketch e il *thesaurus*.

Tutti e tre i software sopra menzionati hanno una serie di funzionalità di base che sono di fondamentale importanza nella ricerca di corrispondenze, tra cui: la possibilità di effettuare ricerche semplici, ricerche utilizzando espressioni regolari o Context Query Language[8], oltre alla possibilità di utilizzare lemmatizzatori come TreeTagger, Freeling, RfTagger. Inoltre, tutti e tre i software dispongono di una speciale API (Application Programming Interface) che consente loro di eseguire operazioni come l'interrogazione o l'inserimento di dati tramite software di terze parti appositamente implementati per le proprie esigenze.

Mentre CWB e SketchEngine dispongono attualmente di un'eccellente documentazione e di una solida comunità di riferimento, lo stesso non si può dire per BlackLab, che mostra ancora debolezze in questi aspetti.

SketchEngine offre anche la funzionalità di Word Sketch, utile per la creazione di elenchi di parole di termini specifici delle diverse arti e che può essere utilizzata per mettere in relazione parole che si riferiscono allo stesso oggetto artistico, per delimitare campi di conoscenza o di riferimento come monumenti, o distinguere tra oggetti e persone e disambiguare i nomi propri che si riferiscono a queste due categorie. Tuttavia, questa funzione non è disponibile nella versione gratuita, ovvero NoSketchEngine.

Alla fine, abbiamo scelto per la nostra ricerca il software di gestione dei corpora SketchEngine. Abbiamo anche provveduto ad effettuare un'installazione di NoSketchEngine sul nostro VPS rendendo così i nostri corpora disponibili

[8] Context Query Language (CQL in breve) è un linguaggio di query originariamente creato per CWB e successivamente implementato negli altri software menzionati.

online[9] agli utenti finali. Attualmente, i corpora contengono otto milione di parole, divise per lingua, come indicato nella tabella seguente.

Tabella 1. Numero attuale di parole per lingua.

Lingua	Parole
Francese	3.165.000
Russo	1.900.000
Inglese	1.036.000
Spagnolo	1.020.000
Tedesco	1.018.000
Italiano	1.009.000

L'utilizzo del software di gestione dei corpora NoSketchEngine consente agli utenti di eseguire vari tipi di ricerche. Un primo tipo di ricerca è la ricerca per parola, in cui gli utenti cercano una parola scritta nel modo in cui si aspettano di trovarla nei corpora; altrimenti è anche possibile cercare il lemma, dove NoSketchEngine fornirà i risultati come un elenco di occorrenze di quel lemma nelle varie forme flesse presenti all'interno del corpus. È anche possibile utilizzare il linguaggio di query CQL per eseguire ricerche più selettive, nonché la possibilità di specificare un particolare POS (Part-of-Speech) da cercare. Ad esempio, l'utente finale può cercare aggettivi vicino alla parola opera o cercare verbi dopo il nome di un pittore, ecc.

Per quanto riguarda i filtri di ricerca, è possibile utilizzare molti dei metadati per limitare l'ambito delle ricerche, dall'intero corpus a una sezione; ad esempio, gli utenti finali possono filtrare per lingua, autore, titolo dell'opera, categoria, sottocategoria e così via.

Di conseguenza, possiamo affermare che NoSketchEngine è un potente strumento per la ricerca all'interno dei nostri corpora utilizzando ricerche generiche o dettagliate e che soddisfa i requisiti del progetto LBC e le diverse esigenze degli utenti specificate in precedenza.

Conclusioni

La realizzazione di questa prima fase dei nostri corpora è da ritenere soddisfacente in quanto ha creato le basi necessarie per i primi lavori e per le ricerche del nostro gruppo (Carpi 2017; Farina, Billero 2018; Billero, Carpi 2018; Lanini, Nicolás 2018; Farina, Frinz 2020; Garzaniti 2020), dimostrando il fondamentale apporto dell'Informatica Umanistica al progetto Lessico dei Beni Culturali.

Senza dubbio quanto realizzato finora si tratta solo di un primo passo verso il raggiungimento di altri prodotti utili alla ricerca nell'ambito della lessicogra-

[9] Consultabile su <http://corpora.lessicobeniculturali.net/>.

fia dell'arte; ad esempio, in futuro dovrebbe essere introdotta la possibilità di etichettare i testi a un livello molto più avanzato rispetto a quanto fatto finora, utilizzando standard universalmente riconosciuti. In particolare, poiché i nostri corpora sono strettamente legati ai nomi propri di persona e luogo, è ipotizzabile che ISNI possa essere utilizzato applicando un'etichetta a ciascuno dei nostri nomi per raccogliere i nomi propri di tutte le varianti in tutte le lingue presenti nei corpora.

Bibliografia

Billero R. 2020, *Cultural Heritage Lexicon: A Case Study*, in Pano Alamán A., Zotti V. (eds.), *The language of art and cultural heritage: a plurilingual and digital perspective*, Cambridge Scholars Publishing, Newcastle upon Tyne: 86-103.

Billero R., Carpi E. 2018, *Corpora e terminologia artistica: il caso del corpus spagnolo LBC*, «CHIMERA Romance Corpora and Linguistic Studies», 5(1): 85-91.

Billero R., Nicolás Martínez M.C. 2017, *Nuove risorse per la ricerca del lessico del patrimonio culturale: corpora multilingue LBC*, «CHIMERA Romance Corpora and Linguistic Studies», 4(2): 203-216.

Carpi E. 2017, *El lenguaje para fines artísticos: traducciones de tondo al español*, in Curado A. (ed.), *EPiC Series in Language and Linguistics, 3, LSP in Multi-disciplinary contexts of Teaching and Research. Papers from the 16th International AELFE Conference*: 79-84, <https://doi.org/10.29007/wx3m>.

CWB, <http://cwb.sourceforge.net/>, (10/2020).

Farina A. 2016, *Le portail lexicographique du Lessico plurilingue dei Beni Culturali, outil pour le professionnel, instrument de divulgation du savoir patrimonial et atelier didactique*, «Publif@rum», 24, <http://www.farum.it/publifarum/ezine_articles.php?art_id=335>.

Farina A., Billero R. 2018, *Comparaison de corpus de langue «naturelle» et de langue «de traduction»: les bases de données textuelles LBC, un outil essentiel pour la création de fiches lexicographiques bilingues* in Iezzi D.F., Celardo L., Misuraca M. (eds.), *JADT 2018 – International Conference on Statistical Analysis of Textual Data, Roma, 12-15 giugno 2018*, UniversItalia, Roma: 108-116.

Farina A., Flinz C. 2020, *Analisi linguistica comparativa dei corpora LBC. La visione del patrimonio fiorentino francese e tedesco: l'esempio del Duomo*, in Farina A., Funari F. (eds), *Past in Present / Le passé dans le présent / Il passato nel presente*, FUP, Firenze: 75-98.

Garzaniti M. 2020, *Il termino russo friag e le sue radici nelle relazioni culturali e artistiche fra la Russia e l'Italia*, in Pano Alamán A., Zotti V. (eds.), *The language of art and culture heritage: a plurilingual and digital perspective,* Cambridge Scholars Publishing, Newcastle upon Tyne: 104-119.

ISNI, <http://www.isni.org/>, (10/2020).

Lanini L., Nicolás Martínez M.C. 2018, *Verso un dizionario corpus-based del lessico dei beni culturali: procedure di estrazione del lemmario* in Iezzi D.F., Celardo L., Misuraca M. (eds.), *JADT 2018 – International Conference on Statistical Analysis of Textual Data, Roma, 12-15 giugno 2018*, UniversItalia, Roma: 411-418.

NextCloud, <https://www.nextcloud.com/>, (10/2020).

Numerico T., Vespignani A. 2003, *Informatica per le scienze umanistiche*, il Mulino, Bologna.

Schmid H. 1995, *Improvements in Part-of-Speech Tagging with an Application to German* in *Proceedings of the ACL SIGDAT-Workshop*, Dublin.
Schmid H. 1994, *Probabilistic Part-of-Speech Tagging Using Decision Trees* in *Proceedings of International Conference on New Methods in Language Processing*, Manchester.
SketchEngine, <https://www.sketchengine.eu/>, (10/2020).
Svensson P. 2010, *Landscape of Digital Humanities*, «Digital Humanities Quarterly», IV (1), <http://digitalhumanities.org/dhq/vol/4/1/000080/000080.html> (11/2020).
Zotti, V., Pano Alamán A. 2017, *Introduzione*, in Zotti V., Pano Alamán A. (eds.), *Informatica umanistica. Risorse e strumenti per lo studio del lessico dei beni culturali*, FUP, Firenze: 7-15.