BIBLIO THECAE .it

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA
DIPARTIMENTO DI BENI CULTURALI

Nello Cristianini - Thomas Lansdall-Welfare
Gaetano Dato - Marco Menato*

Contea Principesca di Gorizia e Gradisca, 1873-1914: creazione e analisi di un corpus digitale di periodici nel loro contesto storico

1. *Introduzione*

L'uso di strumenti informatici negli studi umanistici risale almeno agli anni Quaranta del Novecento, con il lavoro di padre Roberto Busa, che usò le prime macchine IBM per creare il suo Index Thomisticus; lungo questa strada padre Busa sviluppò una serie di metodi e domande che ancora oggi sono rilevanti.[1] Già nel 1966 venne fondata la rivista «Computers and the Humanities» dedicata alla «application of computer methods to humanities scholarship».[2] Nell'ultima decade l'espressione 'umanistica digitale', ispirata alla denominazione inglese *digital humanities*, è emersa come il termine più diffuso in Italia per qualificare questo campo di studi, che pure include, all'inverso, l'applicazione di conoscenze e metodi degli umanisti allo studio dell'informatica.

La promessa delle biblioteche digitali non è solo quella di miglio-

* NC, TLW e GD sono stati supportati da ThinkBIG. Ringraziamo anche FMP digitization team alla British Library.

1 Busa 1980.

2 <http://www.historyofinformation.com/detail.php?entryid=4214>.

rare la conservazione dei documenti e facilitarne l'accesso da parte degli utenti. Il nuovo medium di rappresentazione ci consente di porre nuove domande di ricerca, alle quali è possibile rispondere, per esempio, esaminando le relazioni statistiche fra i contenuti di migliaia di documenti, un tipo di analisi accessibile con relativa velocità solo tramite l'uso dell'informatica e che invece richiederebbe un tempo enorme senza le moderne tecnologie.

Per giungere a questo punto, si deve poter passare dal supporto analogico (carta o microfilm) a quello digitale, poi trasformare le immagini in testo digitale editabile, e infine analizzare statisticamente quei testi.

L'introduzione di tecniche di acquisizione di massa di testi digitali, e quelle di analisi su larga scala che vanno talvolta sotto il nome di Big Data, si sono combinate ripetutamente nella creazione di sottocampi, per esempio quello chiamato *Culturomics*,[3] il cui obiettivo è di individuare e studiare proprietà statistiche macroscopiche di prodotti culturali quali i database di testi. In questa direzione è l'idea del 'macroscopio', proposta da Graham-Milligan-Weingart,[4] che si propone di usare metodi informatici per scoprire regolarità, *network* o anomalie in *dataset* prodotti in ambito letterario, storico, archivistico, artistico o nell'incrocio fra diverse discipline umanistiche e sociali.

Del resto già padre Busa aveva compreso che, lungi dall'essere solo una automazione di un processo manuale, questa metodologia avrebbe consentito nuove domande e metodi di ricerca.

Franco Moretti ha dato un nome a un nuovo modo di usare testi digitali senza sottoporli allo scrutinio tipico della lettura critica: *Distant Reading*. Il suo esempio principale - il cambiamento nella lunghezza e la forma dei titoli dei romanzi durante il XIX secolo - ci mostra chiaramente come la macchina possa aiutare a trovare delle relazioni - attese o inattese che siano - ma anche come sia sempre l'interpretazione dello studioso a conferire loro un significato.[5]

3 Michel et al. 2011.
4 Graham Milligan Weingart 2015.
5 Moretti 2013.

Roberto Franzosi, con la sua *Quantitative Narrative Analysis*, ha dimostrato come i giornali del passato possano fornirci un nuovo modo di studiare la cultura e i suoi cambiamenti, un modo che in un articolo più recente ha definito: «una terza via verso il passato», alternativa sia alla lettura critica che all'uso di dati sociologici.[6]

Un uso di tecniche di Intelligenza Artificiale ispirato a tali approcci è stato recentemente applicato allo studio di giornali storici.[7] L'idea alla base di questi lavori è quella di individuare eventi, transizioni e continuità macroscopiche, su una scala che non sarebbe accessibile a studiosi individuali, consentendo di accedere ai contenuti di migliaia di documenti in breve tempo. L'informazione estratta da questo metodo non è contenuta in alcun specifico testo e sfugge alla comune lettura, ma emerge dalle relazioni statistiche, le quali possono evidenziare differenze temporali o geografiche, oppure la salienza di certi termini o concetti all'interno del corpus. È tale approccio a rendere necessario l'uso di un 'macroscopio', per usare il termine di Graham, Milligan e Weingart.

L'esecuzione di studi di informatica umanistica incontra inevitabilmente alcune criticità. Uno - fondamentale - è la differenza culturale tra i due settori che ne forniscono le basi culturali, che non si risolve solo in una questione terminologica, ma di valori. In particolare, nelle discipline umanistiche è essenziale che ogni osservazione venga interpretata all'interno del proprio contesto, una procedura che non è automatizzabile, e che richiede una comprensione storica dei dati, della loro origine, e della loro analisi.

In ogni caso, la lezione dell'esempio di Moretti sulla lunghezza dei titoli è chiara: ogni studio di *digital humanities* non può prescindere da una comprensione del contesto culturale che ha prodotto le fonti digitalizzate. La terza via di accesso al passato, proposta da Franzosi, ha reso questa lezione ancora più attuale, ponendola come soluzione

6 Franzosi 2017.

7 Lansdall-Welfare et al 2017a; Lansdall-Welfare et al, 2017b; Dzogang et al. 2017; Flaounas et al. 2010; Dzogang et al. 2018.

alla disputa metodologica tra Elton e Fogel, fautori l'uno di un percorso narrativo e l'altro di un percorso scientifico / quantitativo.[8]

Kirsch è ancora più diretto,[9] nella sua critica a *Uncharted*,[10] il volume scritto dai creatori del *Google Ngram Viewer*, il servizio della nota società informatica che consente gratuitamente di verificare la frequenza di una o più parole nel tempo fra i libri digitalizzati da Google. In *Uncharted* i suoi autori affermano che la frequenza della citazione di ciascun anno nel database è in assoluto più alta in quello stesso anno così che, ad esempio, ci sono più volumi che contengono "1950" fra i libri pubblicati in quel medesimo anno che in qualsiasi altro periodo; su tale base gli autori teorizzano l'esistenza di una *forgetting curve*, il fatto che gli eventi che avvengono in un determinato anno, vengano ricordati maggiormente nel corso dell'anno stesso, e poi sempre più dimenticati man mano che ce se ne allontana. Kirsch osserva però come l'inevitabile presenza nel colophon dell'anno nella dichiarazione di copyright può essere sufficiente a giustificare quel picco, fatto che mina alla base ogni riflessione in *Uncharted* sulla *learning curve*.

La conclusione, ancora una volta, è che senza meta-informazione (informazione sui dati e la loro origine) non si può fare digital humanities.

1.1. *Dalle biblioteche digitali all'umanistica digitale*

In questo articolo presentiamo il risultato di un ciclo completo di applicazione dei metodi dell'umanistica digitale, dalla creazione del corpus, ovvero digitalizzazione dei supporti analogici e estrazione dei testi digitali, fino all'analisi statistica, la visualizzazione dei dati riguardanti alcuni specifici temi e la loro interpretazione sotto il profilo storiografico. Infine, discutiamo le limitazioni di questo approccio metodologico e tracciamo possibili futuri percorsi di studio.

8 Franzosi 2017.
9 Kirsch 2014.
10 Aiden - Michel 2013.

Uno degli scopi di questo articolo è di mostrare come si possa intraprendere uno studio statistico di larga scala di giornali storici. Per questo, molti dei segnali analizzati hanno la funzione di verificare e calibrare il metodo, e si riferiscono a eventi ben noti e facilmente circoscrivibili. Successivamente applichiamo il metodo anche all'analisi di fenomeni più complessi e dai contorni più sfumati.

Il corpus in questione è formato dai giornali italiani e sloveni stampati a Gorizia tra il 1873 e il 1914. Tra i casi affrontati con maggiore dettaglio, mostriamo come eventi importanti - quali il passaggio della cometa di Halley - siano chiaramente individuabili da un'analisi statistica, ma altri eventi siano meno facili da individuare, quali i dettagli delle trasformazioni dei rapporti fra italiani e sloveni nel Goriziano. Questo articolo non compie un'analisi storica completa del periodo, ma si limita a mostrare alcuni esempi del tipo di informazioni che si possono ottenere dal corpus rispetto ai maggiori temi storiografici riguardanti l'area nord adriatica nella *belle epoque*.

2. *Gorizia e la sua stampa nel 1800*

2.1. *Gorizia, Gorica, Görz - La Contea Principesca di Gorizia e Gradisca*

Da secoli, nei territori a nord dell'Adriatico, si incontrano il mondo latino, lo slavo e il germanico. Più o meno al centro di questa area, sulle rive del fiume Isonzo, sorge Gorizia.

La città e il suo circondario sono abitati da popolazioni di lingua italiana e slovena, e fra il 1500 e il primo conflitto mondiale esse erano parte dei domini asburgici, trovandosi al confine occidentale di quest'ultimo per gran parte della sua esistenza. Nel corso del XIX secolo il confine si mosse in diverse occasioni e lo stesso avvenne dopo ciascuna delle guerre mondiali. Oggi l'Isontino è diviso fra Italia e Slovenia, con l'area urbana prevalentemente all'interno dei confini ita-

liani. Del resto Gorizia,[11] fondata nell'alto medioevo, è stata per gran parte della sua intera storia un città di confine, con una composizione etnica variabile nel corso del tempo. La attraversa il fiume Isonzo, che gli sloveni chiamano *Soča* e che scorre per 136 km dalle Alpi al mare Adriatico collegando il monte Triglav, luogo centrale nell'identità nazionale slovena, ai territori costieri che tradizionalmente furono sotto l'influenza della Repubblica di San Marco. Gorizia rappresenta dunque uno dei luoghi centrali nelle relazioni fra italiani e sloveni.

Durante il periodo abbracciato da questa ricerca Gorizia e l'intero corso dell'Isonzo erano parte della Principesca Contea di Gorizia e Gradisca,[12] area a sua volta incorporata al Litorale Austriaco,[13] un'entità amministrativa creata nel 1849 e che comprendeva anche Trieste e la penisola istriana.

Il centro della città di Gorizia è stato principalmente abitato da una popolazione di lingua italiana, specialmente dopo il tardo medioevo, mentre nelle campagne vi si sono stabiliti in maggioranza gli sloveni, che chiamano il centro abitato *Gorica*. Sino al primo conflitto mondiale, inoltre, risiedeva nel capoluogo isontino anche un certo numero di austriaci, nella cui lingua la città è denominata *Görz*. La componente italiana era a sua volta suddivisa fra coloro che parlavano il friulano e quanti invece si esprimevano in un dialetto di impronta essenzialmente veneta, rispettivamente situati in maggior misura nelle pianure a ovest di Gorizia e a sud nei territori costieri. Peraltro, i confini storici del Friuli sono proprio l'Isonzo a oriente e il Tagliamento a occidente. Il nome di questa regione, così come quelli del fiume I*sonzo-Soča*, e

[11] Il primo documento storico che si sia mai riferito a Gorizia è datato 28 aprile 1001. È un testo latino che cita un 'villaggio che nella lingua degli slavi si chiama Gorizia' (*unius ville que Sclavorum lingua vocatur Goriza*) (Marušič 2005, p. 7; Cavazza 2001, p. 3). In sloveno, 'Gorica' (pron. Goriza) significa letteralmente 'piccolo monte'.

[12] Slo. *Poknežena grofija Goriška in Gradiščanska*, ted. *Gefürstete Grafschaft Görz und Gradisca*.

[13] Slo. *Avstrijsko Primorje*, ted. *Österreichisches Küstenland*.

della regione del Litorale-*Primorje*, hanno ispirato la denominazione dei principali periodici dell'area isontina.

Durante il periodo abbracciato da questa ricerca la maggioranza italiana del centro abitato di Gorizia, come lingua d'uso quotidiano, parlava di fatto il friulano se erano forti i legami con le aree rurali in direzione del Friuli o un dialetto giuliano di area veneta se era più forte l'identità cittadina. Essi erano spesso membri della classe media o dell'aristocrazia. La campagna circostante vedeva invece una prevalenza di contadini sloveni che si esprimevano in dialetto sloveno. Sia italiani che sloveni comprendevano comunque le rispettive lingue standard ed erano infatti queste ultime a essere usate nei periodici interessati da questa ricerca.

Molti funzionari pubblici del governo centrale, così come alcuni imprenditori, erano invece austriaci. Nel corso dell'Ottocento, numerosi turisti provenienti dalle più fredde regioni dell'Austria e del resto dell'impero passavano lunghi periodi a Gorizia, a volte scegliendola come luogo d'elezione per vivere in serenità gli ultimi anni della propria esistenza; Gorizia godeva infatti della fama di luogo dal clima mite ed estremamente salubre, che si diffuse specialmente dopo la guerra del 1866 e la perdita del Lombardo-Veneto da parte di Vienna.

Un'accresciuta urbanizzazione della popolazione rurale e lo sviluppo economico del tardo Ottocento determinò una notevole variazione dell'antico equilibrio etnico-sociale, che spinse le identità nazionali italiana e slovena a un più acceso livello di scontro. In base ai censimenti asburgici, il numero degli abitanti di Gorizia crebbe dai 16.659 del 1869 ai 29.291 del 1910. L'aumento accelerò proprio nell'ultimo decennio, dato che furono conteggiati 23.765 abitanti nel 1900. Nello stesso torno di tempo coloro che dichiaravano di usare l'italiano come lingua d'uso calò dai 16.112 individui nel 1900, a 14.812 dieci anni più tardi; al contempo coloro che dichiaravano di usare lo sloveno passarono da 4.754 a 10.790. Nel complesso della contea coloro che si esprimevano in sloveno erano invece sempre stati maggioranza; nel 1910 erano 154.564, mentre coloro che affermavano di parlare princi-

palmente in italiano erano 90.146.[14]

Al contempo, l'estensione del diritto di voto concesse una maggiore influenza alle classi sociali inferiori e quindi alla popolazione delle campagne nella politica rappresentativa. Mentre le autorità asburgiche furono via via più tolleranti delle espressioni della vita culturale nazionale di italiani e sloveni, la coesione etnica dell'Austria-Ungheria entrò in una progressiva crisi che segnò le ultime decadi della sua esistenza e riguardò non soltanto Gorizia, ma tutte le regioni multietniche dell'impero. Ciononostante, nell'Isontino il conflitto nazionale in età asburgica non registrò mai episodi di particolare violenza o efferatezza fino alla guerra, mantenendosi nei limiti di un confronto che, per quanto aspro, rimase all'interno delle dinamiche della politica e della cultura.[15]

2.2. *La stampa locale*

La Gazzetta Goriziana fu il primo periodico pubblicato a Gorizia, un settimanale venduto nel biennio 1774-75. Negli anni successivi sorsero altri giornali locali che però furono prevalentemente di breve durata. Ci furono inoltre dei tentativi di pubblicare dei periodici in tedesco ma, dovendo contare su un troppo ristretto numero di lettori, non riuscirono mai a sopravvivere a lungo. La fioritura della stampa goriziana avvenne sostanzialmente nella seconda metà dell'Ottocento grazie a un'accresciuta scolarizzazione e una maggiore complessità della struttura sociale della regione, pertanto sia la popolazione di lingua italiana che quella di lingua slovena costituirono una più solida base di consumatori di carta stampata. Ciascun gruppo nazionale disponeva allora di propri quotidiani, oltre che di rivendite specializzate nelle pubblicazioni dei rispettivi gruppi nazionali.[16] Tale processo fu particolarmente favorito dalle riforme costituzionali del dicembre 1867, le

14 Fabi 1991, p. 252-254; Marušič 2005, p. 45-46; Kalc 2013, p. 684-701.
15 Fabi 1991, p. 9-10, 32-35; Marušič 2005, p. 7-12; Ferrari 2002, p. 313-318.
16 Gorian 2010; Feresin 2007, p. 14-17; De Grassi 1982, p. 55.

quali modificarono la legislazione riguardante la libertà di espressione in modo da consentire, molto più facilmente che in precedenza, la fondazione di nuovi periodici, che da allora poterono molto più liberamente rappresentare le differenti prospettive etniche e politiche che si sviluppavano nell'Impero Asburgico e quindi nell'area di Gorizia.[17]

Il panorama politico lungo tutta la storia della zona di confine nord adriatica fra Otto e Novecento fu peraltro connotata dall'incrocio di due aspetti fondamentali: ideologia e nazione. Queste sono le chiavi per interpretare la politica della regione sino al termine della Guerra Fredda, senza dimenticare le specificità che caratterizzarono di volta in volta le varie epoche e le singole unità territoriali.

A Gorizia, negli anni Settanta dell'Ottocento le due principali ideologie presenti nelle istituzioni e nei periodici locali erano quella liberale e il conservatorismo cattolico. In seguito alla enciclica Rerum Novarum emessa da Papa Leone XIII nel 1891, gran parte dei cattolici abbracciarono il pensiero cristiano-sociale, che spingeva i credenti a un maggiore impegno politico a favore dei più bassi strati sociali, fatto che al contempo consentì di controbilanciare l'attrazione del socialismo. Il movimento cristiano-sociale ottenne presto un notevole successo nella contea di Gorizia, mentre, all'inizio del Novecento, i socialisti riuscirono a mettere radici soprattutto nel monfalconese, grazie alla presenza dei cantieri navali e di altre strutture industriali.

Le identità nazionali seguivano invece la frattura etnica che divideva i maggiori gruppi presenti a Gorizia, innanzitutto italiani e sloveni, ma anche gli austro-tedeschi avevano un proprio contenuto spazio di rappresentanza. I due assi ideologico-nazionali della politica goriziana costituivano una combinazione di almeno quattro punti cardinali che soddisfacevano i maggiori orientamenti della politica locale, cui si affiancava il ristretto gruppo di rappresentanti politici austriaci.

Le identità nazionali influenzarono un ulteriore aspetto alquanto significativo per la politica locale nel periodo considerato: la fedeltà

17 Ferrari 2002, p. 342; Horel 2015, p. 88-90.

alla casa d'Austria e all'unità dell'impero asburgico. Tanto più ci si accosta alla Prima guerra mondiale, tanto più le nazioni non germaniche dell'impero ricercavano maggiori espressioni di autonomia, se non vere e proprie proprie forme di separatismo, come quello rivendicato dagli irredentisti italiani..[18]

La stampa, libera di esprimersi fin tanto che le autorità non ritenessero che stesse turbando troppo il precario equilibrio etnico, che fosse offesa la monarchia o che fosse messo in discussione l'assetto territoriale dello stato, diede voce ai diversi attori sulla scena della politica goriziana. I giornali che durarono più a lungo presero posizione rispetto ai punti cardinali della politica isontina: liberali e cattolici, italiani e sloveni.

Passiamo dunque a descrivere i cinque più duraturi e influenti giornali goriziani, che furono in vendita per gran parte del periodo che va dal 1873 al 1914. Essi ricoprirono, in forma non del tutto omogenea e coerente, le quattro principali posizioni politiche appena descritte, come mostrato nella Tabella 1; gli anni di pubblicazione di ciascuna testata sono invece riassunti nella Figura 2.

	Liberali	Cattolici
Italiani	Corriere Friulano / Corriere di Gorizia	Eco del Litorale
Sloveni	Soča (1871-1899) Soča (1899-1914)	Soča (1871-1899) Gorica Primorski List

Tabella 1 - Le posizioni politiche dei cinque giornali interessati dalla ricerca.

[18] Fabi 1991, p. 12-46; Ferrari 2002, p. 340-75; Marušič 2005, p. 239-344, Kacin-Wohinz - Troha 2000, p. 69-79. L'ultimo testo citato rimanda all'unica edizione a stampa della relazione finale della "Commissione storico-culturale italo-slovena", la quale operò per definire una comune interpretazione circa le relazioni fra sloveni e italiani tra 1880 e 1956. La relazione fu redatta in italiano, sloveno e inglese. La commissione, che si incontrò fra il 1993 e il 2000 prese le mosse da un'iniziativa congiunta delle autorità statali dei due paesi, e vi parteciparono alcuni dei maggiori storici italiani e sloveni esperti delle questioni prese in esame.

Corriere di Gorizia / Il Corriere Friulano [CFG]. Nonostante i liberali italiani fossero riusciti a mantenere sempre la maggioranza in consiglio comunale e ad amministrare la città, le testate che si riferivano alla loro area politica subirono sempre una certa opposizione da parte delle autorità del governo centrale, a causa della propria linea circa gli sloveni, le relazioni politiche fra Roma e Vienna e per il sostegno al processo unitario italiano.

L'Isonzo fu il primo giornale dei liberali italiani a uscire a Gorizia dopo le riforme asburgiche del 1867 e fu stampato dall'ottobre 1871 al 1880, quando le forze governative ne imposero la chiusura. Alcuni anni dopo i suoi editori, capeggiati dall'intellettuale italiana Carolina Luzzatto, fondarono al suo posto nel 1883 il quotidiano *Il Corriere di Gorizia*. Un ulteriore intervento delle autorità nel 1899 spinse gli editori a cambiare denominazione alla testata, che prese il nome di *Il Friuli Orientale.* Nel 1901 il gruppo capeggiato da Luzzatto iniziò anche a pubblicare un altro giornale, *Il Corriere Friulano*, e dopo alcuni mesi di coesistenza dei due quotidiani, *Il Friuli Orientale* venne chiuso lasciando dunque *Il Corriere Friulano* a rappresentare il punto di vista dei liberali italiani fino al 1914, allorché a causa del conflitto la testata terminò le pubblicazioni, in seguito all'entrata in vigore delle leggi dello stato di guerra, che limitavano nuovamente la libertà di stampa.[19]

Il nostro corpus comprende l'intera serie de *Il Corriere di Gorizia* e *Il Corriere Friulano* ma non *L'Isonzo* e *Il Friuli Orientale*, abbracciando quindi gli anni che vanno dal 1883 al 1914, all'esclusione di 16 mesi tra il gennaio 1900 e l'aprile 1901. I numeri pubblicati in questo breve torno di tempo fanno parte di un'altra collezione di microfilm conservata dalla Biblioteca Statale Isontina (BSI), e potranno essere aggiunti al corpus in un prossimo futuro.

[19] De Grassi 1982, p. 58-60, 70-71; Ferrari 2002, p. 356-357, 366.

Eco del Litorale [EDL]. Nell'ottobre del 1871, due settimane dopo la nascita della testata dei liberali italiani *L'Isonzo*, la diocesi di Gorizia patrocinò la fondazione del giornale italiano *Il Goriziano* (da non confondersi con un'altra pubblicazione locale, ad opera di un gruppo di liberali italiani di orientamento piuttosto radicale e data alle stampe nel 1876-78). Dal 1873 il foglio cambiò il proprio nome in *L'Eco del Litorale* e per gran parte degli anni in cui esistette una delle sue più note firme fu padre Domenico Alpi. Il giornale si caratterizzò subito per la sua netta posizione antiliberale e a sostegno della Chiesa e dell'Impero, diffondendosi innanzitutto nelle campagne. Tuttavia a partire dagli anni Novanta del XIX secolo, *L'Eco del Litorale* abbracciò gradualmente le nuove tendenze cristiano-sociali emerse con l'enciclica Rerum Novarum e divenne di fatto il principale sostenitore nell'opinione pubblica del movimento politico che da essa fu ispirato, capeggiato nell'Isontino da don Luigi Faidutti.

L'Eco era inoltre molto spesso obiettivo delle polemiche sollevate dalle testate via via amministrate da Luzzatto. Nonostante la lealtà alla casa d'Austria, anche il giornale dei cattolici goriziani subì le pressioni delle autorità governative, che ne imposero la temporanea sospensione nel periodo intorno al quale i rispettivi governi sottoscrissero nel 1882 gli atti definitivi che portarono alla costituzione della Triplice Alleanza; *L'Eco del Litorale* si era infatti spinto a criticare apertamente la scelta del Governo di suggellare l'alleanza fra Austria-Ungheria ed Italia, la nazione, quest'ultima, a cui non si perdonava di aver sconfitto e annesso lo Stato Pontificio circa dieci anni prima.

La testata cattolica mutò periodicità nel tempo, muovendosi fra pubblicazioni quotidiane, bi o trisettimanali. Per timore dell'imminente conflitto con l'Italia dall'aprile del 1915 essa fu pubblicata a Vienna, ma gli ultimi numeri vennero editi a Trieste nel 1918.[20]

Abbiamo digitalizzato l'intera serie de *L'Eco del Litorale* dal 1874 alla primavera del 1915, dunque tutti i numeri stampati a Gorizia.

[20] Feresin 2007-2008, p. 17; Medeot 1981, p. 29-40; De Grassi 1982, p. 74-75.

Non abbiamo incluso *Il Goriziano*, il quale tuttavia resta disponibile in microfilm presso la BSI e anch'esso potrà in futuro essere aggiunto a questo corpus.

Soča, Gorica, Primorski List. I liberali e i cattolici sloveni tentarono, all'inizio del periodo considerato, di allearsi sotto la comune bandiera della nazione slovena. Insieme fondarono nel marzo 1871 quello che presto divenne il principale giornale sloveno della contea il quale, ispirandosi al suo principale corso d'acqua, prese il nome di *Soča*. Quattro anni dopo cattolici e liberali sloveni fondarono anche l'organizzazione politica unitaria *Sloga*. Quello sloveno restava però un panorama politico complesso e sfaccettato, dominato da forti personalità che spesso non scendevano a compromessi, e diviso da fratture non solo ideologiche, ma anche generazionali. La stessa area cattolica era frammentata sin dall'inizio fra gli orientamenti clerical-conservatore e progressista; quest'ultimo abbracciò in seguito il messaggio della Rerum Novarum. Nonostante le temporanee scissioni del 1872-75 e del 1889-92, in cui la componente cattolica cercò di pubblicare dei propri giornali, l'alleanza cattolico-liberale sopravvisse di fatto sino al 1899 quando si giunse a una rottura ufficiale. I cristiano-sociali sloveni iniziarono subito a pubblicare il *Gorica*, diretto dalla carismatica figura di Anton Gregorčič, già a capo del *Soča* nel 1882-89 e nel 1892-99. A sua volta il *Soča*, rimase sotto il controllo dei leader liberali Andrej Gabrščček e Henrik Tuma (Tuma divenne poi socialista e lasciò il giornale nel 1908), per poi concludere la propria esperienza nel 1915.

I cattolici conservatori, allo scopo di pubblicare un giornale che si rivolgesse a tutti i cattolici del Litorale (Sl: *Primorje*) fondarono a Trieste nel 1893 il *Primorski List*, attaccando i cristiano-sociali come criptosocialisti. Ciononostante, l'anno successivo la testata iniziò a essere pubblicata a Gorizia e dal 1898 iniziò anche a dar vita a una collaborazione con il *Soča,* cominciando pure a supportare il movimento cristiano-sociale. In seguito alla rottura fra cattolici e liberali del 1899 tentò, inutilmente, una fusione con il *Gorica* nel 1900. L'arcivescovo

Missia però, contrario alla eccessiva proliferazione di giornali cattolici che peraltro supportavano tutti la linea cristiano-sociale, forzò sia la chiusura del *Primorski List* che del *Gorica,* all'inizio del 1914, per sostituirli col *Goriški list* (non incluso nel corpus). Esso ebbe vita breve e concluse le propria esperienza l'anno dopo a causa della guerra.[21]

Soča, *Gorica*, e *Primorski List* sono stati digitalizzati della Biblioteca Nazionale e Universitaria di Lubiana e sono liberamente disponibili online.[22]

2.3. La Biblioteca Statale Isontina

Il barone Carl von Czoernig scriveva nel 1873:

> Tra gli istituti scientifici dobbiamo nominare l'i.r. Biblioteca degli studi e il Museo provinciale. I primi passi per l'allestimento della Biblioteca si fecero nel 1819 a proposito della riorganizzazione del Liceo filosofico che era stato sciolto in seguito all'invasione francese del 1810; il nocciolo era costituito dalla collezione di libri del soppresso collegio dei gesuiti. Alla fine del 1872 la biblioteca, che è diretta da un custode insieme con un amanuense ed è dotata di 1.000 fiorini annui, possedeva 10.159 opere in 17.975 volumi.[23]

Qualche notizia in più, con un tono velatamente nazionalistico, si ritrova nel saggio, sempre del 1873, scritto dal podestà Alessandro de Claricini.[24]

Venti anni prima, nel 1853, la Biblioteca degli studi era stata inserita nell'importante *Handbuch Deutscher Bibliotheken di* Julius Petzholdt.[25] La Biblioteca risale infatti agli inizi dell'Ottocento, allor-

21 Marušič 2005, p. 250-256, 330-344; Medeot 1981, p. 29-30.
22 <www.dLib.si>.
23 Von Czoernig 1987, p. 47.
24 De Claricini 1873.
25 Petzholdt 1853, p. 148. Nel testo originale venne riportato quanto segue: "Gorz. Gymnasialbibliothek war 1843 im Besitze von 7098 Banden und einer jahrlichen Dotation von 50 Fl. C. M. zu Anschaffungen. Die Leserzahl betrug 1714 Personen".

ché, dopo l'occupazione francese, il governo austriaco decise di riformare l'istruzione liceale e di trasformare la biblioteca ginnasiale in una istituzione pubblica di cultura nel 1819. Solo nel 1822 fu emanato il decreto aulico di costituzione, ma l'apertura al pubblico, a causa della disorganizzazione dei cataloghi e della mancanza di personale tecnico, avvenne solo nel novembre 1825. Nel sistema bibliotecario asburgico la Biblioteca degli studi (*Studienbibliothek*) veniva generalmente aperta in quelle città che le autorità ritenevano di maggior rilevanza culturale; condizione essenziale era l'esistenza di un ginnasio al quale la biblioteca veniva istituzionalmente collegata. Nel momento in cui Petzholdt scrisse il suo volume, la biblioteca goriziana era la più piccola e la più giovane, le altre erano invece nelle cittadine austriache Linz, Salisburgo, Klagenfurt, nella slovena Lubiana e a Olmütz, in territorio ceco.[26]

Nelle città più importanti, erano funzionanti le Biblioteche universitarie, delle quali la più vicina a Gorizia si trovava a Graz, mentre al vertice dell'amministrazione bibliotecaria stava quella che oggi è la Biblioteca Nazionale di Vienna. La *Studienbibliothek* di Gorizia, caduta sotto la mannaia della propaganda nazionalistica italiana come del resto altre istituzioni di origine asburgica, cessò formalmente le sue funzione nel 1915, quando il bibliotecario Bratina trasferì il patrimonio raro e di pregio in Austria e chiuse la Biblioteca. La storia della *Studienbibliothek* goriziana deve essere ancora seriamente indagata: sono rimasti l'archivio, i registri inventariali, il catalogo per autori e per materie, parzialmente trasferito nella Biblioteca Digitale Italiana - Cataloghi storici, e il patrimonio librario, che, contrassegnato dalla sigla di collocazione SB, è ora concentrato quasi completamente in un magazzino della attuale Biblioteca Statale Isontina (BSI), che ne ha ereditato le funzioni. La biblioteca degli studi ammonta oggi ad oltre 30 mila volumi ed opuscoli, compresi i periodici, le carte geografiche, i libri antichi e i manoscritti: non è stato mai conteggiato il numero dei

26 Ibidem.

volumi andati dispersi nella Prima Guerra quando il materiale ritenuto di pregio fu messo in salvo prima in Austria e poi a Firenze nella Biblioteca Laurenziana.

Arrivata l'amministrazione italiana, tutti gli uffici dello stato asburgico passano allo stato italiano; questo spiega spiega l'anomalia di una biblioteca statale a Gorizia, e non negli altri capoluoghi della regione, fuorché Trieste, dotati solo di biblioteche civiche: in un certo senso è una Biblioteca d'*ancien regime*, come la maggior parte delle biblioteche statali, provenienti dagli stati preunitari.

Invero, anche Gorizia aveva, ed ha, formalmente, una biblioteca civica. Nel 1893 il Comune, dopo anni di discussioni, inaugurò la Biblioteca Civica, ma a ben leggere le carte la scelta ebbe più intendimenti politici che culturali; del resto si trattava pur sempre di una biblioteca italiana in una città dell'impero asburgico. La vita della Civica, almeno dal punto di vista amministrativo, non fu facile, dato che mancavano risorse economiche e il personale, oltre a una sede fissa. Nel 1919 in una Gorizia coperta da macerie, il sindaco Giorgio Bombig decise collocare la Civica all'interno di Palazzo Werdenberg, sede della Biblioteca Governativa, che aveva nominalmente preso il posto della *Studienbibliothek* e che dal 1967 venne denominata BSI. Anche l'Amministrazione Provinciale compì la medesima scelta per la sua Biblioteca-Archivio. Regista di tale brillante operazione biblioteconomica fu Carlo Battisti (1882-1977), trentino, docente universitario e bibliotecario a Vienna che, dopo una rocambolesca vita militare sotto l'Austria, approdò, grazie alla sua esperienza e alla padronanza sia del tedesco che dell'italiano, alla direzione della nuova Biblioteca di Stato goriziana nel luglio del 1919. La Biblioteca Provinciale rimase unita alla Statale fino all'aprile 1941, mentre per la Civica nulla è cambiato e da allora si trascina l'annuale rinnovo di convenzione con la BSI.

Le raccolte della Civica, costituite *in primis* dalla biblioteca privata dell'erudito Giuseppe Domenico Della Bona (1790-1864), riguardano particolarmente il Friuli orientale e la Venezia Giulia, comprese Istria e Dalmazia, e in genere la letteratura d'evasione per adulti e ragazzi,

ed è questo l'unico aspetto che ancora caratterizza le raccolte della Civica. In considerazione dell'origine dellaboniana sono interessanti i manoscritti e le edizioni sei-settecentesche (molto meno le cinquecentine). La vita della Civica ebbe un momento importante nel 1973, quando le fu assegnato, per volere testamentario della sorella Paula, il materiale manoscritto, grafico e a stampa, del filosofo goriziano Carlo Michelstaedter (1887-1910). Da allora il fondo è stato visitato da centinaia di studiosi ed oggi conserva oltre 2300 documenti che a vario titolo riguardano Michelstaedter.[27]

Negli ultimi anni sono confluite nell'Isontina numerose biblioteche private, che coprono differenti ambiti disciplinari e forse anche per questo motivo sono rimaste sempre elevate, nonostante la vicinanza di altre istituzioni bibliotecarie, le consultazioni e i prestiti locali e interbibliotecari. La scelta, nel 1998, di inserire il maggior numero possibile di dati bibliografici nel catalogo del Servizio Bibliotecario Nazionale ha ovviamente offerto ad una platea sempre più ampia la possibilità di effettuare ricerche, a cominciare dai fondi storici, cioè la *Studienbibliothek*, il fondo Gesuitico e delle cinquecentine.

Si deve a Carlo Battisti la scelta di impegnare la Biblioteca anche sul versante editoriale; nel 1923 uscì infatti il primo numero della rivista "Studi Goriziani" che pur con interruzioni e difficoltà finanziarie è giunta al numero 113 (il n. 105 comprende gli indici generali). La rivista, ora con cadenza annuale, è stata per decenni l'unica sede scientifica per gli studi di storia locale e oggi, nonostante la presenza dell'editoria specializzata in questo settore, mantiene intatto il prestigio scientifico, oltre a garantire alla Biblioteca il ricevimento in cambio di oltre 200 riviste di ambito storico edite dalle deputazioni di storia patria italiane e dalle principali istituzioni di ricerca storica tedesche, austriache e slovene. L'attività editoriale della Biblioteca non è limitata unicamente alla rivista, ma comprende la collaborazione con altre case editrici per la pubblicazione di cataloghi di fondi librari e di

27 <http:// www.michelstaedter.beniculturali.it>.

mostre d'arte.

Dal punto di vista della biblioteca digitale, si è appena concluso il progetto di digitalizzazione di 14.407 unità librarie, facenti parte del fondo *Studienbibliothek* e cinquecentine, portato avanti da *Google Books* sotto la direzione tecnica della Biblioteca nazionale centrale di Roma, progetto al quale hanno partecipato le Nazionali centrali di Roma e di Firenze insieme ad alcune biblioteche dipendenti dal Ministero beni culturali (vedi De Simone 2019).

3. Il corpus digitale

3.1. Giornali italiani

Carta e microfilm. I giornali goriziani del XIX e XX secolo sono stati sempre raccolti e conservati nel medesimo edificio, che un tempo era sede della Biblioteca Governativa e oggi della BSI. Fra 1984 e 1999 BSI condusse un programma di microfilmatura dei propri periodici che produsse un totale di 563 bobine, contenenti le testate italiane più rappresentative pubblicate nella propria area geografico-culturale. Il programma fu anche volto a colmare le numerose lacune del posseduto, ricorrendo ad altre biblioteche locali quando possibile. Pertanto la collezione dei giornali in microfilm risulta più completa di quella cartacea, costituendo una risorsa fondamentale per le odierne esigenze di digitalizzazione.[28]

Digitalizzazione delle Immagini**.** Abbiamo digitalizzato l'intera serie de *Il Corriere di Gorizia* e *Il Corriere Friulano*, comprendendo così il periodo fra 1883 e 1914. Risulta tuttavia lacunosa l'annata 1900, a causa di una diversa denominazione della testata, chiamata allora *Il Friuli Orientale*, che non abbiamo incluso. Complessivamente il pro-

[28] De Simone 1996.

cesso di digitalizzazione dei microfilm dei giornali liberali italiani ha prodotto delle immagini digitali corrispondenti a 21.855 pagine.

Abbiamo inoltre digitalizzato l'intera serie de *L'Eco del Litorale* dal 1874 al 1914, producendo delle immagini digitali per un totale di 25.611 pagine.

Abbiamo digitalizzato complessivamente 42 bobine di microfilm da 35 mm, ciascuna lunga 100 piedi e capace di contenere approssimativamente 600 immagini a doppia pagina, utilizzando il microfilm scanner *Wicks & Wilson 8850*. Le immagini prodotte sono stati file .tiff non compressi in scala di grigio con qualità 300 dpi, convertite successivamente in .jpeg per *Quality Control* (QC). Il processo di scannerizzazione della singola coppia di pagine dal microfilm aveva prodotto degli ampi bordi bianchi intorno al testo riprodotto, che sono stati eliminati tramite un processo automatico. Successivamente ogni immagine è stata divisa in due in modo che ogni file contenga una sola pagina di giornale; questa procedura è stata fatta semi-manualmente a scaglioni prima del QC.

Quest'ultimo ha incluso il controllo logico che in questo caso ha significato la verifica della corrispondenza del numero di immagini presenti in una singola bobina con quelle presenti nella cartella di file relativa, e il controllo della dimensione di tutti i file. Il controllo fisico delle immagini ha incluso il controllo visuale per i livelli di luminosità, la correttezza del processo di eliminazione dei bordi bianchi, e la verifica dei nomi e dei metadata delle cartelle. Tali procedure di QC costituiscono una condizione fondamentale per la buona riuscita complessiva del progetto e sono state completamente documentate, consentendo così la tracciabilità di ogni passaggio.

Complessivamente, l'intero processo di digitalizzazione ha prodotto 47.466 immagini corrispondenti ad altrettante singole pagine di giornale (EDL - 25.611 pp, CFG - 21.855 pp).

Annotazione delle immagini. Ciascuna delle 42 bobine ha circa 1130 pagine singole di giornale, ma nel corso del tempo i periodici

interessati dalla ricerca variarono il numero di pagine e la frequenza, pertanto il lasso di tempo coperto da ciascuna bobina varia sensibilmente, essendo state a loro volta sempre utilizzate in toto senza riprodurre le unità archivistiche originali, per ragioni di economia. Non abbiamo quindi potuto mettere in atto alcuna procedura automatizzata per assegnare la data corretta e il numero esatto a ciascuna immagine. Le date, altresì, si trovano generalmente solo in prima pagina e non sempre sono di facile lettura. Siamo stati perciò costretti ad assegnare manualmente data e numero di pagina a ciascuno dei 47.466 file/pagina, procedura che ha richiesto molto tempo. È possibile che in questa fase ci siano stati degli errori, e non abbiamo operato un completo passo di QC, a parte il ri-controllo manuale di tutte le pagine che violavano i seguenti vincoli di consistenza interna: l'ordine delle date doveva comunque essere coerente con l'ordine fisico delle immagini in bobina, non dovevano esserci ampi periodi di tempo privi di pagine, i giorni della settimana dovevano combaciare con la data. Tale metodo di annotazione non è perfetto ma riteniamo che sia adeguato per l'analisi delle tendenze statistiche, considerato che abbiamo scelto di operare a livello di trimestri, come descritto meglio in seguito. Inoltre ci ha consentito di stabilire/stimare che la parte del corpus relativa a CFG contiene 5.475 distinti numeri del periodico (e quindi date), mentre l'EDL ne abbraccia 6.356.

Optical Character Recognition (OCR). Per convertire le immagini delle pagine di giornale in testo editabile abbiamo utilizzato il software Abbyy FineReader, versione 12,[29] specificando che il testo da elaborare è stato scritto in italiano. FineReader ha automaticamente tentato, dove possibile, di adeguarsi alla distribuzione del testo in molteplici colonne per pagina, e di suddividere a sua volta la pagina e il testo in articoli; tuttavia abbiamo scelto di considerare come unità fondamentale dell'analisi la singola pagina, senza scendere al livello

[29] <https://www.abbyy.com/en-gb/finereader12/en/>.

dei singoli articoli né abbiamo voluto estrapolare o evidenziare i titoli di ogni articolo. Inoltre, benché FineReader potesse distinguere fra articoli, tabelle e immagini, abbiamo deciso di usare per l'analisi tutto il testo disponibile, che fosse quello degli articoli o all'interno di tabelle, e di scartare le immagini riconosciute come tali dal software OCR. Questo passaggio ci ha consentito di conteggiare 67.068.101 parole per la parte del corpus relativa al CFG e 50.429.470 parole per quella estratta dall'EDL.

La nostra versione di FineReader non consente però di conoscere la stima della precisione del riconoscimento del testo originale da parte dello stesso software, la quale a sua volta è composta da due valutazioni, una riguardante il corretto riconoscimento del carattere, *Character Error Rate* (CER), e l'altra il corretto riconoscimento della parola, *Word Error Rate* (WER). Abbiamo pertanto calcolato tali stime selezionando casualmente dieci articoli, in modo da rappresentare entrambe le testate e diverse epoche, e trascrivendoli manualmente, allo scopo di ottenere una *ground truth* da comparare al risultato di FineReader. Dei dieci articoli casualmente selezionati, uno non è stato però processato correttamente dal software in quanto conteneva un'immagine di una certa dimensione, la quale ha spinto lo stesso programma a ritenere l'intero articolo un'immagine, annullando l'intero processo di OCR e non producendo alcun testo digitale; pertanto abbiamo a nostra volta scartato questo articolo dal conteggio dei due tassi di errore. Sulla base dei restanti nove articoli, abbiamo misurato per FineReader un CER del 24,6% e un WER del 23,1%, che sono in linea con quanto riscontrato in letteratura.[30]

Tali errori sono essenzialmente provocati da tre fattori: problemi di conservazione, quali polvere su supporto cartaceo o pellicola, limiti del processo di scannerizzazione, che può essere tratto in inganno da caratteri stampati nella pagina retrostante visibili in trasparenza, e ambiguità fra caratteri non perfettamente stampati, ad esempio per le

30 Lansdall-Welfare et al. 2017.

similitudini fra le lettere "è" e "ò".

Abbiamo scelto di limitarci a un'analisi statistica e di non utilizzare tecniche di analisi del testo più sofisticate come il *Natural Language Processing* (NLP) o *Information Extraction* (IE), in quanto queste avrebbero necessitato o di un testo più pulito o di un corpus più ampio. Metodi di NLP sarebbero andati oltre il semplice conteggio delle parole, analizzando la struttura delle frasi e valutando il significato delle parole nel loro contesto, ma per essere affidabili devono poter contare su frasi intere prive di errori di digitalizzazione. L'IE sarebbe stato utile nell'estrazione del contenuto delle liste e delle tabelle, ma sarebbe stato altrettanto vulnerabile agli stessi errori rispetto alla più semplice analisi statistica che abbiamo invece deciso di elaborare. Sia l'NLP che l'IE possono essere usati con database imperfetti, ma almeno il corpus sarebbe dovuto essere molto più vasto e ridondante di quello invece in nostro possesso, in modo che tale ridondanza avrebbe potuto compensare il più basso tasso di risultati utili dovuto agli errori di digitalizzazione presenti nel corpus. Per tali ragioni abbiamo preferito concentrarci su una più semplice analisi statistica, che include word clouds e serie temporali delle frequenze relative di parole e frasi. Ad ogni modo riteniamo che in futuro algoritmi più avanzati saranno capaci di estrarre informazione di più alta qualità da questo corpus, particolarmente se riusciremo ad estendere il corpus aggiungendo altre testate.

3.2. Giornali Sloveni

Dato il panorama essenzialmente bilingue della regione, abbiamo scelto di aggiungere al nostro corpus anche tre importanti giornali sloveni dello stesso periodo allo scopo, inoltre, di poterli comparare con i periodici italiani che abbiamo prima descritto.

I giornali *Soča*, *Gorica*, e *Primorski List* sono stati digitalizzati dalla Biblioteca Nazionale e Universitaria della Slovenia e sono disponibili online. DI essi abbiamo incluso rispettivamente 3342, 1472 e 892 nu-

meri, o 38.153.317, 16.189.845, e 9.662.362 parole. La digitalizzazione di tali testate si è svolta nel corso del progetto europeo IMPACT (IMProving ACcess to Text). L'elaborazione OCR é stata fatta utilizzando Abbyy FineReader, versione 10; è stato dichiarato un CER fra il 15% e il 30%, a seconda della tipologia di caratteri usate nel supporto cartaceo originale (Jerele et al, 2011). I tre giornali sloveni che abbiamo incluso in questo studio sono stati resi disponibili dalla Biblioteca Digitale della Slovenia come materiale di dominio pubblico.[31] Da parte nostra abbiamo semplicemente fatto uso del testo digitale già a disposizione degli utenti attraverso il sito web della Biblioteca.

Corpus	**Parole**	**Edizioni**	**Anni**
Gorica	16.189.845	1472	16 (1899-1914)
Primorski list	9.662.362	892	22 (1893-1913)
Soča	38.153.317	3342	44 (1871-1914)
EDL	50.429.470	6356	42 (1873-1915)
CFG	67.068.101	5475	31 (1883-1914)

Tabella 2 - Numeri delle diverse edizioni, parole e anni disponibili per ciascun corpus.

3.3. Analisi statistica di serie temporali testuali

Andamenti statistici, continuità e cambiamenti nella salienza di diverse tematiche contenute in corpus di giornali (o libri) storici possono essere valutate sulla base delle variazioni delle frequenze nelle uso di determinate parole se particolare cura viene prestata alla scelta di parole adeguate.[32] Abbiamo dunque sviluppato la nostra analisi utiliz-

31 <https://www.dlib.si/Help.aspx>.
32 Nicholson 2012; Michel et al. 2011; Lansdall-Welfare 2017.

zando serie temporali costruite sulla stima di frequenze relative di parole e frasi, calcolata su periodi di tre mesi, per avere dati sufficienti da ridurre gli errori statistici.[33] La figura 3 mostra le parole più frequenti di ciascuna testata. Notiamo che il nome di Gorizia / Gorica e' molto frequente, come ci si aspetta, e che le testate italiane menzionano spesso anche Trieste, EDL anche la Chiesa.

Essendo tanto l'italiano, quanto lo sloveno lingue prettamente flessive, lo stesso concetto può essere espresso da un alto numero di parole (leggermente) diverse. Per procedere correttamente all'analisi statistica di un testo nelle lingue di questo tipo morfologico è necessario rimuovere le inflessioni, e preservare solamente la radice delle parole con una procedura automatica denominata *stemming*. Abbiamo attuato questa procedura usando un algoritmo integrato al software *Snowball* per entrambe le lingue del corpus;[34] in questo modo abbiamo sostituito i vocaboli con le loro radici: per esempio, in italiano '*parola, parole*', diventano '*parol*', mentre in sloveno '*beseda, besede*' diventano '*besed*'.[35] Il processo di *stemming* in sloveno ha tenuto anche presente l'uso dei casi, che invece nelle lingue neoromanze sono andati perduti. Per un'analisi statistica, e non linguistica, del testo, questa informazione è sufficiente a stimare se una parola cambia frequenza - indipendentemente da caso, genere e numero.

Successivamente il testo è stato ripartito in n-gram della lunghezza di uno, due e tre vocaboli. Per esempio, chiamiamo 3-grammi quelli che talvolta si chiamano tri-lemmi, ovvero tre parole in serie, in questo

---

33 Per il metodo con cui abbiamo stimato la frequenza relativa si veda in particolare: Lansdall-Welfare Cristianini 2018.

34 *Snowball* per l'italiano: <snowball.tartarus.org/algorithms/italian/stemmer.html>; per lo sloveno: <snowball.tartarus.org/archives/snowball-discuss/att-0670/01-slo.proc>.

35 Abbiamo inoltre rimosso gli articoli e le preposizioni in italiano seguite da apostrofo e anteposte ai vocaboli, utilizzando un filtro per le elisioni contenente nel *package* 'Lucene', estraendo ad esempio: "c", "l", "all", "dall", "dell", "nell", "sull", "coll", "pell", "gl", "agl", "dagl", "degl", "negl", "sugl", "un", "m", "t", "s", "v", "d".

caso tre radici di parole. Per chiarezza, "giornale storico italiano" e "giornali storici italiani" danno entrambi origine al tri-gramma "giornal- storic- italian-". Tutti gli n-gram con una frequenza inferiore a 10 sono stati scartati: la ragione di questa operazione è dovuta alla presenza di errori prodotti nel corso del processo di OCR, che abbiamo visto essere inevitabili nel trattamento di giornali storici. Abbiamo perciò rimosso tutti gli n-gram che erano quindi poco frequenti, dato che buona parte di questi o erano parole digitalizzate erroneamente, e risultano essere prive di significato oppure sono parole talmente rare da non avere molto interesse nello studio di vasti trend statistici. Le dimensioni del lessico risultante da queste procedure può essere visto nella Tabella 3.

Corpus	**1-grams**	**2-grams**	**3-grams**	**n-grams totali**
Gorica	53.683	166.412	93.273	313.368
Primorski list	25.891	111.519	66.422	203.832
Soča	90.425	363.337	240.266	694.028
EDL	90.786	395.821	234.248	720.855
CFG	116.793	492.933	316.155	925.881

Tabella 3 - Dimensioni del vocabolario di ciascuna testata

Per ciascuno degli n-gram rilevati abbiamo dunque generato una serie temporale della frequenza relativa. Dobbiamo inoltre sottolineare che dal punto di vista statistico, non abbiamo dati sufficienti per stimare la frequenza relativa di ogni parola a livello giornaliero, né tantomeno a quello mensile. Per tale ragione abbiamo scelto di impostare

l'analisi a livello trimestrale. Questo scelta ci ha consentito di non aver bisogno di ripartire il testo in articoli, un passaggio complesso e al contempo foriero di errori statistici, e di essere accurati nell'applicare manualmente le date a ciascuna delle pagine digitalizzate. Fondendo quindi il contenuto testuale di tre mesi di pubblicazioni di ciascuna testata in un singolo *time point* per il quale le frequenze di ogni n-gram viene computato, vengono così minimizzate altre possibili distorsioni nell'analisi complessiva dei dati.

Rimarchiamo pertanto che le procedure di analisi che abbiamo stabilito di fatto stimano la probabilità dell'uso di una determinata parola nel corso di ogni trimestre per ogni singolo giornale. Anche se una parola non può essere esattamente rilevata in tutte le occasioni in cui fu usata dagli autori dei testi originali, a causa degli errori in che abbiamo prima descritto, la probabilità di frequenza della parola nel trimestre può comunque essere stimata; caso analogo è quello del calcolo delle probabilità di una moneta non imparziale nel favorire testa o croce avendo a disposizione il risultato di un numero finito di lanci.

In questo modo abbiamo potuto rappresentare la salienza di ciascuna parola (o di una espressione di due o tre parole) del corpus nel corso del tempo in base alla frequenza relativa di ciascun n-gram per ogni trimestre.[36] Ne sono risultati complessivamente 2.857.964 serie temporali, rappresentanti una stima della frequenza relativa di ciascun n-gram presente in ciascun trimestre di ogni giornale del corpus. In questa sede utilizzeremo alcune di tali serie temporali allo scopo di mostrare come questa metodologia possa mettere in luce tendenze statistiche, continuità e trasformazioni nei 41 anni compresi dal corpus.

Nel complesso, dalla collezione di 42 microfilm relativa ai due giornali italiani abbiamo estratto 47.466 pagine e 207.579 serie temporali di singole parole, che sono state espresse da 168 trimestri di EDL e

[36] Per ogni n-gram estratto dal testo originale, abbiamo stimato la frequenza relativa durante periodi di tre mesi, seguendo il metodo descritto in: Lansdall-Welfare - Cristianini 2017.

122 di CFG. Se inoltre consideriamo i tre giornali sloveni che abbiamo incluso nella nostra analisi, possiamo affermare che complessivamente il corpus è formato da 181.503.095 parole e comprende 164 trimestri dal 1873 al 1914, come indicato nella Tabella 4.

3.4. Panoramica dell'analisi statistica tramite word clouds

Sia per proporre una panoramica generale del corpus che abbiamo costruito, e sia per effettuare un controllo dell'integrità dei processi condotti nel corso della ricerca, abbiamo prodotto delle *word clouds* per ciascuno dei cinque giornali. Inoltre vogliamo con ciò facilitare da parte del lettore la visualizzazione di quale tipo di parole siano state prodotte nel corso della nostra indagine, di quale ordine di errori possano essere rilevati, e che genere di informazioni sia possibile estrarre con le tecniche che abbiamo intrapreso. Dopo aver rimosso le *stop words*, abbiamo ottenuto le *word clouds* mostrate nella Fig. 3, che mostrano le parole più frequenti in ciascuna delle cinque testate nel corso della loro vita editoriale. Come c'era da aspettarsi 'Gorizia-Gorica' risulta essere tra le parole più frequenti, tranne per il cattolico EDL, dove è la parola 'Chiesa' ad essere la più frequente; si nota altresì la presenza marcata di Trieste nei due giornali italiani. Le *word clouds* consentono inoltre di visualizzare il livello di errore prodotto in fase di OCR, elemento purtroppo ineliminabile in questo genere di analisi statistica.

4. Analisi testuale delle serie temporali

Intendiamo ora mostrare alcuni esempi di esplorazione del corpus mediante l'analisi statistica, con lo scopo di mostrare come quest'ultima possa integrarsi all'interpretazione storica.

La variazione della frequenza relativa di una parola può indicare i periodi in cui la parola stessa è stata più o meno utilizzata e così

permettere di comprendere verso cosa l'opinione pubblica stesse prestando più o meno attenzione. Tuttavia, persino 180 milioni di parole possono non essere sufficienti per certi tipi di analisi statistica, pertanto dobbiamo anche mettere in conto di non poterci aspettare sempre un segnale del tutto affidabile dalla stima della frequenza relativa delle parole. Ciò è dovuto al fatto che in ogni lingua la maggior parte delle parole vengono usate con una frequenza abbastanza ridotta, la quale può essere misurata efficacemente solo in un campione sufficientemente vasto. Perché la metodologia qui proposta resti efficace, è necessario dunque concentrare l'analisi statistica solo sulle parole più frequenti, per di più scegliendo quelle dal significato meno ambivalente, o sugli episodi e le personalità maggiormente rappresentativi. In ogni caso è buona norma cercare di corroborare i segnali statistici che possono emergere dal database con un'esplorazione diretta delle fonti, leggendo un campione selezionato di testi originali. D'altra parte il dibattito all'interno della comunità accademica su quale sia il miglior modo di esplorare i corpus digitali è tutt'ora aperto; gli approcci più vicini al presente studio sono quelli del *distant reading*[37] e della *culturomics*[38] i quali entrambi mirano a fare ciò che abbiamo appena proposto, oltre a sottolineare l'importanza di un continuo dialogo fra analisi statistica e lettura dei testi originali allo scopo di rendere più precisa la prima e più efficace e meno dispersiva la seconda.[39]

Gli esempi di analisi che ora vi proponiamo mirano proprio a mostrare la potenzialità di queste metodologie, che sono tra le più efficaci per sfruttare al meglio le fonti a stampa digitalizzate.

Procederemo inizialmente verificando la traccia lasciata do alcuni noti eventi dell'epoca, allo scopo di verificare come il corpus risponda alle più semplici sollecitazioni del ricercatore. Dato che le pubblicazioni coinvolte da questo studio vennero alla luce in un'epoca caratterizzata da forti trasformazioni di ordine sociale, tecnologico e

37 Moretti 2013.
38 Michel et al. 2011.
39 Nicholson 2012; Lansdall-Welfare 2017.

culturale, che si svolsero in un territorio complesso, incontro di diversi popoli, gli esempi successivi cercheranno di sondare le potenzialità della metodologia proposta per gli studiosi interessati a tali tematiche.

4.1. La cometa di Halley

I passaggi di ben due comete furono visibili nei cieli europei del 1910. La prima, del tutto inaspettata, apparse nella volta celeste già il primo mese dell'anno; si trattava di una cometa non periodica che fu semplicemente battezzata "Grande cometa del 1910". La più nota e da tempo annunciata cometa di Halley apparse invece fra aprile e maggio. Tutti i giornali del mondo ne parlarono, pertanto tali eventi costituiscono dei segnali perfetti per verificare i metodi e gli strumenti che abbiamo impiegato.

A Gorizia, il passaggio della cometa fu considerato da alcuni con sospetto. Molti temevano che essa fosse foriera di cattive notizie, tuttavia i giornali del corpus trattarono l'argomento assumendo una prospettiva essenzialmente scientifica, come questo breve articolo intitolato "La paura della cometa - La fine del mondo!", apparso ne L'*Eco del Litorale*:

> La cometa di Halley ha destato terrore fra le popolazioni rurali slovene e croate della Carniola, del territorio di Trieste e della Dalmazia. Secondo rapporti pervenuti al Governo, la paura è così grande e la convinzione del prossimo finimondo così diffusa che parecchi contadini pensano di vendere i loro beni e di darsi alla pazza gioia che tanto fa lo stesso. Insomma una ripetizione dei terrori del 1000, ma con una rassegnazione più allegra. Ciò posto, il Ministero dell'istruzione mandò un ordinanza ai governatori della Carniola, di Trieste e della Dalmazia perché provvedano a tranquillare [sic] le popolazioni a mezzo dei maestri e dei parroci, spiegando popolarmente nella scuola e dal pulpito la teoria delle comete. Un apposito opuscolo si distribuirà ai maestri e ai preti.[40]

[40] *La paura della cometa - La fine del mondo!*, «Eco del Litorale», 16/04/1910, p.2.

Nella Figura 4 possiamo vedere come il nostro corpus mostri chiaramente il passaggio della cometa, suggerito dalla traccia lasciata dalla frequenza relativa della parola "Halley" nella stampa locale di Gorizia. Inoltre, visto che il nome dell'astronomo britannico veniva scritto allo stesso modo in entrambe le lingue rappresentate nel corpus, siamo così in grado di verificare contemporaneamente l'efficacia dell'analisi tanto nei giornali italiani che in quelli sloveni.

4.2. L'imperatore

L'imperatore Francesco Giuseppe regnò sopra i domini asburgici dal 1848 al 1918. Il prestigio del suo ruolo imponeva inevitabilmente alla stampa locale di riferirvisi con la massima deferenza. Era del resto proibito per legge criticare o sbeffeggiare il sovrano e la famiglia reale,[41] così che, oltre ad alcuni giovani liberali italiani di tendenze irredentistiche, nessuno dei redattori delle maggiori testate goriziane avrebbe facilmente rischiato la prigione per poter mettere apertamente in discussione la casa d'Austria.[42]

Le visite di Francesco Giuseppe alla regione furono sempre un evento di primissimo piano. Visitò Gorizia in ben cinque occasioni, negli anni 1850, 1857, 1875, 1882 e il 29 settembre 1900, quando fu invitato dall'amministrazione municipale per celebrare i 400 anni di Gorizia asburgica.[43] Ogni visita dell'imperatore nel periodo coperto dal corpus ha lasciato un evidente picco della parola imperatore sul grafico.

L'incoronazione di Francesco Giuseppe avvenne il 2 dicembre 1848, per cui ogni dieci anni da quella data l'anniversario godeva di particolare copertura dalla stampa, fatto anch'esso testimoniato dai picchi del grafico in Figura 5.

41 Horel 2015, p. 89.
42 De Grassi 1982, p. 57-58.
43 Agostinetti 1981, p. 42.

4.3. Il cinema

Gli anni fra il 1873 e il 1914 furono un periodo di grandi trasformazioni, segnato dalla seconda rivoluzione industriale, che si tradusse non solo nella comparsa di numerose nuove invenzioni tecnologiche e di nuovi e più efficienti modi di produzione, ma anche nel riassetto generale delle società occidentali, che influenzarono sempre più il resto del mondo, e nell'emersione del ruolo delle masse nella storia. Una delle grandi novità del periodo, che forse più di tutte ne sintetizzò le caratteristiche salienti, fu il nuovo mass medium dell'epoca, il cinema.

La prima proiezione cinematografica goriziana avvenne l'8 dicembre 1896, presso l'Hotel Central. Due giorni dopo il Corriere di Gorizia ne riportò la notizia, descrivendo alcune delle famose scene che fino ad oggi caratterizzano la memoria collettiva sulla nascita del cinema:

> Le vedute di questo cinematografo sono varie; p. e. esso ci mostra una sfida di donne, della ginnastica infantile, una bagnante, l'arrivo di un treno ferroviario, il movimento dei passeggieri [sic] ecc. Questa della ferrovia è anzi una delle vedute più interessanti. Si vede il convoglio in arrivo, poi i conduttori che aprono gli sportelli, la discesa dei passeggieri [sic], a chi si piglia una valigia, chi un cagnolino ecc., tutto molto chiaro e molto bene, tanto che specialmente questo quadro della ferrovia fu calorosamente applaudito.[44]

Nel medesimo luogo tredici anni dopo l'imprenditore Josip Medved inaugurò un vero e proprio cinema, che chiamò Central Bio, fondendo il nome dell'Hotel a quello di un antesignano dell'invenzione dei fratelli Lumiere, il bioscopio. Fu cura del proprietario che programmi stampati in tedesco, italiano e sloveno, fossero sempre a disposizione della clientela per tutti gli spettacoli. Tuttavia già l'anno prima un altro impresario sloveno, Andrea Kumar, aveva aperto l'Edison, la prima sala cinematografica della città, che fu appositamente

44 *Il cinematografo al salone Dreher*, «Corriere di Gorizia», 10/12/1896, p. 3.

costruita per questo scopo.[45]

4.4. Associazioni sportive e culturali italiane e slovene

L'associazionismo italiano e sloveno costituì uno dei principali veicoli di espressione delle identità nazionali in epoca asburgica, seguendo modelli culturali comuni a tutta la mitteleuropa. Già prima delle riforme costituzionali del 1867, le associazioni nazionali sorsero a Trieste e Lubiana per favorire la pratica delle varie attività sportive; a Gorizia gli italiani fondarono l'*Unione Ginnastica* nel 1868 e gli sloveni il *Sokol* nel 1887. Dagli anni Settanta venne poi alla luce la Sloga, una società che già abbiamo ricordato aver avuto lo scopo di coordinare la componente cattolica e quella liberale della comunità slovena, favorendone le espressioni culturali in senso più ampio, e di cui il giornale *Soča* fu il principale portavoce. Le società culturali nel senso pieno del termine, capaci anche di organizzare strutture scolastiche per la formazione delle nuove generazioni nelle rispettive lingue nazionali, sorsero però solo negli anni Ottanta. Gli italiani di Gorizia, quasi contemporaneamente agli altri connazionali stanziati nei principali centri del Litorale, fondarono nel 1885 la *Pro Patria*; nel 1890 le autorità viennesi accusarono l'organizzazione di sostenere l'irredentismo e la resero fuorilegge, ma ben presto riaprì sotto il nome di *Lega Nazionale*. Gli sloveni dal canto loro, sempre nel 1885 costituirono la *Ciril in Metod*, prima a Lubiana e poi in tutte le principali città abitate dagli sloveni, compresa Gorizia. Il riferimento ai due santi considerati i cristianizzatori del mondo slavo, Cirillo e Metodio, si deve all'enciclica Grande Munus emessa da papa Leone XIII nel 1880, la quale favorì un particolare devozione fra i cattolici sloveni nei due predicatori dell'alto medio evo.[46]

L'esistenza delle associazioni appena citate è facilmente visibile nel nostro corpus dei grafici raccolti da Figura 7 e Figura 8. Il picco del

45 Mlakar-Turel 2010, p. 138.
46 Ferrari 2002, p. 356-357; Redivo 2005, p. 19-36; Filipič 2010.

1880-1 per Ciril in Metod conferma inoltre la considerazione dell'enciclica papale.

Influenza

Influenza - Sappiamo che c'era stata una pandemia di influenza (chiamata la Asiatica) nell'inverno 1889 - 1890. A Gorizia l'epidemia era arrivata come nel resto d'Europa, e i giornali sloveni usavano di fatto la stessa parola italiana (Fig 9).

Epidemie agricole

L'economia dell'Isontino si basava largamente sulla produzione agricola. La zona era ed è rinomata per la qualità dei suoi vini. Alla fine degli anni 1890 giunse in Europa la Filossera, un parassita della vite che mise in pericolo l'economia di vaste aree. La questione fu risolta solamente con la sostituzione di tutte le piante, che dovettero essere innestate su una base di vite americana. La Peronospora è una piaga della vite più costante, che si cura con il verderame, giunta in Italia negli anni 1880 (Fig 10).

5. Conclusioni

Il bel saggio di Franzosi *A Third Road to the Past?*[47] mette in relazione i metodi dell'umanistica digitale al confronto metodologico tra Geoffrey Elton e Robert Fogel contenuto nel saggio che insieme pubblicarono nel 1983.[48] In tale opera, i due storici discutono della strada che meglio può condurci a una comprensione del passato: quella narrativa o quella quantitativa? Elton preferisce una storia narrativa, Fogel una quantificabile. Franzosi propone l'idea che l'umanistica digitale abbia il potenziale di riconciliare queste due visioni.

La lettura distante dei giornali storici goriziani ci restituisce l'im-

47 Franzosi 2017.
48 Elton Fogel 1983.

magine di una città e di un territorio alle prese con numerose tensioni e grandi cambiamenti. Le *time-series* relative alle società sportive - per esempio - all'occhio esperto parlano di tensioni etniche, ma solo quando lo storico è consapevole di come queste fossero associate ai movimenti nazionali. L'arrivo del cinema, dell'elettricità, o delle automobili ci consentono di comparare il tenore di vita a Gorizia con quello di Vienna o Venezia. Ma le *time series* di cinema ed elettricità da sole non ci dicono se ci fossero differenze tra le diverse regioni dell'Impero e del vicino Regno. Le *time series* delle epidemie agricole ci mostrano diversi momenti di grande preoccupazione per le coltivazioni di vite, ma non ci dicono – da sole – quali comunità dipendessero maggiormente dalla produzione di vino. Insomma, possiamo formare un ritratto di questo territorio a partire dai contenuti della sua stampa locale, solo quando abbiamo già sufficienti conoscenze per metterli in un contesto. Ma quando questo contesto è presente, lo storico può accedere ai benefici di quasi cinquantamila pagine di giornale in pochi secondi.

È chiaro che il semplice conteggio della frequenza di parole e frasi non sarà mai un sostituto della storiografia. Le tecniche descritte in questo articolo non ambiscono a guidare l'interpretazione dei fenomeni storici, la fase forse più creativa e importante del lavoro dello storico, ma possono fornire un utile supporto alla ricerca, analisi e comprensione delle fonti e dei rapporti di causalità. Possono servire a eliminare ipotesi, e a formarne di nuove, a concentrare l'attenzione verso determinate direzioni.

Con sufficienti quantità e qualità di testo digitalizzato, oggi è già possibile estrarre nomi di persone, luoghi e organizzazioni, ricostruendo reti sociali e relazioni tra individui;[49] è inoltre possibile estrarre il *sentiment* (atteggiamento positivo, negativo o neutro) con cui si discutono certi argomenti,[50] anche da giornali di lingua diversa tradot-

49 Lansdall-Welfare et al. 2017b; Sudhahar et al. 2015.
50 Lansdall-Welfare et al. 2014.

ti automaticamente.[51] Altre tecniche consentono di misurare segnali relativi a possibili disuguaglianze di genere,[52] e di scoprire parole che presentano strutture periodiche.[53]

Tra breve sarà possibile estrarre dai giornali storici informazioni specifiche, come le cause di morte o i prezzi di numerosi prodotti, mentre al momento ciò è possibile solo con testi senza errori di digitalizzazione. Nel corso del tempo sarà consentito estrarre informazioni di qualità sempre maggiore, da una base di giornali sempre crescente, fino a quando gli studiosi saranno in grado di interrogare un archivio di giornali storici con domande specifiche quali: quanto costavano le patate a Gorizia nel 1810?

A quel punto sarà chiaro che il ruolo della digitalizzazione va oltre la preservazione degli originali e non è limitato a evitare agli studiosi una visita in biblioteca o in archivio.

Il ruolo della digitalizzazione, e delle analisi di umanistica digitale, sarà dunque sempre più frequentemente volto a consentire la consultazione di vaste quantità di fonti in numerose lingue, facendo domande precise, e ricevendo fonti rilevanti come risposta. Quello che gli umanisti digitali faranno con quelle risposte sarà certo al di là delle ambizioni e delle capacità di qualsiasi sistema informatico: scoprire, comprendere e spiegare avvenimenti e idee che hanno influenzato la storia e la cultura.

In Italia l'umanistica digitale sembra aver sviluppato negli ultimi due decenni numerose applicazioni nel campo della conservazione del patrimonio di biblioteche e archivi, nel settore museale, della linguistica, della stilometria e dell'archeologia. Anche nel campo della storiografia sono stati fatti alcuni interessanti esperimenti, anche se l'approccio digitale risulta al momento meno diffuso rispetto ad altri ambiti degli studi umanistici. A titolo di esempio tra gli 86 interventi

51 Sudhahar Cristianini 2018; Flaounas et al. 2010.
52 Jia et al. 2016; Lansdall-Welfare et al. 2017a.
53 Dzogang et al. 2017.

presso il convegno AIUCD (Associazione per l'Informatica Umanistica e la Cultura Digitale) del 2018 a Bari, solo quattro erano in *Digital History*.[54]

Anche per tale motivo questo articolo, allo scopo di stimolare il dibattito e la ricerca nel campo dell'umanistica digitale, cerca di ripercorrere l'intero ciclo coinvolto dalla nostra ricerca, dalla produzione delle fonti digitali alla loro analisi nel contesto degli studi sull'area goriziana e il confine nord Adriatico.

Numerosi sono ormai in Italia i patrimoni di giornali storici digitalizzati, a cominciare da giornali nazionali come La Stampa, Il Corriere della Sera e L'Unità che nell'ultimo decennio hanno messo a disposizione di tutti, gratuitamente, le proprie edizioni storiche. Lo stesso si può dire di molte altre pubblicazioni periodiche regionali lungo la penisola.

Tutto ciò potrà confluire o essere confrontato con altri grandi database internazionali quali 'Europeana', 'British Newspaper Archive', 'Chronicling America', 'Project Gutenberg'. Quello che non è ancora chiaro, invece, è il modo migliore di integrare diverse fonti di testo tra loro, magari in lingue diverse, e come integrare questi dati nel più ampio contesto storico che li ha generati. Questa è una delle maggiori sfide dell'umanistica digitale: fornire agli studiosi uno strumento nuovo, senza però dimenticare che le discipline umanistiche hanno esigenze fondamentalmente diverse da quelle scientifiche - e le une non possono ridursi alle altre.

Insomma, una strada che ci conduca a comprendere meglio il nostro passato, e a raccontarlo in modo sensato, può avvalersi anche di metodi quantitativi, specialmente quando sarà possibile estrarre dai testi informazioni complesse. Quella direzione è già stata presa, ma ha bisogno di svilupparsi ulteriormente. Migliorare i metodi di OCR, estendendoli anche ai testi manoscritti tramite lo HTR[55] (*Handwritten*

[54] Spampinato 2018.

[55] <https://transkribus.eu/wiki/index.php/Handwritten_Text_Recognition_Workflow>.

Text Recognition), darà un contributo notevole a questa ambizione.

Quanto ai giornali goriziani, i prossimi passi di questo progetto saranno in due direzioni: digitalizzare i giornali italiani mancanti che si riferiscono a Gorizia per il medesimo periodo, ma anche accostarli e integrarli ai giornali stampati in zone limitrofe, in direzione del Friuli e del resto della Venezia Giulia. Molte di queste risorse sono già state digitalizzate e altre sono presenti in microfilm nelle biblioteche locali - ad esempio i periodici *Il Gazzettino Popolare* e *L'Isonzo*.

Un'ulteriore direttrice di sviluppo di questo lavoro porterà a raffinarne i metodi statistici, adattandoli alle sfide poste dall'analisi di testo digitale corrotto da errori di OCR. Questa è una sfida tecnica che va accettata se vogliamo che l'umanistica digitale possa mantenere la sua promessa di aprirci una nuova strada verso il passato.

Ci sono molte strade che conducono alla comprensione del passato.

Noi, ispirandoci al metodo di Franzosi, abbiamo cercato di combinarne più d'una, sperimentando numerosi metodi fino a trovare quelli che appaiono essere i più fecondi. L'analisi di database di documenti, libri, giornali, e magari un giorno anche dei testi manoscritti, possono semplificare alcuni compiti degli storici e ampliarne il raggio d'azione, mentre arricchiscono l'arsenale di bibliotecari e archivisti, consentendo a tutti di operare su quantità vastissime di documenti come mai era accaduto prima.

Gli strumenti digitali possono dunque espandere l'azione umana, allargarne gli orizzonti, ma non sono altro che tecnologia, e acquistano senso solo negli scopi per cui vengono impiegati. Essi non privano gli studiosi del piacere della scoperta ma offrono nuove possibilità, e comunque, alla fine, spetterà sempre alle persone, e non alle macchine, trarre le lezioni che il passato potrà impartirci.

Illustrazioni

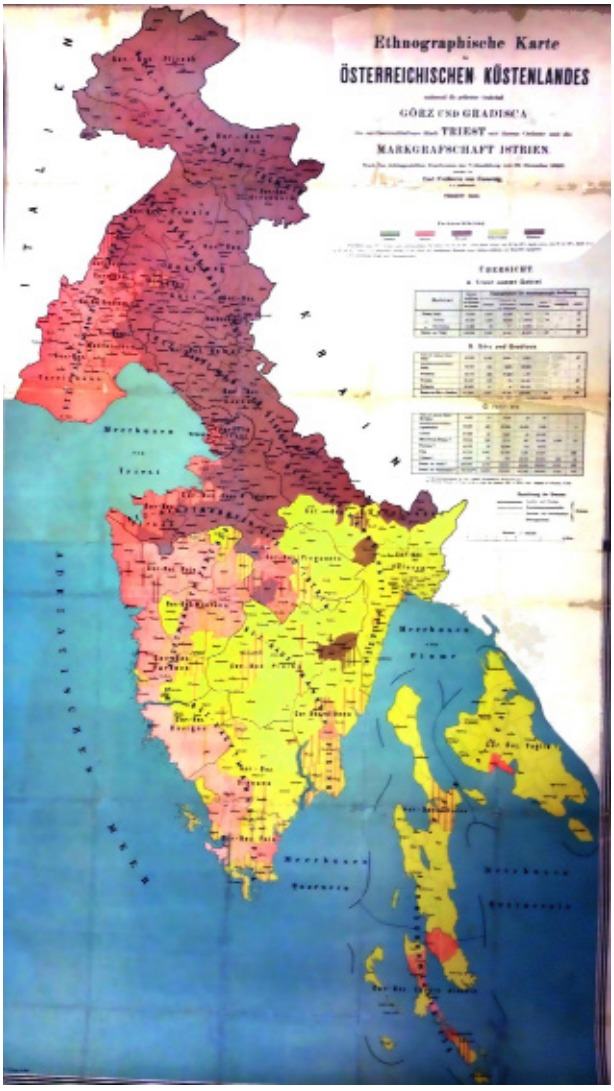

Fig. 1. Mappa della Contea Principesca di Gorizia e Gradisca – *Ethnographische Karte Österreichischen Küstenlandes umfassend die gefürstete Grafschaft Görz und Gradisca, die reichsunmittelbare Stadt Triest mit ihren Gebiete und die Markgraftschaft Istrien* di Carl von Czörnig (Trieste 1885)

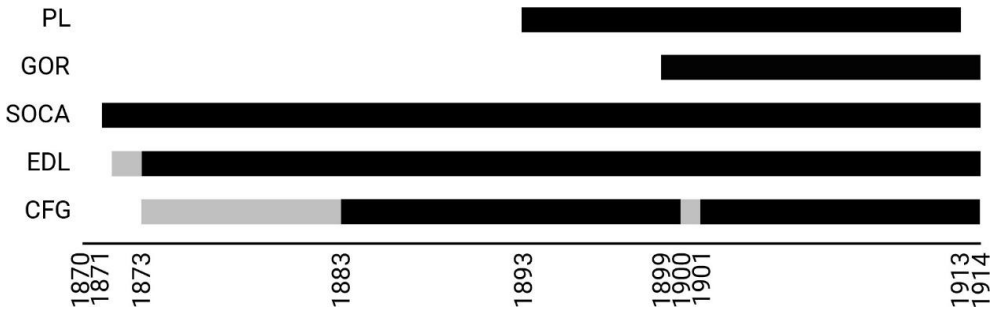

Fig. 2. Gli anni di pubblicazione delle 5 testate

Fig. 3. Le *word-clouds* con le parole più frequenti in ciascuna testata, danno un'idea del tipo di informazioni (ma anche di errori introdotti dalla fase OCR, e di approssimazioni introdotte dal *pre-processing*) che si trovano nel corpus digitale. Notiamo come la parola Gorizia / Gorica sia tra le più frequenti

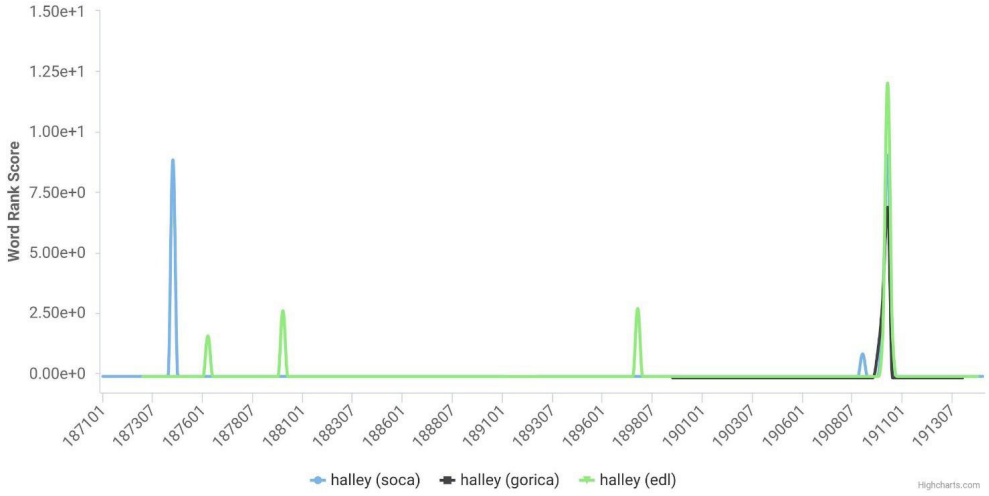

Fig. 4. Frequenza relativa della parola Halley fra 1899 e 1913

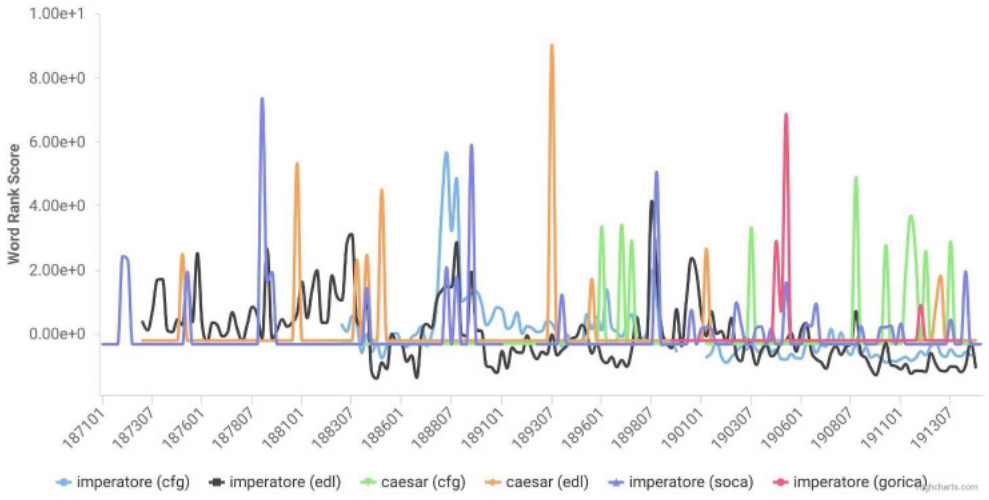

Fig. 5. Frequenza relativa della parola imperatore (*sl*: *cesar*)

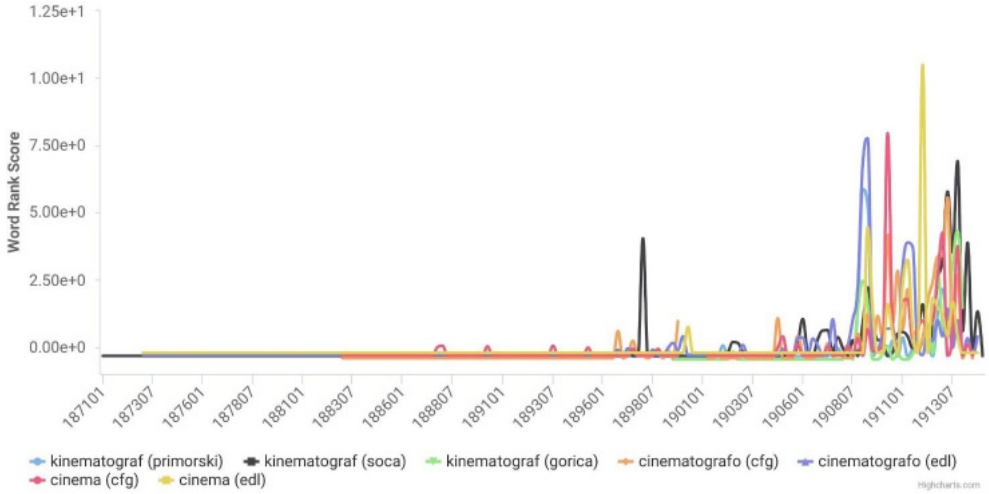

Fig. 6. Frequenza relativa degli *stems 'kinematograf-'*, *'cinematograf-'* e *'cinem-'* dal 1871 al 1915

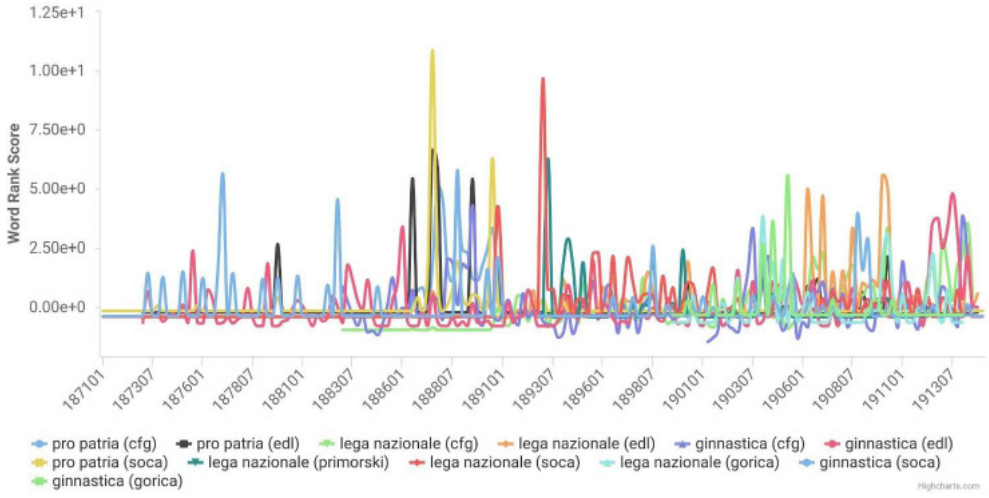

Fig. 7. Frequenza relativa di alcune delle principali associazioni nazionali italiane di Gorizia: Lega nazionale, Unione Ginnastica, Pro Patria

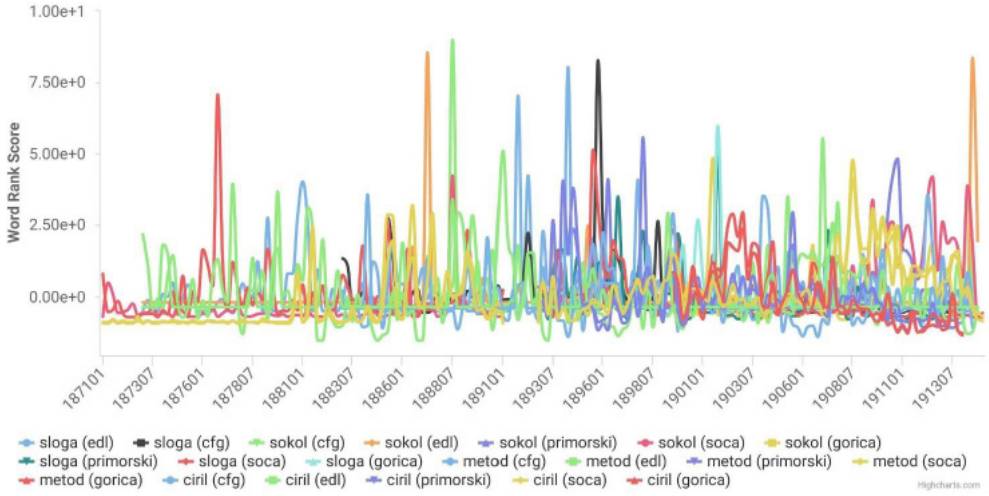

Fig. 8. Frequenza relativa di alcune delle principali associazioni nazionali slovene di Gorizia: *Ciril in Metoda*, *Sokol*, *Sloga*

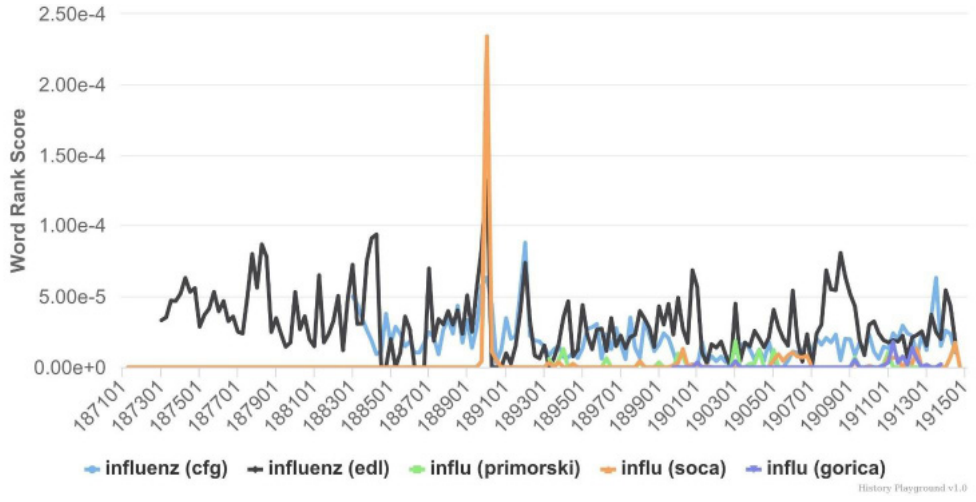

Fig. 9. La frequenza relativa della parola influenza nei giornali di entrambe le lingue

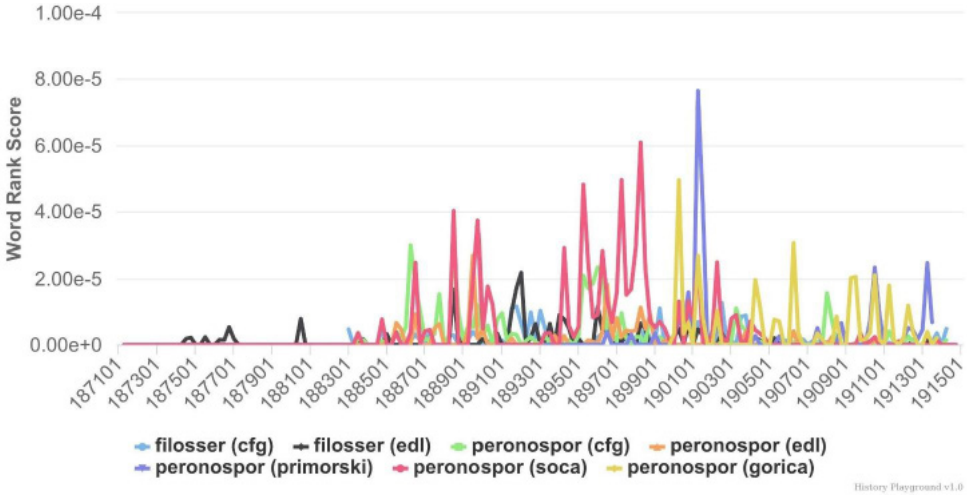

Fig. 10. La frequenza relativa delle parole filossera e peronospora

Bibliografia

Aiden - Michel 2013 = Erez Aiden - Jean-Baptiste Michel, *Uncharted: big data as a lens on human culture*, New York, Riverhead Books, 2013.

Agostinetti 1981 = Nino Agostinetti, *L'attività dei cattolici isontini nel primo ventennio del Novecento*, in *I cattolici isontini nel XX Secolo*, v. 1, *Dalla Fine dell'800 al 1918*, Gorizia, Le Casse Rurali e Artigiane della Contea di Gorizia, 1981.

Busa 1980 = Roberto Busa, *The annals of humanities computing: the index thomisiticus*, «Computers and the Humanities», 14 (1980).

Busa 1992 = Roberto Busa, *Half a century of literary computing: towards a "new" philology*, «Historical Social Research / Historische Sozialforschung», vol. 17 (1992), n. 2, p. 124-133.

Cavazza 2001 = Silvano Cavazza, *Gorizia e il territorio: considerazioni intorno al millenario goriziano*, «Il Territorio», 16 (2001), 2, p. 3-12.

De Claricini 1873 = Alessandro de Claricini, *Gorizia nelle sue istituzioni e nella sua azienda comunale durante il triennio 1869-1871: ricordo del podestà Alessandro nob. de Claricini ai diletti suoi concittadini*, Gorizia, Tip. Seitz, 1873.

De Grassi 1982 = Marino de Grassi, *Catalogo dei periodici stampati o editi nella Contea di Gorizia e Gradisca conservati nelle biblioteche pubbliche isontine 1774-1918*, «Studi Goriziani», 55-56 (1982), p. 51-104.

De Simone 1996 = Giuliana de Simone, *Catalogo dei periodici posseduti in microfilm dalla Biblioteca statale isontina*, «Studi Goriziani», 84 (1986), p. 131-144.

De Simone 2019 = Giuliana De Simone, *Il Progetto Google Books alla BSI*, «Studi Goriziani», 112 (2019), p. 52-56.

Dzogang et al. 2017 = Fabon Dzogang - Thomas Lansdall-Welfare - FMPN Team - Nello Cristianini, *Discovering periodic patterns in historical news*, «PloS one», 11 (2017) <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0165736>.

Dzogang et al. 2018 = Fabon Dzogang - Stafford Lightman - Nello Cristianini, *Diurnal variations of psychometric indicators in Twitter content*, «PloS one», 13 (2018) <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0197002>.

Fabi 1991 = Lucio Fabi, *Storia di Gorizia*, Padova, Il Poligrafo, 1991.

Feresin 2007-2008 = Vanni Feresin, *Fra Settecento e Novecento : la stampa a Gorizia*, «Isonzo Soča», 75-76, (2007-08), p. 14-21.

Ferrari 2002 = Liliana Ferrari, *Gorizia Ottocentesca, fallimento del progetto della Nizza Austriaca*, in *Storia d'Italia. Le regioni dall'Unità ad oggi*, v. 1, *Il Friuli Venezia Giulia*, a cura di Roberto Finzi, Claudio Magris, Giovanni Miccoli, Torino, Einaudi, 2002, pp. 313-375.

Filipi 2010 = Igor Filipi, *Stepišnik in Sveta Brata Ciril in Metod*, «Bogoslovni vestnik», 70, n. 1 (2010), p. 83-93.

Flaounas et al. 2010 = Ilias Flaounas - Marco Turchi - Omar Ali - Nick Fyson - Tijl De Bie - Nick Mosdell - Justin Lewis - Nello Cristianini, *The Structure of EU Mediasphere*, «PloS one», 12 (2010) <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0014243>.

Flaounas et al. 2012 = Ilias Flaounas - Omar Ali - Thomas Lansdall-Welfare - Tijl De Bie - Nick Mosdell - Justin Lewis - Nello Cristianini, *Research methods in the age of digital journalism*, «Digital Journalism», 1 (2012-2013) <https://www.tandfonline.com/doi/full/10.1080/21670811.2012.714928>.

Fogel Elton 1984 = Robert Fogel - Geoffrey Elton, *Which road to the past? Two views of history*, New Haven, Yale University Press, 1984.

Franzosi 2010 = Roberto Franzosi, *Quantitative narrative analysis*, Thousand Oaks, SAGE, 2010.

Franzosi 2011 = Roberto Franzosi, *On quantitative narrative analysis*, in James A. Holstein - Jaber F. Gubrium, *Varieties of narrative analysis,* Thousand Oaks, SAGE, 2011.

Franzosi et al. 2012 = Roberto Franzosi - Gianluca De Fazio - Stefania Vicari, *Ways of measuring agency: an application of quantitative narrative analysis to lynchings in Georgia (1875-1930)*, «Sociological Methodology», 42, n. 1 (2012) <https://journals.sagepub.com/doi/ab-

s/10.1177/0081175012462370?journalCode=smxa>.

Franzosi 2017 = Roberto Franzosi, *A third road to the past? Historical scholarship in the age of big data*, «Historical Methods: A Journal of Quantitative and Interdisciplinary History», 50, n. 4 (2017) <https://www.tandfonline.com/doi/abs/10.1080/01615440.2017.1361879>.

Gorian 2010 = Rudj Gorian, *Gazzetta Goriziana Editoria e informazione a Gorizia nel Settecento*, Trieste, Deputazione di storia patria per la Venezia Giulia, 2010.

Graham Milligan Weingart 2015 = Shawn Graham - Ian Milligan - Scott Weingart, *Exploring Big Historical Data: The Historian's Macroscope*, London, Imperial College Press, 2015.

Horel 2015 = Catherine Horel, *Austria-Hungary 1867-1914*, in Robert Justin Goldstein - Andrew M. Nedd, *Political Censorship of the Visual Arts in Nineteenth-Century Europe*, Basingstoke, Palgrave Macmillan, 2015.

Jerele et al. 2011 = Ines Jerele - Tomaž Erjavec - Daša Pokorn - Alenka Kavčič-Čolić, *Optical character recognition of historical texts: end-user focused research for Slovenian books and newspapers from the 18th and 19th century*, in *6th SEEDI Conference: Proceedings 16-20 May 2011, Zagreb, Croatia*.

Jia et al. 2016 = Sen Jia - Thomas Lansdall-Welfare - Saatviga Sudhahar - Cynthia Carter - Nello Cristianini, *Women are seen more than heard in online newspapers*, «PloS one», 11 (2016) <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0148434>.

Kacin-Wohinz - Troha 2000 = Milica Kacin-Wohinz - Nevenka Troha (a cura di), *Slovensko-italijanski odnosi 1880-1956. Poročilo slovensko-italijanske zgodovinsko-kulturne komisije / Rapporti italo-sloveni 1880-1956. Relazione della commissione storico-culturale italo-slovena / Slovene-Italian relations 1880-1956. Report of the Slovenian-Italian historical and cultural commission*, Lubiana, Nova revija, 2000.

Kalc 2013 = Alks Kalc, *Vidiki razvoja prebivalstva Goriške-Gradiške v 19. stoletju in do prve svetovne vojne / Some aspects of the demographic development in Goriška-Gradiška from early 19th century to WWI*, «Acta Histriae», 421 (2013) <https://www.dlib.si/details/URN:NBN:SI:-

DOC-F2SZUWYI>.

Kirsch 2014 = Adam Kirsch, *Technology Is Taking Over English Departments*, 2014 <https://newrepublic.com/article/117428/limits-digital-humanities-adam-kirsch>.

Lansdall-Welfare et al. 2014 = Thomas Lansdall-Welfare - Saatviga Sudhahar - Giuseppe Veltri - Nello Cristianini, *On the Coverage of Science in the Media: A Big Data Study on the Impact of the Fukushima Disaster*, in *Proceedings of the 2014 IEEE International Conference on Big Data*, New York, 2014. p. 60-66.

Lansdall-Welfare et al. 2017a = Thomas Lansdall-Welfare - Saatviga Sudhahar - James Thompson - Justin Lewis - FindMyPast Newspaper Team - Nello Cristianini, *Content analysis of 150 years of British periodicals*, «Proceedings of the National Academy of Sciences», 114-4, 2017 <https://www.pnas.org/content/114/4/E457>.

Lansdall-Welfare et al. 2017b = Thomas Lansdall-Welfare - Saatviga Sudhahar - James Thompson - Nello Cristianini, *The Actors of History: Narrative Network Analysis Reveals the Institutions of Power in British Society Between 1800-1950*, International Symposium on Intelligent Data Analysis, Cham, Springer, p. 186-197.

Lansdall-Welfare - Cristianini 2017 = Thomas Lansdall-Welfare - Nello Cristianini, *History Playground: A Tool for Discovering Temporal Trends in Massive Textual Corpora*, arXiv preprint, 04-06 (2017).

Marušič, B. 2005 = *Pregled politične zgodovine Slovencev na Goriškem (1848-1899)*, Nova Gorica, Goriški Muzej, 2005.

Medeot 1981 = Camillo Medeot, *Panorama Politico*, in *I Cattolici Isontini nel XX Secolo*, v. 1, *Dalla Fine dell'800 al 1918*, Gorizia, Le Casse Rurali e Artigiane della Contea di Gorizia, 1981.

Michel et al. 2011 = Jean-Baptiste Michel - Yuan Kui Shen - Aviva Presser Aiden - Adrian Veres - Matthew K. Gray - The Google Books Team - Joseph P. Pickett - Dale Hoiberg - Dan Clancy - Peter Norvig - Jon Orwant - Steven Pinker - Martin A. Nowak - Erez Lieberman Aiden, *Quantitative analysis of culture using millions of digitized books*, «Science», 331-6014 (2011), p. 176-182.

Mlakar - Turel 2010 = Liliana Mlakar - Annalisa Turel*, Storia di Gorizia*, Pordenone, Biblioteca dell'immagine, 2010.

Moretti 2013 = Franco Moretti, *Distant Reading*, London,Verso, 2013.

Nicholson 2012 = Bob Nicholson, *Counting culture; or, how to read Victorian newspapers from a distance*, «Journal of Victorian Culture», 172 (2012), p. 238-246.

Petzholdt 1853 = Julius Petzholdt, *Handbuch deutscher Bibliotheken*, Halle, H. W. Schmidt, 1853.

Redivo 2005 = Diego Redivo, *Le trincee della Nazione: cultura e politica della Lega Nazionale 1891-2004*, Trieste, Edizioni degli Ignoranti Saggi, 2005.

Spampinato 2018 = Daria Spampinato, *Prefazione* in *Settimo Convegno Annuale AIUCD 2018 Bari, 31 gennaio-2 febbraio 2018 Book of Abstracts*, Bari-Bologna, Associazione per l'Informatica Umanistica e la Cultura Digitale, 2018.

Sudhahar et al. 2015 = Saatviga Sudhahar - Gianluca de Fazio - Roberto Franzosi - Nello Cristianini, *Network analysis of narrative content in large corpora*, «Natural Language Engineering», 32, n. 1 (2013) <https://www.cambridge.org/core/journals/natural-language-engineering/article/network-analysis-of-narrative-content-in-large-corpora/7B1FFB891E8B3751016B2AE46FCF76C1>.

Sudhahar - Cristianini 2018 = Saatviga Sudhahar - Nello Cristianini, *Detecting Shifts in Public Opinion: A Big Data Study of Global News Content*, in *Advances in Intelligent Data Analysis XVII. IDA 2018. Lecture Notes in Computer Science, vol 11191*, edited by Wouter Duivesteijn, Arno Siebes, Antti Ukkonen, Springer, 2018.

Von Czoernig 1987 = Carl von Czoernig, *Gorizia, la Nizza austriaca,* Cassa di risparmio di Gorizia, 1987, (tit. orig.: *Görz: Oesterreich's Nizza: nebst einer Darstellung des Landes Görz und Gradisca*, Braumüller, 1873-74).

Abstract

Le biblioteche digitali consentono non soltanto di migliorare la conservazione dei documenti e di facilitarne l'accesso da parte degli utenti, ma anche di sperimentare nuovi metodi; ad esempio è possibile esaminare in tempi ridotti le relazioni statistiche tra i contenuti di migliaia di documenti, operazione pressoché inaccessibile ai metodi tradizionali. Il passaggio chiave resta quello della conversione dal supporto analogico, carta o microfilm, a quello digitale, includendo la trasformazione delle immagini del testo stampato in testo digitale: solo così è possibile analizzare statisticamente quei testi, analisi che del resto non può prescindere dal contesto storico della loro produzione e da altre fonti. In questo articolo, descriviamo in dettaglio il processo di creazione di un corpus digitale formato dai giornali italiani pubblicati a Gorizia tra il 1873 e il 1914. Questo include la digitalizzazione, l'estrazione del testo editabile, il processo di annotazione e l'analisi statistica delle risultanti serie temporali. I dati così ottenuti vengono comparati con un corpus di giornali sloveni stampati nella stessa città e nello stesso periodo, già digitalizzati dalla Biblioteca Nazionale Slovena. L'analisi delle 47.466 pagine di giornali italiani ci consente di dimostrare il tipo di informazioni che possono essere estratte da un corpus digitale, evidenziando l'importanza di operare all'interno di un contesto storico e comparativo. Questo esempio di umanistica digitale plurilinguistica ci consente di individuare le tracce statistiche di profonde transizioni culturali che hanno avuto luogo in un'area geografica e un periodo storico molto complessi, il cui studio non può prescindere da una particolare attenzione alle trasformazioni culturali, tecnologiche e sociali.

biblioteche digitali; giornali; periodici; digitalizzazione

Digital libraries allow not only to improve the preservation of documents and to facilitate access by users, but also to experiment with new methods;

for example, it is possible to examine the statistical relationships between the contents of thousands of documents in a short time, an operation almost inaccessible to traditional methods. The key step remains that of converting from analogue support, paper or microfilm, to the digital one, including the transformation of images of the printed text into digital text: only in this way is it possible to statistically analyze those texts, an analysis that cannot be separated from the historical context of their production and from other sources. In this article, we describe in detail the process of creating a digital corpus formed by Italian newspapers published in Gorizia between 1873 and 1914. This includes digitization, editable text extraction, annotation process and statistical analysis of the resulting time series. The data thus obtained are compared with a corpus of Slovenian newspapers printed in the same city and at the same time, already digitized by the Slovene National Library. The analysis of the 47.466 pages of Italian newspapers allows us to demonstrate the type of information that can be extracted from a digital corpus, highlighting the importance of operating within a historical and comparative context. This example of multilingual digital humanism allows us to identify the statistical traces of profound cultural transitions that have taken place in a very complex geographical area and historical period, whose study cannot ignore a particular attention to cultural, technological and social transformations.

Digital libraries; *newspapers*; *journals*; *digitization*