UNIVERSITÀ DEGLI STUDI DI CATANIA

DIPARTIMENTO DI SCIENZE UMANISTICHE

DOTTORATO DI RICERCA IN FILOLOGIA MODERNA

XXVIII CICLO

SALVATORE ARCIDIACONO

ELEMENTI DI LESSICOGRAFIA COMPUTAZIONALE PER UN VOCABOLARIO DEL SICILIANO MEDIEVALE (VSM)

_____

TESI DI DOTTORATO

_____

COORDINATORE:
Chiar.mo prof. ANTONIO DI GRADO

TUTOR:
Chiar.mo prof. MARIO PAGANO

_____

ANNO ACCADEMICO 2015-2016

1. INTRODUZIONE

Qualora si voglia riconoscere alla lessicografia computazionale, se non il privilegio di un'identità disciplinare autonoma[1], una valenza referenziale che denoti un insieme coeso di problemi, si potrebbe avvertire un leggero senso di disorientamento dovuto all'assenza di una tradizione scientifica consolidata[2]. Prima che di delicate interferenze metodologiche o di possibili idiosincrasie concettuali, si tratta, più banalmente, della mancanza di un percorso canonico e della difficoltà insita nel dover in parte rifigurare gli obiettivi, le pratiche di indagine e la natura dei risultati attesi. Lo scardinarsi delle tassonomie scientifiche, in un contesto di forti incroci disciplinari, e la veemente accelerazione digitale scoraggiano, come noto, il ricorso e la piena adesione alle teorie e ai modelli che appena qualche decennio fa rappresentavano lo stato dell'arte.

---

[1] La definizione dello statuto della lessicografia computazionale — per quanto verrà inevitabilmente più volte richiamato nei capitoli seguenti — esula dagli obiettivi del lavoro. Il termine sarà usato in accordo con Granger (2012: 2): «in this volume electronic lexicography is used as an umbrella term to refer to the design, use and application of electronic dictionaries (EDs), which are in turn defined as primarily human-oriented collections of structured electronic data that give information about the form, meaning and use of words in one or more languages and are stored in a range of devices (PC, internet, mobile devices)».

[2] Sulla difficoltà di integrare in un settore disciplinare riconosciuto le tematiche relative all'Informatica Umanistica — e le conseguenti ricadute negative sull'innovazione tecnologica e sulla conservazione del patrimonio culturale — si vedano almeno Roncaglia (2002), Orlandi (2002) e Buzzetti (2014).

Nel nostro caso lo spaesamento sintomatico deriva dalla forte matrice tecnologica del progetto *Artesia* (*Archivio Testuale del Siciliano Antico*[3]), da cui questa ricerca prende avvio nel momento in cui i dati raccolti negli ultimi dieci anni di attività sembrano soddisfare le condizioni per assecondare la vocazione quiescente di un'indagine lessicografica ad ampio raggio.

Muovendo dal ricco complesso informativo di partenza, il problema si configura quindi, in prima istanza, come ricerca di un modello progettuale e come messa a punto di un flusso di lavoro ideale, accompagnata — ma solo in un secondo tempo — dalla realizzazione degli strumenti informatici più adatti al raggiungimento dei risultati previsti. Pur nell'impossibilità di seguire un tracciato prestabilito, ci si dovrebbe domandare, quindi, fino a che punto sia lecito, e in quali termini, reinventare la propria ricerca in ciascuno dei diversi momenti in cui si sviluppa. Cedere alle lusinghe delle nuove tecnologie come a un invito implicito alla libertà assoluta, ed esercitare una completa autonomia di scelte in un vuoto privo di punti di riferimento, pur nel pieno rispetto degli

[3] Così come si legge nella homepage del progetto, <http://artesia.unict.it>, «*ARTESIA* (*Archivio Testuale del Siciliano Antico*) vuole essere un articolato strumento di studio sul siciliano medievale. Rende accessibile alla lettura (fatti salvi i diritti di copyright) e alla interrogazione, con il programma GATTOWEB, oltre ad una selezione di documenti, il corpus, filologicamente attendibile e periodicamente aggiornato, dei testi letterari e paraletterari scritti in volgare siciliano dalle prime attestazioni del XIV secolo sino alla prima metà del XVI. Fornisce di ogni autore e testo una breve presentazione, in modo da coniugare alle istanze della documentazione un approccio storico-critico che metta in evidenza i rapporti con la tradizione latina e con altre tradizioni testuali romanze (toscana, catalana, etc.). Pubblica in versione elettronica studi e ricerche attinenti, anche indirettamente, al siciliano medievale. Contribuisce a porre le basi documentarie per la redazione di un Vocabolario del siciliano medievale». Per una presentazione del progetto si veda Pagano (2009). Il *Corpus Artesia* è accessibile all'indirizzo <http://www.ovi.cnr.it>.

assunti propri del progetto originario, riserva alte probabilità di rischio e l'ingovernabilità dei risultati finali.

In mancanza di un paradigma teorico vincolante, l'unico appiglio è la conoscenza critica dei più recenti sviluppi raggiunti nella prassi lessicografica. L'analisi della situazione corrente, delle *best pratices*, può essere un valido surrogato di un paradigma teorico forte. Nei termini di Gouws (2012: 17) — in un contributo significativamente intitolato «Learning, Unlearning and Innovation in the Planning of Electronic Dictionaries» — l'approccio alla progettazione di un dizionario elettronico si gioca tra un polo "contemplativo" (*contemplative*) e un polo "trasformativo" (*trasformative*)[4]. Se il secondo termine è essenziale all'elaborazione di nuove idee, riservare uno spazio al momento contemplativo giova a evitare che l'innovazione prodotta in un momento particolarmente dinamico per la lessicografia passi inosservata e si disperda.

Quanto detto si rintraccia e conserva piena validità anche in un'ottica disciplinare allargata. Recentemente, il Research Information Network (RIN[5]) ha commissionato una serie di tre *report*, il secondo dei quali dedicato alle scienze umane[6], che

---

[4] «We often make a well-motivated distincion between a *contemplative* and a *trasformative* approach in lexicography. While the former focuses on an investigation of the prevaling situation, the latter opts for a development from a current situation to something new», Gouws (2012: 17).

[5] Il Research Information Network è una commissione fondada *dall'UK higher education funding councils*, i sette research councils e le tre biblioteche nazionali del regno unito. <http://www.rin.ac.uk/>, ultima consultazione: 03 settembre 2015.

[6] AA.VV. (2011). Il report è il risultato di uno sforzo collaborativo tra il RIN e studiosi afferenti a un prestigioso gruppo di centri di ricerca: Oxford Internet Institute (OII), University of Oxford, UCL Centre for Digital Humanities and the Department of Information Studies, UCL, e-Humanities Group, Royal

mirano a conoscere approfonditamente gli approcci più recenti dei ricercatori relativi alla scoperta, l'accesso, l'analisi, la gestione, la creazione, il raffinamento e la disseminazione delle risorse informative nelle scienze umane. Il progetto è stato emblematicamente battezzato *Reinventing Humanities*, ed è significativo che la risposta ai quesiti impliciti nel titolo e negli scopi sia affidata a una raccolta di *case studies*, come a lasciare supporre che i tempi non sono ancora maturi per una sentenza definitiva e che solo l'osservazione delle realtà più innovative può condurre alle risposte desiderate.

Il primo capitolo sarà quindi dedicato al dovuto inquadramento disciplinare. Si ripercorreranno rapidamente le origini dell'informatica umanistica con un triplice scopo: contestualizzare la ricerca nel suo dominio; ritrovare le origini della lessicografia computazionale; costruire una prima mappa di punti di riferimento sul modo di intendere il cambio di paradigma attraverso una riflessione epistemologica che parte dall'incontro tra parole e *bit*. Si introdurrà il concetto di *dizionario elettronico* in senso forte, dove, con questa espressione si vuole indicare la concretizzazione ideale dell'innovazione digitale applicata alla lessicografia.

Il secondo capitolo mira empiricamente alla raccolta di spunti e parametri di riferimento attraverso un'analisi storico-critica delle imprese più significative nel campo della lessicografia storica — dalla retroconversione del *Vocabolario degli Accademici della Crusca*, alle riconversioni dell'*Oxford English Dictionary* (*OED*) e *dell'Anglo-Norman Dictionary* (*AND*), fino ai più recenti dizionari *born digital*. Il capitolo si incarica di sviluppare la parte contemplativa[7] del lavoro ma con una trattazione pragmaticamente

---

Netherlands Academy of Arts & Sciences (KNAW), Maastricht University, Oxford e-Research Centre (OeRC), University of Oxford.

[7] Il riferimento è ancora a Gouws (2012: 17).

finalizzata al reperimento di alcuni rilevanti principi costitutivi del dizionario elettronico.

Il terzo capitolo riprenderà e approfondirà singolarmente alcuni degli elementi salienti del dizionario elettronico, rilevati anche nel corso dell'analisi precedente, cercando di dare a ciascuno il giusto rilievo problematico, superando la contingenza del singolo caso concreto.

Il quarto capitolo completa il quadro con la trattazione del problema della codifica digitale in lessicografia. Oltre ad affrontare il grado zero della realizzazione di un vocabolario elettronico, ha la funzione di introdurre, censire e valutare le alternative tecniche per la successiva applicazione sperimentale.

L'informatica tende spontaneamente a riconfigurare l'indagine in termini tecnici con il rischio di affidare la "responsabilità" al "responso" tecnico[8] e di cadere in una fredda autoreferenzialità, ma è anche vero che solo quando viene calata in un ambiente applicativo una ricerca simile può trovare valore. Privata del dominio di riferimento, si svuoterebbe di validità sperimentale. L'aspetto progettuale[9] in una ricerca nel campo delle *digital humanities* è un elemento di primaria importanza. La parte successiva del lavoro, che focalizzerà l'attenzione sul contesto applicativo reale, è la logica conseguenza della prima e ne rappresenta l'ideale completamento 'sul campo'.

---

[8] Il riferimento è all'intepretazione della tecnica come "assoluto astorico" di Galimberti (2000: 40-41),

[9] «Si è passati a una fase in cui, accanto alla competenza metodologica, linguistica e storico-culturale del singolo (comunque insostituibile), è diventato sempre più importante il progetto e la capacità di selezionare e porre in relazione gli innumerevoli dati e le potenzialità offerte dalla rete» (Antonelli 2011: VIII).

Col quinto capitolo si ritornerà alla fonte documentaria del progetto, il *Corpus Artesia*. A partire dall'analisi della rappresentatività del *Corpus* viene affrontato il problema del bilanciamento del sottocorpus dei testi d'archivio, limite riconosciuto delle precedenti edizioni[10]. La ricerca ha quindi curato la codifica di oltre 60 testi per l'indicizzazione in GATTO (Gestione degli Archivi Testuali del Tesoro delle Origini) e ha condotto alla realizzazione di una nuova versione della base di dati che ha il merito di risolvere il problema del mancato bilanciamento dei testi non letterari. Il lavoro è corredato dalla realizzazione di un'interfaccia pensata per la distribuzione su CD-ROM. La sezione si conclude con alcuni sondaggi lessicali a campione per valutare il contributo dell'aggiornamento alla rappresentazione del lessico della quotidianità e della cultura materiale. Accompagna il lavoro il necessario corredo di documentazione — di prossima pubblicazione nella collana dei «Quaderni di Artesia[11]» e inserito in appendice alla tesi — che dovrebbe seguire il rilascio di ogni risorsa elettronica (uno dei *desiderata* del progetto Artesia).

Il capitolo 6 è dedicato alla definizione microstrutturale del *Vocabolario*, alla quale si assegna il ruolo di nucleo formale per l'impianto lessicografico e lo sviluppo degli strumenti informatici. Il modello di voce viene definito, punto per punto, riadattando il prestigioso modello del *Tesoro della Lingua Italiana delle Origini*[12] e operando una parallela riconversione dell'articolazione logica della voce in XML/TEI[13]. Con il modello XML si perviene a una struttura logica sulla quale poter fondare la realizzazione di tutti i successivi strumenti informatici. Il capitolo si chiude con la

---

[10] Arcidiacono (2011: 286), Pagano (2012: 120); Pagano / Arcidiacono (2013).

[11] Arcidiacono in c.d.s.

[12] <http://www.ovi.cnr.it>. Cfr. §2.7.

[13] Cfr. §4.2.

presentazione dell'architettura e una prima implementazione per un sistema di inserimento guidato, un'interfaccia di redazione collaborativa accessibile on-line, con strumenti di autocompletamento e controllo dell'integrità referenziale.

Per riassumere, questa tesi è idealmente divisa in due parti: un saggio di lessicografia computazionale e una sezione applicativa focalizzata sul progetto del *VSM*. Il lavoro si completa però oltre i confini dell'elaborato cartaceo con una serie di prodotti e applicazioni sperimentali che hanno puntualmente accompagnato i nodi cruciali del percorso: la realizzazione di una nuova edizione della base di dati; l'interfaccia per la pubblicazione del *Corpus*; la documentazione per la nuova risorsa elettronica; la piattaforma web per la redazione l'archiviazione e la consultazione del *VSM*. Questo piccolo sistema di applicazioni e risorse non è un mero complemento al lavoro ma ne fa parte integrante a pieno titolo in una piena e coerente adesione a quel modo di intendere le nuove pratiche di ricerca secondo il quale

> l'Umanistica Digitale sollecita una radicale ridefinizione delle discipline umanistiche come impresa generativa: un'impresa che vede studenti e docenti impegnati a produrre "cose" nel momento in cui essi studiano e mettono in atto differenti pratiche di ricerca. Queste cose non si limitano ai testi (analisi, commentari, narrazioni, critiche), ma comprendono anche immagini, interattività, materiali cross-mediali, software e piattaforme[14].

Per concludere, vorrei affidare il compito di riassumere e sintetizzare lo spirito con il quale ho affrontato questo lavoro a una

[14] Burdick et al. (2012: 23).

citazione di uno studioso che ci ha lasciati da appena qualche mese[15]:

> A lexicographer is a divided soul, part scientist, part tool-builder. The scientist is a linguist, wanting to describe the language. The tool-builder wants to help the user find the information they want, the territory of information science. Lexicography is in the intersection[16].

Questa interpretazione della lessicografia spiega le corrispondenze simmetriche che scandiscono l'intero percorso: alle teorie della rappresentatività linguistica corrisponde l'allestimento del *Corpus Artesia 2015*, alla metalessicografia della microstruttura un modello XML-TEI, alla codifica del *Corpus* il manualetto in appendice. La speranza è che, nell'incontro tra queste linee che

[15] Per la stesura di questo lavoro si è fatto largo uso delle risorse liberamente disponibili sulla sua pagina <https://kilgarriff.co.uk/>, ultima consultazione: 8 novembre 2015. Al ricordo di Kilgarriff va anche unito quello di Alberto Varvaro, scomparso meno di un anno fa. Se non bastasse la straordinaria importanza della sua intera produzione scientifica, la sua ultima opera, il *VSES*, rappresenta oggi il più grande esempio di lessicografia diacronica per il siciliano. La consulenza che ho avuto il piacere e l'onore di prestare per la digitalizzazione del primo volume del *VES*, poi confluito nel *VSES*, costituisce un'esperienza di codifica lessicografica sulla quale si basano molte delle scelte operate in questo lavoro (con riferimento particolare al §4.1). Nel corso dell'ultima revisione sono stato purtroppo costretto ad aggiungere al ricordo David Trotter, uno degli autori più citati in tutto il lavoro, più che per il suo ruolo nella redazione dell'*AND*, per la capacità di saper inquadrare i punti cruciali della lessicografia computazionale e trattarli con estrema semplicità e chiarezza.

[16] Kilgarriff (2012: 26).

corrono parallele, si possa trovare quella lessicografia[17] di cui parlava Kilgarriff.

[17] In senso ampio, visto il contesto propriamente *pre-lessicografico* («the planning stages of a dictionary project», Atkins / Rundell 2008: 15) e la prevalenza di argomenti metalessicografici.

1 ELEMENTI DI LESSICOGRAFIA COMPUTAZIONALE

1.1 NUOVE COORDINATE EPISTEMOLOGICHE

«Think. Il difficile lo facciamo subito, l'impossibile richiede un po' più di tempo[18]». Con questo storico slogan della IBM, Roberto Busa Sj, padre ideale dell'Informatica Umanistica, riuscì a convincere Thomas Watson, fondatore dell'azienda, a tentare la strada del trattamento computazionale della parola, inizialmente ritenuto impossibile.

Lo scettico atteggiamento iniziale di Watson nei confronti delle proposte dello studioso, dietro lo spettro dell'impossibilità tecnica forse nascondeva un istintivo atteggiamento di sfiducia nei confronti di un possibile legame tra parole e calcolo, una prefigurazione – si sarebbe tentati di dire – delle resistenze che le scienze umane opposero per anni all'«indebita intrusione[19]» dei metodi computazionali.

Padre Busa, spinto da un moto di delusione, rispose al rifiuto restituendo al suo legittimo proprietario il biglietto da visita con il motto aziendale stampato vicino al nome del dirigente. Fu quel

[18] Il racconto del colloquio tra padre Busa e Thomas Watson è abbondantemente diffuso in rete (anche su testate autorevoli come lastampa.it, vatican.va), l'origine del resoconto credo vada fatta risalire all'intervista che Stefano Lorenzetto publicò su *Il Giornale* il 3 ottobre 2010 e ancora disponibile all'indirizzo <http://www.ilgiornale.it/news/e-gesuita-cre-link-merito-suo-se-navigate-internet.html>. Ultima consultazione: 18 agosto 2015.

[19] Orlandi (1987: 15).

gesto a indurre Watson a riconsiderare i suoi pregiudizi. Da quel giorno del 1949 in cui, grazie a quel colloquio, si instaurò una fruttuosa interazione tra computer e discipline linguistiche e letterarie, l'applicazione delle tecnologie digitali alle scienze umanistiche non ha mai cessato di obbligare ricercatori e studenti a ripensare l'oggetto e la natura stessa delle loro ricerche e a stimolare l'immaginazione di scenari innovativi, e rivoluzionari.

Roberto Busa divenne, almeno per la tradizione italiana[20], il padre ideale dell'Informatica Umanistica[21]. Lo studioso, grazie ai risultati di quell'accordo, riuscì a concludere il suo lavoro, l'*Index Thomisticus*[22], concordanza lemmatizzata dell'opera di Tommaso d'Aquino, in appena un paio di anni.

A voler ben vedere, se le concordanze possono essere considerate a pieno titolo «comme des dictionnaires particuliers où les définitions sont remplacées par le contexte naturel, qui donne au mot-vedette sa valeur, la signification du mot-vedette étant comprise comme la somme se ses emplois[23]», l'*Index* rappresenta e sancisce idealmente anche l'atto di nascita della lessicografia computazionale.

---

[20] In Italia, agli spogli elettronici di padre Busa presso il Centro per l'Automazione dell'Analisi Linguistica (CAAL) di Gallarate, fondato nel 1946 (Gigliozzi 2003: 4), si aggiunsero, nel 1965, le attività del CNUCE (Centro Nazionale Universitario di Calcolo Elettronico) di Pisa, dove fu istituita una divisione per la linguistica diretta da Antonio Zampolli (cf. Zampolli 2003).

[21] Si veda a questo proposito la breve rassegna bibliografica in Raynaud (2009: 57).

[22] Busa (1951).

[23] Hanon (1990: 1563).

1.2 INFORMATICA E LESSICOGRAFIA

I vocabolari, in tutte le loro forme, si rivelano nei decenni successivi un terreno estremamente fertile per l'informatica e il trattamento algoritmico dei dati. La sinergia è stata possibile grazie a un essenziale isomorfismo tra i rispettivi oggetti trattati[24]:

> l'interrogazione informatica si applica bene agli elementi formali come i grafemi, i lessemi e i morfemi, il che spiega la sua particolare utilità nella lessicografia storica: è facile sapere se la frequenza di una parola cambia in funzione dei parametri diasistematici; si può anche quantificare il contesto sintagmatico e identificare così eventuali fraseologismi[25].

Oggi sarebbe quasi impossibile pensare a un nuovo dizionario escludendo il ricorso alle tecnologie digitali sia sul versante della redazione sia negli utilizzi previsti. Le zone di contatto tra informatica e discipline linguistiche non sempre hanno dato luogo a interazioni così proficue e intense, e questo basta ad affermare che quello tra informatica e lessicografia è diventato un rapporto simbiotico eccezionalmente stabile, duraturo e, per molti aspetti, ormai indissolubile: «today lexicography is largely synonymous with electronic lexicography and many specialists predict the disappearance of paper dictionaries in the near future[26]».

La lessicografia storica non fa eccezione, anzi, sembra averne tratto una rinnovata importanza dai cambiamenti che ha attraversato nel corso della rivoluzione digitale[27].

---

[24] Sinergia catalizzata poi da altri fattori che verranno individuati nel corso di questo capitolo e approfonditi nei seguenti.

[25] Gleßgen (2006: 16).

[26] Granger (2012: 2).

[27] «Mi sembra che la lessicografia storica sia oggi potenzialmente molto più importante di quanto questa non lo fosse in passato. […] Mentre in passato il

Oggi il peso e il volume fisico dei 56 tomi a stampa dell'*Index Thomisticus* sembra esprimere la distanza che ci separa da quella prima fase dei padri fondatori[28]. La potenza di calcolo della IBM aveva contribuito a produrre un'opera rivoluzionaria ma, sulla pagina stampata ritroviamo un dizionario tradizionale, «il quale nella sua rigida monumentalità di opera stampata si comporta come il libro secondo Socrate: se interrogato, non risponde[29]». La comprensibile cognizione dello scarto che intercorre tra la nostra pratica quotidiana del digitale e le rudimentali tecnologie di epoche a noi distanti solleva un valido dubbio: la nascita del calcolatore nel quinto decennio dello scorso secolo, e la spettacolare irruzione nel mondo delle scienze umane sono forse manifestazioni di una rivoluzione ancora in corso, il cui completo compimento sarebbe arrivato in un momento successivo. Se oggi questo momento si trova alle nostre spalle o deve ancora venire non è dato sapere: l'innovazione è per certi aspetti imperscrutabile quando le sue dinamiche sono ancora attive.

---

dizionario era essenzialmente uno strumento per cercare informazioni riguardo a una precisa parola, da un po' di tempo, io insieme ad altri, abbiamo incominciato a sostenere l'idea che, grazie alla tecnologia, i vocabolari storici possono essere considerati come portali per l'accesso all'intero sistema della lingua, e in generale al mondo dell'informazione e della conoscenza. La parola cercata su un dizionario non rappresenterebbe più la fine di una ricerca, ma soltanto l'inizio», Simpson (2013: 36).

[28] Per il periodo compreso tra le origini e gli anni '60 cfr. Lenders (2013).

[29] Nencioni (1987: 157).

1.3 VERSO UN CAMBIO DI PARADIGMA

Se il prodotto dell'applicazione informatica rimanesse virtualmente indistinguibile da quello a cui potrebbero pervenire uno o più operatori umani — pur con parecchio tempo, energie e carta a disposizione — le metodologie computazionali non potrebbero sperare di ottenere più credito di quanto non sia concesso a un puro ausilio meccanico[30]; nel nostro, come in qualunque ambito di applicazione, gli esiti del trattamento computazionale si distinguerebbero solo per semplice scarto quantitativo. Oggi, a fronte di un elevato livello di penetrazione delle tecnologie digitali in numerosi domini di ricerca, le implicazioni sembrano ben più consistenti, al punto di poter affermare senza timore di smentita che «l'applicazione di strumenti informatici alla ricerca umanistica (come d'altra parte a ogni dominio), infatti non è mai metodologicamente innocente[31]». Si fa riferimento da più parti[32] a un presunto cambio di paradigma[33] in corso, «un saut de qualité terriblement grand et terriblement exigeant[34]».

Rimane però da determinare quando l'interazione disciplinare raggiunge un livello di sinergia tale da annunciare la nascita di un nuovo paradigma e per quale motivo. Senza voler entrare nel merito di controversie troppo delicate per questo

---

[30] Cfr. Roncaglia (2002).

[31] Ciotti (2005: 9).

[32] Per citare un caso esemplare, nel recente volume-manifesto di Burdick, Drucker, Lunenfeld, Presner e Schnapp, le *digital humanities* vengono paragonate a una rivoluzione copernicana a vasto raggio, che trasforma anche il ruolo delle istituzioni, delle biblioteche e dei musei (Burdick et al. 2012: 48)

[33] Il riferimento è, ovviamente, a Kuhn (1962). Per una panoramica sulle teorie del progresso scientifico si veda Losee (2009: 200-210).

[34] Busa (1990-1991: XI).

contesto, ci si limiterà, per il momento e per puro scopo pratico, a individuare tre elementi salienti, rilevati in un confronto per contrasto tra l'attuale stato dell'arte e le prime applicazioni: la qualità del trattamento, il completo impiego dei supporti, la nascita di problematiche autonome.

1.3.1 Qualità del trattamento

Agli albori della disciplina, il calcolatore, la cui storia in quel momento non supera le due decadi, sperimenta le prime applicazioni nelle scienze umane, con momenti particolarmente significativi nella generazione di concordanze e indici[35]. Si fa riferimento a questa fase, compresa tra gli anni '50 e '60, con l'espressione di "stage of language data collection", prima fase di trattamento del materiale linguistico con trattamento elettronico del testo (Lenders 2013: 970). Si sviluppano le tecniche di automazione per il trattamento del testo e si comincia a maturare una sensibilità per i sistemi di codifica del testo, effettuata all'epoca su schede perforate.

In molti casi si riscontra un'elaborazione poco matura e, complessivamente, si avverte la mancanza delle moderne tecniche di marcatura e disambiguazione.
Se ne ricava, per contrasto con lo stato dell'arte attuale, una serie di differenze significative, da collocare sull'asse ideale delle metodologie di trattamento:

1. Assenza di *normalizzazione.* Diverse forme grafiche di una stessa parola vengono trattate come elementi differenti e non

[35] «L'usage confond souvent concordance et index. Les index sont des inventaires de mots dotés d'une référence numérique, mais dépourvus de contexte. Les conc. sont souvent munies de donnés statistiques», Hanon (1990: 1563).

sono raccolti sotto una notazione normalizzata.

2. Assenza di *lemmatizzazione.* Le forme flesse di una parola rimangono indipendenti tra loro e non sono accomunate da una relazione con una parola-vedetta o lemma.

3. Assenza di *disambiguazione morfologica.* Gli omografi non sono separati.

4. Mancanza di *disambiguazione semantica.* Le informazioni semantiche sono spesso insufficienti. In relazione allo specifico tipo di dizionario, mancano sinonimi esplicativi o sufficienti contesti.

Sulla base di questi quattro punti si può per il momento concludere — ma l'argomento meriterebbe analisi più approfondite — ipotizzando che i primi trattamenti computazionali fossero stimolati e basati sulle istruzioni fondamentali della programmazione. Non traspare nulla che vada oltre l'applicazione dei cicli e delle operazioni di confronto su stringhe. Anche se oggi disponiamo di istruzioni native in ogni linguaggio di programmazione è difficile dire quanto potessero essere complesse operazioni del genere sulle macchine dell'epoca; venne percepita sicuramente la possibilità immediata di risolvere compiti estremamente laboriosi per mezzo delle procedure iterative (*loop*). Nell'epoca in cui i computer venivano costruiti assemblando assieme migliaia di valvole termoioniche, quello che oggi appare come un semplice *foreach* (ma, si ricordi, che questa istruzione non era presente neppure nel linguaggio C) poteva essere visto come la chiave per accedere a una nuova dimensione di manipolazione intensiva del testo.

1.3.2 Nuovi media

Al raffinamento dell'informazione, se ne è fatto cenno prima, va aggiunta la differenza di *medium*. Basterebbe ricordare il celebre passo McLuhan[36]:

> Per quanto riguarda le conseguenze pratiche, il *medium* è il messaggio. Che in altre parole le conseguenze individuali e sociali di ogni *medium*, cioè di ogni estensione di noi stessi, derivano dalle nuove proporzioni introdotte nelle nostre questioni personali da ognuna di tali estensioni o da ogni nuova tecnologia.

Tornando all'*Index Thomisticus* di padre Busa, l'opera è priva delle carenze tipiche di quel periodo elencate nel paragrafo precedente. Si consideri il trattamento degli omografi [37] o la lemmatizzazione che associa ad ogni lemma un codice numerico progressivo e che prevede, nell'elegante ordinamento della concordanza, trattamenti diversificati a seconda della categoria grammaticale[38]. Al termine del trattamento digitale che l'ha reso possibile, l'*Index* però affida la consultazione del contenuto al classico volume a stampa. Il prodotto finale della ricerca si presenta nella tradizionale veste di carta e inchiostro, azzerando le capacità logiche e "attive" della macchina. La stampa non è quello che oggi

---

[36] McLuhan (1964: 15). Per una trattazione del determinismo tecnologico da un punto di vista vicino al nostro si veda Fiormonte (2003).

[37] Busa (1951, I, 1: XI).

[38] «Intra lemma formae ordinantur diversa ratione. Formae nominales, invariabiles et speciales ordinantur prius iuxta gradum (positivus, comparativum, superlativum), sed intra gradum ordinantur simpliciter iuxta graphiam: quare, eo quod automatum distinguit inter i et j, forma pejori praecedit pejor, pejora, etc. Formae autem verbales semper atque omnes ordinantur ea ratione qua in grammaticae paradigmatibus, ut legenti evidens erit. Ubique tamen ultimo loco adveniem formae stricto sensu compositae, qualis es v.g. maleacta de qua videsis», Busa (1951, II, 1: XI).

ci si aspetterebbe da un trattamento automatizzato, ma i tempi non erano certamente maturi per consegnare al calcolatore anche il compito di custodire e rendere fruibile (e modificabile) il prodotto finale, se non altro per il ridottissimo numero di computer disponibili a quel tempo.

La storia dell'importante opera di Busa non si conclude con quei volumi e nel corso degli anni diventerà oggetto di alcune riconversioni digitali: Busa (1995) e Busa (2005). Confrontando gli estremi del percorso si potrebbe concordare con Barbera (2013: 65) quando afferma che «il risultato finale della sua impresa è BUSA 2005[39]». La pubblicazione on-line appare come il compimento ideale dell'impresa, ne restituisce la piena portata, sia per la facilità di accesso, sia per le numerose possibilità di ricerca che il motore permette di utilizzare e personalizzare. Nell'edizione web cambia l'esperienza di consultazione e viene percepito con maggior evidenza il miglioramento sostanziale. A voler parafrasare una battuta che lo stesso Busa utilizzava a proposito della propria età, la rete rivela che la sua impresa (quella del 1951) non è vecchia, ma solo «giovane da più tempo[40]». Nello iato tra le due versioni dell'*Index Thomisticus* diventa lampante come l'uso della macchina «non come sostituto ma come antecedente della stampa è fondamentalmente errato[41]».

1.3.3 Nuovi nodi problematici

Nella misura in cui si fa della tecnologia un puro strumento pratico, un espediente per sollevare lo studioso dal peso

[39] Busa (1992), col passaggio alle memorie ottiche, appare oggi come una tappa intermedia.

[40] Savoca (1994).

[41] Orlandi (1986: 75).

dell'iterazione, dal tedio dei confronti meccanici, lo spazio disciplinare tradizionale rimane immobile.

Quando si parla — e non solo in lessicografia — di impiego puramente strumentale delle tecnologie digitali si interpreta l'applicazione dei nuovi mezzi come se si trattasse di risolvere problemi vecchi con attrezzi nuovi. La validità di questa interpretazione sarebbe quanto meno auspicabile; oltre all'economicità di una visione che lascia intatti tutti i sistemi di riferimento, visti gli ultimi progressi della tecnologia, dovremmo già essere sul punto di risolvere quasi tutti i nostri problemi. Purtroppo, per quanto le lentezze del sistema postale possano dirsi egregiamente superate dall'e-mail, attualmente le sfide tecnologiche sono ormai altre e non si misurano sui parametri del secolo scorso, perché nel momento in cui alcune esigenze vengono risolte, la tecnologia sposta simultaneamente l'asse di riferimento verso nuove esigenze da soddisfare.

In questi termini il cambio di paradigma trova un chiaro segnale nella rivelazione di una nuova problematicità che finisce per condizionare e modificare il quadro epistemologico dominante. Se un impiego strumentale è realmente possibile, può manifestarsi storicamente solo in una prima fase, come ben evidenziato da Mordenti (2001: 25-28), destinata presto a lasciare il posto a una «rivelazione di altri problemi mai intravisti[42]», mai emersi in precedenza, come segnale di una traslazione di coordinate che concorre a manifestare la specificità disciplinare dell'informatica umanistica.

In conclusione, sembra lecito ritenere che il semplice utilizzo di apparecchiature digitali non implica uno scarto di paradigma. Esiste una soglia, o forse una molteplicità di soglie, oltre la quale avviene una variazione qualitativa: «ciò che sembra

[42] Avalle (1985a: 380), cit. in Mordenti (2001: 27).

essere un fatto quantitativo, ha in realtà un lato qualitativo che influisce sulla nostra conoscenza della storia del lessico e della sua situazione nel presente[43]». La difficile individuazione di questa soglia è di fondamentale importanza nella progettazione di un'impresa che miri a definirsi innovativa.

Quanto visto rimane valido anche per il dizionario informatizzato: l'impiego accidentale del calcolatore non basta a trasportarci nel territorio della lessicografia elettronica nelle forme e nelle tecnologie con cui comincia a definirsi in un processo di affermazione ancora in corso.

1.4 FORME DEL DIZIONARIO ELETTRONICO

Sulla base di quanto affermato si può supporre che il semplice coinvolgimento del calcolatore, nella sostanziale invarianza di metodologie operative ed esiti, non è sufficiente per poter parlare di un dizionario elettronico in senso forte e che occorre spingersi oltre per raggiungere un livello di innovazione realmente significativo.

Il pieno sviluppo delle potenzialità del digitale è percepito come una delle responsabilità professionali di chi concepisce sistemi lessicografici informatizzati:

> If new methods of access (breaking the iron grip of the alphabet) and a hypertext approach to the data stored in the dictionary do not result in a product light years away from the printed dictionary, then we are evading the responsibilities of our profession. Now that we know how to

[43] Pascual (2013).

> build better dictionaries, and how to build them faster, we must be able to exploit this happy situation to the full[44].

Se non tutti i vocabolari sono "elettronici" allo stesso modo, ci si dovrà pur chiedere in quale misura lo sono e cosa li rende più o meno innovativi.
La letteratura degli ultimi anni riporta numerosi tentativi di definizione e classificazione del dizionario elettronico, spesso accompagnati da una scala che misura lo scarto rispetto allo *standard* di partenza — il grado zero — rappresentato dai vocabolari cartacei. Spesso l'operazione è compiuta implicitamente attraverso le periodizzazioni storiche, dove è quasi sempre ravvisabile un generale grado di isomorfismo tra l'evoluzione disciplinare e la scala di innovazione.

Va qui introdotta una precisazione preliminare. Oltre alle molteplici, e a volte contrastanti, connotazioni che si ritrovano nella letteratura, quando si parla di dizionario elettronico vanno distinte due accezioni fondamentali: i dizionari-macchina e i dizionari informatizzati[45]. I dizionari-macchina sono costituiti da liste di parole alle quali possono essere associate informazioni linguistiche e sono creati per essere impiegati in software *dictionary-based*. Sebbene mantengano una relazione strettissima tra repertorio lessicale e le applicazioni che accedono alle informazioni da questo contenute, possono essere progettati e realizzati anche indipendentemente da un software di riferimento per essere utilizzati da applicativi diversi. I dizionari informatizzati, invece, sono i comuni dizionari pensati per un lettore umano (*human oriented*), come i classici dizionari in volume. Da ora in poi col termine dizionario elettronico o

[44] Atkins (1992: 521).

[45] Si veda anche Chiari (2007a: 85).

vocabolario elettronico si farà riferimento solamente a quest'ultimo tipo.

Il ricorrente modello di Cerquiglini, riportato da Pruvost (2000: 188) — che ha avuto una buona ricezione, ed è spesso citato e ripreso, ad es. da De Schryver (2003: 143), Fuertes-Olivera / Niño-Amo (2011: 168), Verlinde / Leroyer-Binon (2009: 6) — articola la storia dei dizionari elettronici in tre fasi, caratterizzate da diverse zone di applicazione dei processi computazionali in termini di progetto lessicografico:

1) lessicografia di dizionari cartacei assistita dal calcolatore;

2) digitalizzazione di dizionari cartacei esistenti;

3) dizionari concepiti e realizzati espressamente per il supporto digitale.

Il primo punto dovrebbe essere ormai un dato scontato in qualunque impresa lessicografica del terzo millennio[46]: l'assistenza del calcolatore è ormai una prassi consolidata in qualunque impresa lessicografica (su carta o meno).

Il secondo punto individua le imprese che possiamo definire di "retroconversione", un settore che per quanto possa sembrare poco innovativo, non ha ancora cessato di manifestare la sua importanza e anche alcune soluzioni tecnicamente all'avanguardia[47].

L'ultimo punto nello sviluppo storico, vertice nella scala ideale di innovazione, dovrebbe essere il modello di riferimento per ogni dizionario informatizzato [48] (*VSM* compreso). Un progetto

[46] Cfr. §3.1.

[47] Cfr. il caso esemplare de La Crusca in rete al §2.3.

[48] Per alcuni, ad es. De Shryver (2003:146), la maggior parte dei dizionari elettronici attualmente in circolazione non può essere ricondotta che alla seconda fase.

lessicografico avveduto non dovrebbe allontanarsi da questa zona di informatizzazione radicale, anche solo per evitare vincoli di *medium*: basti considerare l'asimmetria nelle possibili conversioni tra carta e digitale (pressoché immediata quella dal calcolatore al volume, faticosa e imprecisa quella inversa).

Sebbene la turbinosa rapidità delle generazioni tecnologiche pregiudichi rapidamente anche la validità delle classificazioni lessicografiche basate sul supporto[49], conserva un'incontestabile validità la distinzione tra dizionari *off-line* (o basati su reti locali) e dizionari accessibili via internet[50]. Questi ultimi, nella fluidità e velocità radicale del medium, privi di vincoli tipografici ed editoriali, rimangono costantemente aggiornabili ed espandibili in qualunque punto e in qualsiasi momento.

Ulteriori tipologie di dizionari on-line si ritrovano in Chiari (2012: 97), la quale classifica i dizionari elettronici in cinque categorie:

> a) *dizionari informatizzati*, versioni online di autorevoli dizionari cartacei;
> *b*) *dizionari elettronici*, direttamente creati per essere distribuiti esclusivamente online;
> *c*) *dizionari collaborativi*, creati da utenti ordinari in progetti volontari;
> *d*) *strumenti di lessicografia computazionale*, detti anche *dizionari macchina*, sono pensati come database lessicali o basi di conoscenze finalizzate non tanto alla consultazione da parte di utenti, ma all'uso e integrazione in applicazioni computazionali;

[49]Per una rassegna delle classificazioni basate sul supporto (e per un ulteriore tentativo già compromesso dal passare del tempo) si veda De Shryver (2003: 147).

[50] Tanto che alcuni, come Gouws (2011: 17), limitano l'uso del termine 'vocabolario elettronico' solo ai dizionari on-line.

e) *aggregatori di fonti lessicografiche*, come dictionary.com e thefreedictionary.com.

Se i punti *a* e *b* corrispondono, rispettivamente, ai punti 2 e 1 del modello di Cerquiglini e al punto *d* ritroviamo i già citati dizionari macchina, i punti *c* ed *e* elencano due tipi speciali di vocabolari che hanno acquisito una dimensione autonoma con l'affermarsi di internet. Anche se questi due tipi di dizionario non sono rigorosamente pertinenti ai nostri scopi, i principi su cui si basano (la dimensione collaborativa[51] e l'integrazione organica di risorse eterogenee) sono da ricondurre integralmente alla rivoluzione digitale e li ritroveremo applicati, sebbene sotto altre forme, nell'aggregazione dinamica dei lessici del *DMF* e nel *crowdsourcing* dell'*OED*.

Specificamente mirato al grado di innovazione effettiva, Tarp (2012) propone una valida classificazione caratterizzata dalla estrosa metafora dei mezzi di trasporto — «an appealing vocabulary, already widely adopted, for talking about e-dictionaries[52]».

Copycats: nella categoria delle 'fotocopie' rientrano le retroconversioni effettuate non voce per voce ma pagina per pagina e tipicamente pubblicate su file PDF.

Faster horses: il termine fa riferimento a una frase attribuita a Henry Ford[53]. I dizionari in questa categoria fanno uso della tecnologia digitale semplicemente per ridurre i tempi di accesso alle informazioni, anche attraverso l'uso di link o stringhe di

[51] Dove la collaborazione è intesa principalmente come contributo collettivo di non professionisti. Si veda a questo proposito Fuertes-Olivera (2009).

[52] Kilgarriff (2012: 26).

[53] «If we had asked peolpe what they wanted, they would have said faster horses», Tarp (2012: 59).

ricerca. Tarp include nella categoria anche funzioni di lemmatizzazione di forme flesse e varianti ortografiche. A questa categoria, secondo l'autore, va ricondotta la maggior parte dei dizionari elettronici attualmente in circolazione, molti dei quali non sono retroconversioni ma sono stati concepiti direttamente per la pubblicazione in digitale, tuttavia sono basati su modelli tradizionali accettati acriticamente. I dati sono ancora organizzati in articoli statici, modellati pedissequamente sulle entrate dei dizionari a stampa.

Ford modello T: nella categoria intitolata al leggendario modello di automobile che nel 1908 rivoluzionò la fisionomia della produzione industriale moderna, Tarp annovera quei dizionari che riescono a unire la rapidità di accesso alla capacità di adattare dinamicamente gli articoli ai bisogni degli utenti e spesso funzionano attraverso la rete con la possibilità di riadattare i dati (anche in questo caso in funzione dei bisogni degli utenti). La generazione dinamica dei contenuti è ripiegata sulla personalizzazione dei contenuti in base ai bisogni degli utenti, in linea con la *function theory of lexicography*, sviluppata da Henning Bergenholtz e dai suoi collaboratori presso il Centre for Lexicography della Aarhus School of Business dell'University of Aarhus (di cui Tarp fa parte).

Rolls Royce: l'auto di lusso simboleggia qui l'evoluzione del dizionario modello T, con l'eccezione che mentre questo prevedeva la personalizzazione dell'informazione come modalità di accesso a un database precostituito, il dizionario Rolls Royce consente di ri-rappresentare l'informazione e di creare informazioni nuove a partire da quelle già esistenti. Secondo Tarp non sono ancora stati creati dizionari meritevoli di rientrare in questa categoria[54].

[54] Di opinione diversa Verlinde (2012) nello stesso volume.

Questa ricerca proseguirà su una strada diversa da quella tracciata dalla scuola di Aarhus e affiderà eventuali funzioni di adattamento ai bisogni dell'utente a una corretta strutturazione dell'informazione. Pur se improntato alla *function theory of lexicography* la classificazione di Tarp ha comunque una validità generale e condivisibile se gli ultimi due punti vengono considerati per le caratteristiche di dinamicità dell'informazione e per la capacità di generare contenuti nuovi sulla base di informazioni preesistenti.

2 CASI DI STUDIO

2.1 PER UN'ANALISI STORICO-CRITICA

Il tentativo di descrivere il legame tra lessicografia e informatica deve comprendere un gran numero di variabili disperse su un dominio complesso e dimostrare una certa agilità nel districarsi tra le infinite situazioni concrete in cui storicamente la simbiosi si è manifestata. Una via percorribile è tracciata dall'identificazione di fenomeni macroscopici in imprese dalla particolare valenza innovativa. Rintracciarli e quantificarne il peso equivale a sintetizzare un percorso storico-critico in chiave metalessicografica[55] tra alcune tappe scelte, per osservare i modi in cui il settore reagisce agli stimoli delle nuove tecnologie in un momento particolarmente dinamico. Come prodotto secondario, l'analisi potrebbe rintracciare un insieme di tasselli utili per la ricostruzione della storia della nascente lessicografia elettronica, ancora una volta in prospettiva metalessicografica[56].

---

[55] Sulla scorta di Büchi (1996:2), si dovrà distinguere una "métalexicographie syncronique", concentrata sui dettagli del dizionario, una "métalexicographye macrodiachronique", rivolta all'analisi della filiazione dei dizionari, e una "métalexicographye microdiachronique", da condurre sul doppio asse della pianificazione e produzione del dizionario e della ricezione.

[56] Naturalmente la liceità dell'operazione è subordinata ancora una volta al riconoscimento di un'identità qualificante per la lessicografia computazionale, questione ancora da definire ma non di certo in questa sede. Certo è che l'analisi dei fattori di contatto tra lessicografia e informatica può fornire degli elementi a supporto della caratterizzazione delle questioni legate all'ambito di ricerca.

Esiste però un motivo più immediato che spinge a individuare i fattori responsabili della catalisi: individuare le regioni in cui il contatto si è manifestato con spontanea efficacia e osservarle, di volta in volta, nei due diversi momenti della pratica redazionale e della consultazione, rivela quante e quali applicazioni sono suscettibili di apportare immediati e concreti avanzamenti al contesto applicativo per il *VSM*. Per ciascun punto individuato, sarà allora necessario comprendere le dinamiche profonde e, se possibile, enuclearne gli espetti salienti, isolandoli dai fattori accidentali. Attraverso questa strada si avrà un primo censimento dei punti sui quali rivolgere pragmaticamente l'attenzione all'interno del progetto e, in prospettiva, dei momenti da valorizzare, dei limiti da spingere per giungere a nuovi livelli di innovazione propri del dizionario elettronico.

Per i casi in cui il dizionario studiato è stato codificato in XML è stato selezionato e riportato uno *snippet* (un frammento di codice d'esempio). La letteratura specialistica non è prodiga di esempi di codifica dai quali trarre ispirazione per lo sviluppo di un nuovo modello, come per altro già lamentato da Budin / Majewski / Mörth (2012: 4). Nell'analisi caso per caso è stato possibile trovare esempi utili sparsi nella documentazione on-line o nelle opzioni di visualizzazione della voce, con il risultato di raccogliere una piccola collezione di impieghi alternativi di marcatura della voce.

2.2 LIMITI E PARAMETRI DELL'INDAGINE

2.2.1 Tecnologie di rilevo

L'analisi storico-critica illumina, naturalmente, solo una parte dello scenario; rimarranno in ombra tutte le tecnologie che non hanno ancora espresso il loro potenziale. Solo per fornire

alcuni esempi immediati, l'analisi trascurerà tecnologie come la sintesi vocalica o le tecniche di *data mining* e gli algoritmi di intelligenza artificiale. Il fatto che non abbiano beneficiato di applicazioni costanti o significative in imprese pertinenti ai nostri fini non le esclude dalla definizione del dizionario elettronico e non ne svaluta il potenziale nel medio e nel lungo periodo. Il loro eventuale ruolo andrà però riconosciuto in momenti diversi, attraverso lo studio e l'applicazione sperimentale in contesti concreti.

Per converso, non tutte le tecnologie rilevate andranno annoverate di diritto tra i fondamenti per la costituzione del dizionario elettronico. In alcuni casi si ha la chiara percezione che, seppure indispensabili nella realizzazione del dizionario elettronico e senza dubbio decisive per l'esplosione quantitativa, certe tecniche (senza volerne sminuirne l'importanza) siano più vicine all'ideale cassetta degli attrezzi che alle fondamenta epistemologiche della disciplina [57] e una volta compiuto il compito per le quali sono progettate spariscono dietro il risultato finale. Si considerino, ad esempio, la *computer vision* e il riconoscimento delle immagini, che hanno portato alla realizzazione dei diffusissimi programmi per il riconoscimento ottico dei caratteri (OCR), fondamentali in tutte le imprese lessicografiche che prevedono una fase di retro-digitalizzazione [58]. Grazie anche a strumenti di revisione assai

[57] Sono «aspetti puramente tecnici» ad es. per Gandin (2005: 145).

[58] Le immagini acquisite al computer, comprese quelle provenienti dalla digitalizzazione di volumi a stampa, sono costituite da una matrice di punti (i pixel) a ciascuno dei quali sono associate le informazioni relative ai colori (nel caso più comune, rosso, verde e blu). Per convertire le pagine digitalizzate in testo elettronico, manipolabile con i programmi di videoscrittura o per l'*information retrieval*, occorre isolare tramite algoritmi le matrici di punti che raffigurano i diversi caratteri che compongono l'immagine e associarle a un carattere di uno schema di codifica di riferimento. Allo stato attuale i software che svolgono questo compito hanno una percentuale di affidabilità che si aggira

potenti, questi software riescono ad abbattere drasticamente i tempi di digitalizzazione in formato testo e sono ormai diventati indispensabili sia per digitalizzare e rendere interrogabili vecchi dizionari a stampa, sia per costituire i corpora sui quali vengono costruiti nuovi dizionari elettronici[59]. In merito all'ultimo punto è da citare il progetto per il *Nuovo Vocabolario dell'Italiano Moderno e Contemporaneo*, per il quale è in corso di allestimento il corpus di riferimento[60] con il coordinamento dell'Accademia della Crusca. Nel corso dell'ultimo anno l'unità di Catania [61] ha digitalizzato e convertito in formato testo parecchie migliaia di pagine per i generi testuali della letteratura per l'infanzia, testi scientifici e galatei. L'esperienza diretta in questo progetto ha dimostrato l'impossibilità di ottenere risultati simili con tecniche diverse e nulla potrebbe far dubitare dell'efficacia della tecnologia,

---

tra il 90 e il 95% (a seconda del programma e del testo da analizzare). Il sistema OCR spesso prevede la possibilità di essere "addestrato" sul testo specifico da riconoscere. Dopo un riconoscimento iniziale, il programma presenta il risultato a un operatore umano che corregge gli eventuali errori; da queste correzioni il programma "impara" e migliora il rendimento nel riconoscimento delle pagine seguenti, arrivando a sfiorare percentuali di errori di ordine inferiore all'1%.

Si veda anche Ciotti (1995: 159-160).

[59] Per Garrett et al. (2015: 51) vale anche l'inverso, con un particolare legame tra tecnologie e comunità linguistica: «Outside of dictionary compilation, part-of-speech tagged corpora are a sine qua non for many language related technologies such as automatic translation, speech recognition, auto-completion, optical character recognition, etc. If these technologies are ever to be available to Tibetan speakers then more must be done to create part- of-speech tagged Tibetan corpora».

[60] Corpus di riferimento per un Nuovo Vocabolario dell'Italiano moderno e contemporaneo. Fonti documentarie, retrodatazioni, PRIN 2012.

[61] Per la quale ho prestato servizi di consulenza e pianificato un flusso di lavoro per la digitalizzazione (per mezzo di scanner piani e planetari), riconoscimento ottico dei caratteri, revisione manuale, controllo qualità e codifica in formato XML/TEI.

ma i materiali finora prodotti e il metodo di riconoscimento in sé non ci dicono nulla sulla natura e sul futuro del vocabolario o sul tipo di utilizzo che questo farà del corpus e queste tecniche, una volta compiuto il loro compito, tendono a sparire dietro al risultato finale.

Analoghe considerazioni andranno fatte sui supporti di memorizzazione ottici o magnetici, rivoluzionari per aver liberato i dizionari dalla massa consistente dei volumi a stampa ma ora superati (nelle performance e nel potenziale lessicografico) dalla memorizzazione on-line, nel *cloud*, o dalle più recenti memorie flash. Sarà da considerare la dematerializzazione in generale[62] ma non la singola tecnologia.

Non è facile riconoscere e descrivere ampie trasformazioni dall'interno, nel preciso momento storico in cui ci troviamo e immersi nell'incessante avvicendarsi degli strumenti tecnologici. La storia ci insegna inoltre che le nuove tecnologie e i relativi *standard* non sono mai stabili all'inizio[63] e, da questo punto di vista, la stretta finestra temporale attraverso cui possiamo guardare non offre all'interpretazione un panorama ampio e stabile. Per questi motivi, oltre alla necessaria prudenza, è raccomandabile una sensibilità che (per quanto possibile) si orienti verso gli elementi più profondi e di largo impatto e tralasci i dettagli minuti e una disponibilità ad accettare alti margini di incertezza.

2.2.2 Definizione del campo di indagine

Il rigore nella selezione dei casi di studio è minato alle fondamenta dai confini sfumati del panorama di nostro interesse. Il

[62] Di cui si tratterà in relazione al sistema informativo (§3.1).

[63] Un esempio su tutti, pertinente con la testualità digitale, è la guerra di formati per gli *e-book* e la mancanza di uno *standard* affermato. Per una rassegna si veda Sechi (2010: 112-172).

criterio di inclusione che sembrerebbe più scontato, cioè selezionare i dizionari storici del dominio romanzo con una buona versione elettronica, non può funzionare per svariati motivi.

A voler perseguire il rigore sulla prima condizione (essere un dizionario storico) si finirebbe per omettere il *Vocabolario degli Accademici della Crusca*, che tradizionalmente apre le trattazioni sulla lessicografia diacronica italiana ma che dizionario storico, a rigore di termini, non lo era in origine e non lo è neppure adesso. Con buona pace del dovuto ossequio, nulla impedirebbe di rifiutare il consueto tributo al nostro monumento lessicografico, ma la nostra indagine perderebbe l'occasione di riportare un eccellente esempio di retroconversione, col suo portato di soluzioni tecniche e che dimostra come la digitalizzazione possa trarre nuove informazioni anche da un insieme di dati già ampiamente scandagliato.

Il secondo requisito (essere o avere una buona edizione elettronica) nella logica formale verrebbe definito come una condizione ambigua, per la quale cioè non è possibile stabilire una condizione di verità se non a patto di definire formalmente l'aggettivo "buona"[64]; se poi consideriamo che lo scopo dello studio è di descrivere le proprietà di un buon dizionario elettronico, si ricadrebbe nella trappola della circolarità. La bontà dell'edizione elettronica — che ha già l'effetto di tagliare fuori tutte le iniziative informatizzate solo nella redazione ma non nel prodotto finale — verrà valutata elasticamente secondo due criteri: il miglioramento sostanziale rispetto all'edizione a stampa (reale o ideale nel caso di dizionari *born-digital*); adeguatezza agli *standard* tecnici che definiscono lo stato dell'arte attuale. Quest'ultimo criterio esclude,

[64] In termini rigorosi, l'aggettivo "buona" non identifica una proprietà (cfr. Devlin 1997: 249).

tra gli altri[65], le edizioni su CD-ROM[66] del *Godefroy*[67] o del *Tobler-Lommatzsch*[68].

Alcuni importanti esempi di dizionari etimologici o storico-etimologici (come il *FEW* e il *DEAF*) indurrebbero ad allargare l'indagine all'intera lessicografia diacronica. Accettato un simile allargamento, si toglierebbe dignità di studio a quella zona di irriducibile e indistricabile sovrapposizione tra lessicografia diacronica e sincronica[69]; a farne le spese sarebbero alcuni considerevoli esempi di lessici sincronici assolutamente pertinenti con i nostri interessi come nel caso del *DMF*[70] e dell'*AND*[71].

Se questo non fosse sufficiente a escludere l'ipotesi, anche limitando rigidamente alla lessicografia sincronica l'ambito di

---

[65] Si tratta precisamente di dizionari sincronici (cfr. Trotter 2013) ma del dovuto allargamento del campo di indagine si tratterà poco più avanti.

[66] Lo stato dell'arte attuale è dominato dal dizionario on-line (cfr. §3.5).

[67] Blum (2002), per il quale si vedano anche le recensioni di Matsumura (2003a e 2003b). Interessante, ma distante dai nostri fini immediati, il progetto *Bibliographie Godefroy* sviluppato dall'ATILF (<http://www.atilf.fr/BbgGdf/>).

[68] Blummenthal / Stein (2002). La versione digitalizzata in formato immagine disponibile con un trattamento OCR grezzo disponibile sul sito, <http://www.uni-stuttgart.de/lingrom/stein/tl/>, non è da prendere in considerazione. La versione on-line alla data attuale non funziona correttamente (viene visualizzato il lemmario ma le voci restituiscono un errore 404 di pagina non trovata). Il sistema impiegato per la realizzazione del dizionario è OpenCms, un gestore di contenuti (Content Management System) di uso generico e non specialistico (<http://www.opencms.org/>).

[69] «A clear distinction between synchronic lexicography, and the presentation of historical and etymological data is rarely achievable. […] Dictionaries of past periods of the language, even if they themselves present a synchronic overview, necessarily yeld information which is inherently historical and often of use in etymological dictionaries» (Trotter 2013: 663).

[70] Cfr. Martin (1998: 967).

[71] Ivi, p. 664.

riferimento, il risultato sarebbe privo di rappresentatività perché privo delle sue opere più importanti e paradigmatiche:

> Il *FEW*[72], il *LEI*[73], il *DI* o anche il *DEAF* ed il *DEM*, cioè i dizionari che costituiscono i riferimenti metodologici della scienza attuale, non includono l'informatica nella loro redazione. Sfruttano le banche dati esistenti per migliorare la documentazione antica, ma la loro concezione resta indipendente dai mezzi informatici[74].

Ancora a proposito dei riferimenti metodologici, limitare la selezione al dominio romanzo escluderebbe l'*OED*. Come si potrebbe poi bilanciare la ricchezza di iniziative dell'area gallo-romanza[75] con la povertà dell'area romanza orientale?

Per sfuggire a questa sorta di problemi ci si affiderà a una 'logica *fuzzy*', a una selezione più elastica di pochi e selezionati casi esemplari, restringendo le opere esaminate ai casi che soddisfano, di volta in volta, la pertinenza con l'obiettivo della ricerca e la qualità della realizzazione informatica. Avrà minor peso nella scelta, ma inciderà nell'ordinamento, il perseguimento di una trattazione organica lungo un percorso che dalla retroconversione conduce ai dizionari *born-digital*, fino al vertice ideale del *TLIO* (modello programmatico per il *VSM*).

---

[72] Il dato sarà presto da aggiornare grazie alla retroconversione in cantiere all'ATILF: Pagina del progetto: <http://stella.atilf.fr/gsouvay/scripts/mep.exe?CRITERE=eFEW;ISIS=mep_few.txt;OUVRIR_MENU=5;s=s12523ad0;ISIS=mep_few.txt>. Ultima consultazione: 7 novembre 2015.

[73] Il *LEI* è disponibile solo su PDF all'indirizzo <http://www.uni-saarland.de/lehrstuhl/schweickard/lei/pubblicazioni.html>. Ultima consultazione: 15 ottobre 2015.

[74] Gleßgen (2006: 17).

[75] Trotter (2013: 663).

2.3 IL *VOCABOLARIO DEGLI ACCADEMICI DELLA CRUSCA*

Cominciare con il *Vocabolario degli Accademici della Crusca*, nelle sue cinque impressioni[76], è un modo per approdare nel territorio del digitale attraverso un ponte ideale tra antico e moderno con un buon esempio di digitalizzazione e annotazione di un vocabolario tra i più studiati. L'informatizzazione del *Vocabolario degli Accademici della Crusca* vale qui come esempio eccellente di versione elettronica di un dizionario cartaceo progettato e definitivamente completato sulla carta; studiare le specificità dell'impresa non ne riduce il valore paradigmatico ma, anzi, svela alcuni aspetti di portata generale utili per meglio descrivere lo scarto tra edizione a stampa ed edizione elettronica.

Come già affermato, il dizionario degli *Accademici della Crusca* non era di certo storico nelle intenzioni degli accademici. Se oggi è accostato ai dizionari diacronici — tra i quali viene annoverato come «primo autorevole esempio non solo per

[76] 1612, 1623, 1691, 1724-1738 e 1863-1923; quest'ultima incompleta per decreto ministeriale di Giovanni Gentile (per una ricostruzione con valore di documento storico cfr. Barbi 1935).

l'italiano, ma per tutta l'Europa[77]» o «diretto capostipite dei dizionari storici[78]» — è grazie al gran numero di citazioni con cui arricchisce le sue 30.000 voci, tratte da un canone di autori selezionato con lo scopo di definire un modello di lingua. Ma dedurre la norma dall'uso degli scrittori, a rigor di termini, non dovrebbe essere una condizione sufficiente per parlare di dizionario storico, perché il *Vocabolario* degli accademici contiene citazioni «non degli scrittori in genere, sì dei buoni scrittori; e quel buoni basta ad escludere il fine storico[79]». Gli autori non "purgati" non trovano posto nel fondo dei citati, e con questi sono esclusi tutti i forestierismi, i neologismi e le innovazioni contenute nei loro testi; degli autori approvati, invece, sono state scartate tutte le voci che «a parere degli accademici erano morte e non potevano sperare in una risurrezione[80]». Rimane però incontestabile la linea di indagine empirica[81], singolarmente vicina[82] alla linguistica *corpus-based*[83] —

---

[77] Marello (1996: 87). Il Vocabolario, come noto ebbe impatto sovrannazionale; gli accademici, di fatto, «realizzando la prima edizione del loro *vocabolario*, misero gli altri popoli del continente sulla via per giungere a una concezione più definita e a un migliore equipaggiamento delle proprie lingue» (Sabatini 2014: 16). Per una recentissima panoramica sulla fortuna del *Vocabolario degli Accademici della Crusca* in Europa si veda Maraschio / Poggi Salani (2014: 60-66).

[78] Serianni (1999: 8), cit. in Beltrami (2009b: 3) a cui si rimanda per un approfondimento del concetto attraverso la tradizione lessicografica italiana.

[79] Pasquali (1941: 46).

[80] Ivi, p. 47.

[81] Come risaputo, l'attenzione all'attestazione, inaugurata dal Vocabolario della Crusca, è un carattere che si trasmetterà alla tradizione lessicografica italiana, pur nel mutamento funzionale dell'esempio, che in molti casi si allontanerà progressivamente dalla citazione-autorità della Crusca, per la quale cfr. Beltrami (2009b: 3).

[82] La posizione è di Sabatini (2006) e Sabatini (2007) ma è largamente condivisibile (si veda ad es. Barbera (2013).

che, come si vedrà nel §3.6, è estremamente vicina ai paradigmi attuali:

> è evidentissima […] l'intenzione di pervenire alla formazione di un vero corpus universale, capace di rappresentare tutta la lingua […] L'idea di un canone ristretto di autori modello da imitare per produrre altri testi modello si era trasformata nel concetto di un vero corpus integrato e bilanciato, dal quale si cercava di estrarre tutta la lingua[84].

Negli ultimi anni l'Accademia ha dimostrato un'attenzione sempre crescente verso le nuove tecnologie e per l'attività on-line[85]. Tra queste il progetto *Lessicografia della Crusca in rete*, finanziato dalla Presidenza del Consiglio dei Ministri[86] — avviato nel 2001 e concluso nel 2006 — grazie al quale sono state informatizzate e pubblicate on-line (ad accesso gratuito) le cinque edizioni del vocabolario[87]. La *Lessicografia*, «insieme una biblioteca digitale e un dizionario elettronico[88]», rende disponibili per la consultazione i 22 volumi[89] delle 5 impressioni nella loro interezza[90].

---

[83] Di cui si avrà modo di parlare più in dettaglio al §3.6.

[84] Sabatini (2006: 34-35).

[85] Per una descrizione del nuovo portale dell'Accademia cfr. Biffi (2013).

[86] L'idea di schedare elettronicamente il *Vocabolario* maturò nel corso degli anni '70. Nel 2001 fu resa disponibile la prima impressione on-line sul sito web della Scuola Normale di Pisa. Per una storia del progetto si veda Fanfani (2014: 79-81).

[87] La pagina di accesso è disponibile sul portale dell'accademia, all'indirizzo <www.accademiadellacrusca.it> (ultima consultazione: 2 novembre 2015) o all'indirizzo <www.lessicografia.it> (attualmente «in fase di allestimento»; ultima consultazione 2 novembre 2015).

[88] Biffi (2014: 115).

[89] 1 volume per la prima impressione, 1 volume per la seconda, 3 per la terza, 6 per la quarta e 11 per la quinta.

Alle origini dell'iniziativa troviamo l'urgenza del «superamento della consultazione tradizionale del *Vocabolario* degli accademici per una "lettura" complessa e attenta del medesimo[91]», avanzata nel corso del *I Convegno Nazionale del Lessico Tecnico e delle Arti e dei Mestieri (Cortona, 28-30 maggio 1979)*, troppo presto quindi per poter attribuire al concetto di digitalizzazione l'autosufficienza e l'autoreferenzialità a cui siamo oggi abituati[92]. Il punto di partenza erano le concrete caratteristiche lessicografiche peculiari del *Vocabolario* e un obiettivo preciso da raggiungere: far emergere i termini dell'uso e tecnico-scientifici da una struttura lessicografica che non li restituisce esplicitamente. In altre parole, all'interno del *Vocabolario* esiste un «lessico nascosto[93]» composto da:

---

[90] Le prime 4 edizioni sono state digitalizzate, trascritte e annotate; la quinta è attualmente in corso di marcatura (come nuovo progetto inaugurato nel 2012).

[91] Sessa (1982: 272).

[92] Oggi una digitalizzazione si giustifica di per sé, sulla base di numerosi valori consolidati e implicitamente riconosciuti come l'accessibilità, la preservazione, la divulgazione. Se mai si potesse parlare per il 1979 di una connotazione definita per il concetto di "digitalizzazione", è chiaro che sarebbe molto diversa da quella di oggi. La reti e i supporti di memorizzazione non erano ancora in grado di distribuire i prodotti della ricerca come oggi e il "computer" non era ancora diventato "personal" (un problema sostanzialmente di disponibilità del mezzo informatico, come per Gleßgen 2006: 15). Alcuni studiosi, come Trotter (2011: §1), sono spinti da queste premesse a postdatare «the advent of computing in the Humanities from 1980s onward. I appreciate that Humanities computing started earlier than the date suggested here, but for the purpose of this project [l'Anglo-Norman Dictionary] (and for the vast majority of researchers in the UK at least), the 1980s saw the emergence of affordable personal computers capable of significantly changing working methods». Se il *Times* ci fornisce una data simbolica con la consacrazione del *personal computer* a personaggio dell'anno nel 1982, rimane tuttavia da stabilire se in Italia la diffusione effettiva del PC non sia avvenuta più tardi.

[93] Sessa (1982: 274).

1. sottolemmi senza rimando
2. voci di definizioni, locuzioni e proverbi
3. significati del lemma registrati altrove
4. definizioni non evidenziate nell'articolo del lemma

Tentando di reinterpretare in chiave formale: la macrostruttura a livello di lemmario è "cieca" nei confronti di certi elementi (quello che si verifica ad es. nei punti 1, 3; alcuni sottolemmi non hanno una segnalazione nel lemmario e spesso sono inclusi in voci con le quali non intrattengono alcun legame semantico o etimologico); l'informazione è spesso dispersa e l'accesso a certi tipi di dati non è diretto; si aggiunga il ruolo della citazione, specialmente in voci prive di definizione e una segmentazione della microstruttura non sempre limpida a causa di una segnalazione tipografica degli elementi poco precisa (per es. la caratterizzazione non uniforme dei sottolemmi).

Per accedere alla piena ricchezza del *Vocabolario* è stato quindi formulato il concetto di *rovesciamento*, un rimaneggiamento sistematico dei contenuti del *Vocabolario* per far emergere e rendere ricercabile tutto il lessico in esso contenuto e non solo quello indicizzato nel lemmario[94], «uno "smembramento" dei lemmi nelle loro diverse componenti, da marcare elettronicamente per potervi poi compiere indagini mirate [95] ». Il concetto di "rovesciamento" — rimasto un termine di riferimento nell'impresa di digitalizzazione del *Vocabolario della Crusca* — in generale, è un'azione propedeutica per il passaggio a un vocabolario elettronico in senso forte[96], la cui applicazione basilare è nella

[94] Ibidem.

[95] Fanfani (2014: 79).

[96] La strutturazione dell'informazione e il "rovesciamento" della voce classica aprirà la trattazione sulle proprietà costitutive del dizionario elettronico al §3.1. Non si adotterrà il termine in quanto troppo legato alla singola impresa.

riconversione in digitale di dizionari tradizionali ma i cui principi giacciono a fondamento anche nei dizionari *born digital* (che, se fatti bene, dovrebbero nascere già "rovesciati").

Si aggiungerà che una lettura "rovesciata" è agevolata dalla semplice ricerca *full text* di qualunque edizione in formato testo, ma che per consentire le indagini mirate a cui si è fatto riferimento occorre una marcatura strutturale. A questo proposito, il rovesciamento de *La lessicografia della Crusca in rete* è stata effettuato con una marcatura in XML/TEI[97], basata su una selezione di 32 marcatori. Gli strumenti di interrogazione e gli schemi di marcatura sono stati calibrati, di volta in volta, su ciascuna delle impressioni. La cura nei confronti dell'interpretazione dell'oggetto codificato ha sollecitato un adattamento minuto delle metodologie informatiche sulla specificità testuale del *Vocabolario*:

> Anche la *tokenizzazione* del testo — l'indicizzazione per l'individuazione delle forme — è stata gestita con accortezza diacronica in relazione alla gestione degli apostrofi e degli asterischi; e la punteggiatura, opportunamente indicizzata, può essere ricercata al pari delle forme in tutti i tipi di interrogazione, in modo da allargare il campo di indagine linguistica anche ad alcuni aspetti sintattici[98].

La ricerca avanzata permette di selezionare il modo di presentazione della voce tra due tipi di trasformazione XLS, una modalità in testo non formattato e XML[99]. Il sorgente è quindi facilmente visualizzabile e analizzabile:

---

[97] Cfr. §4.2.

[98] Fanfani / Biffi (2006: 414).

[99] Si noti che come opzione definitiva il form di ricerca avanzata imposta l'evidenziazione dei contesti che si traduce nell'immissione all'interno dell'XML del *tag* <EVIDENZCONTESTO>.

```
<entry>

<form>
      <orth>ACQUISTARE</orth>
</form>


<sense>

     <def>
          Venire in posession di quel che si cerca, o
          che giustamente conviene all'opere, che si
          fanno.
          Lat$. <foreign lang="latino">acquirere,
          consequi, adipisci</foreign>.
     </def>

     <eg>
           <quote>
                 <abbr>
                       <ref>
                       <linkABBR
                       href="reflist.jsp?search_mode=ref&
                       amp;1=checkbox&amp;psearch=Bocc.">
                             Bocc$. </linkABBR>
                       </ref>
                       Introd$. n$. 10.
                 </abbr>
                 E così faccendo si credevano ciascuno a
                 se medesimo salute acquistare.
           </quote>
     </eg>

     <eg>
           <quote>
                 <!-- … -->
           </quote>
     </eg>

</sense>

</entry>
```

I dettagli dei *tag* generici (<entry>, <sense>) verranno discussi nel cap. 6.3. Tra le specificità del caso è possibile notare:

- Le «corrispondenze "paretimologiche", o, per meglio dire, semantiche, latine e greche, che hanno la funzione, come gli esempi, di limitare e scandire i campi semantici[100]», sono marcate con il *tag* *foreign* e i relativi valori dell'attributo "lang".
- L'imprecisione nelle citazioni, endemica nella lessicografia storica, viene risolta scomponendo la referenza in due parti: una parte fissa, con riferimento ad autore e opera e una parte variabile, per indicare il punto dell'opera citato. La marcatura della parte fissa ha lo scopo di uniformare e di mettere in corrispondenza univoca tutte le citazioni a una data opera.

Non per ultimo, il vocabolario permette di passare dinamicamente dalla lettura in formato testo a quella in formato immagine, a garanzia della fedeltà all'originale e come dovuta edizione fotografica dell'originale antico (funzionalità auspicabile in lavori del genere che incide positivamente sulla valutazione dell'operazione[101]).

La qualità dell'immagine, per la verità, non è superlativa: le pagine sono affidate a file poco definiti, in scala di grigi e compressi in formato GIF (non di certo il più adatto per preservare i dettagli dei caratteri).

---

[100] Sessa (1982: 271).

[101] Si consideri, per es., quanto avviene nelle valutazioni parallele dei CD-ROM del Godefroy di (Blum 2002) e del Tobler-Lommatzsch (Blumenthal / Stein 2002) dove viene spesso notata l'assenza delle immagini a differenza del secondo. Si vedano per es. Trotter (2013: 664), Matsumura (2003a e 2003b).

Il ruolo che il *Vocabolario degli Accademici della Crusca* ha ricoperto nella storia della lingua italiana conferisce a questo caso particolare di digitalizzazione uno speciale rilievo "autoreferenziale": l'edizione elettronica permette di conoscere sistematicamente e quantificare più in profondità il *Vocabolario* stesso e il lessico che contiene, restituendone un'immagine quanto mai dettagliata, «una nuova lettura, destrutturata e capillare[102]» di uno dei principali protagonisti della nostra storia culturale e linguistica[103].

2.4 L'*OXFORD ENGLISH DICTIONARY* (OED)

L'*Oxford English Dictionary*, indiscusso monumento lessicografico per la lingua inglese [104], è uno dei principali riferimenti per lessicografia storica. Nencioni (1955), nella prima relazione all'Accademia della Crusca per il progetto di un vocabolario storico per la nostra lingua, ne sottolineò «il rigore del metodo lessicografico» e «l'integralità della registrazione» e ne fece un termine di confronto assieme al *Thesaurus Linguae Latinae*; o, ancora sulla stessa linea ma molti anni dopo, Beltrami (2013) sintetizza l'ideale metodologico della lessicografia storica parafrasandone il titolo della prima versione in «an Italian dictionary on historical principles».

Se queste premesse puramente lessicografiche non dovessero bastare, l'*OED* è una pietra miliare della lessicografia elettronica,

[102] Fanfani (2014: 79).

[103] Cfr. Sabatini (2006: 35-36) e Fanfani (2014: 75-81).

[104] L'*OED* oggi costodisce la storia e i significati di oltre 600.000 parole. Una rapida introduzione alle novità apportate dal vocabolario e alla ricezione della prima edizione in Mugglestone (2000).

nel glorioso passato come nella sperimentazione attuale. Potrebbe bastare a renderlo tale la colossale opera di digitalizzazione, avviata negli anni '80[105], o per il traffico che genera oggi il sito che lo ospita o per l'approccio collaborativo in rete. Ma a rendere l'impresa un caso singolare è il contributo che diede negli anni '80 allo sviluppo dei linguaggi di marcatura; uno dei rari punti di contatto tra lessicografia e informatica in cui la trasformazione non avvenne in maniera unidirezionale e l'influenza tra le due discipline fu reciproca.

Dopo la conclusione della redazione del supplemento moderno, nel 1986, la Oxford University Press si interrogò sul futuro del dizionario e decise di trasferirne l'intero contenuto su supporto digitale. L'operazione fu realizzata digitando manualmente i 67 milioni[106] di caratteri che componevano l'opera, impegnando 120 dattilografi per oltre 18 mesi. In quella prima battitura fu applicata una prima marcatura per consentire future funzioni di ricerca e aumentare le possibilità di trattamento del testo. Furono segnalati, per esempio i tipi di carattere e proprietà strutturali come l'inizio dell'entrata. In quegli anni l'*OED* non si limitò a trasferire le proprie informazioni su supporto digitale ma diede un impulso allo sviluppo dei linguaggi di marcatura attraverso una partnership con il progetto per la realizzazione dello *standard* Generalized Mark-up Language (SGML), mettendo a disposizione il database del dizionario.

La prima edizione su supporto ottico fu un prototipo del 1989, la prima versione commerciale fu lanciata sul mercato nel

---

[105] Cfr. Lee Berg / Gonne / Tompa (1990-1991).

[106] <http://public.oed.com/about/the-oed-and-innovation/>, ultima consultazione 19 ottobre 2015. Lee Berg / Gonne / Tompa (1990-1991: 32) riportano numeri più alti: oltre 500 milioni di caratteri.

1992. L'impresa cominciò quindi a intraprendere lo spoglio elettronico dei testi, attraverso una collaborazione con l'Oxford Text Archive.

Già nel 1994 il progetto si indirizzò verso la pubblicazione on-line. Nel marzo del 2000 l'*OED* divenne il «first major national, historical dictionary to present itself on the Internet[107]».

Le voci sono codificate in XML e vengono periodicamente aggiornate ogni tre mesi. Non è stato possibile trovare sorgenti di esempio[108], ma le caratteristiche fondamentali di marcatura sono conformi ai criteri generali delle risorse lessicografiche della Oxford University Press, «ensuring the best possible accessibility for computational analysis[109]».

Dai criteri elencati dall'editore si ricavano una serie di linee guida che meritano di essere qui riportate sinteticamente, almeno nei punti più importanti, e che andranno tenute in conto nella rappresentazione digitale della microstruttura che si affronterà nel capitolo 6.3: tutte le risorse devono essere marcate completamente (*full tagging*) per consentire l'estrazione di unità informative specifiche; i *record* contenenti i sensi per la disambiguazione vanno rigorosamente separati; è raccomandata la raccolta di un numero consistente di citazioni; è raccomandata l'inclusione di informazioni sintattiche; è previsto l'uso di marcatori lessicali per la creazione di tassonomie; ogni risorsa presuppone una DTD per la

---

[107] *About the OED and Innovation*: <http://public.oed.com/about/the-oed-and-innovation/>, ultima consultazione 19 ottobre 2015.

[108] È stata effettuata una richiesta per un campione di esempio tramite il servizio di supporto on-line. L'*e-mail* di conferma della comunicazione, che assicurava una risposta «as soon as possible, and normally within 10 days» non è stata seguita da ulteriori messaggi.

[109] <https://global.oup.com/academic/rights/digital-reference/?cc=it&lang=en&>, ultima consultazione: 26 novembre 2014.

standardizzazione del formato di rappresentazione dei dati e l'integrazione di risorse diverse.

Nel caso specifico dell'*Oxford English Dictionary* gli elementi predisposti al trattamento automatizzato sono:

- Full listing of all possible syntactic forms, tagged to show relationships to headword.
- Encoding of morphological behaviour at individual sense level.
- IPA pronunciation for every form.
- Full morphological data for spelling variants, plus straightforward links to *standard* forms[110].
- Flexible codification of over 10,000 phrasal verbs and other multi-word units, allowing easy identification of real-world variations.
- Classification of over 80,000 words/senses under 200 subject domains.
- Semantic relationships between nouns and senses codified under WordNet-compatible taxonomy.

Meritano un approfondimento le tecniche di classificazione degli ultimi punti. Quando nel 2000 l'*OED* on-line è stato lanciato come complemento digitale all'*OED*, tra gli elementi tecnici e multimediali che lo distinguevano dalla controparte in volume, apparve un sistema di classificazione basato su *tag*: le parole sono raggruppate attorno a etichette che le classificano per soggetto, uso, regione, origini. Ciascun raggruppamento costituisce una "categoria".

[110] Le tecniche per implementare questa caratteristica torneranno più volte nel corso dei prossimi capitoli.

Il vocabolario è inoltre in grado di generare dinamicamente una rappresentazione grafica della prima attestazione di parole, le *timelines*[111]. La funzionalità si può utilizzare in combinazione con le categorie per mostrare, per esempio, la data di ingresso di tutte le parole italiane nel lessico della lingua inglese[112]. La capacità del sistema di renderizzare graficamente incroci complessi di dati ricavati dal vocabolario è una conseguenza del sistema di marcatura con cui sono state strutturate le voci; la corretta rappresentazione di ogni informazione in XML organizza coerentemente l'informazione in una struttura esplicita su cui gli algoritmi di estrazione delle informazioni possono indirizzare le chiamate[113].

Il vocabolario permette di navigare tra le fonti più citate (dal *Times* a Shakespeare) e di ordinare le fonti per citazioni totali, citazioni per prima attestazione e citazioni per significati. È anche possibile estrarre una classificazione tassonomica dei lemmi e dei significati, creando un indice semantico dei contenuti del vocabolario, il cosiddetto *Historical Thesaurus*. L'accesso ai dati può essere effettuato anche attraverso un web-service[114].

L'*OED* è anche un dizionario fortemente informatizzato anche nel lavoro di redazione. Contestualmente al progetto di digitalizzazione degli anni '80 fu sviluppato un software per la gestione dei dati che nel tempo fu riadattato come sistema di edizione per la revisione del 1993; il software accompagnò il vocabolario fino alla pubblicazione on-line del 2000. Intorno al 2003 lo sviluppo di un nuovo sistema di editing divenne

---

[111] *Guide to Timelines*: <http://public.oed.com/how-to-use-the-oed/guide-to-timelines/>. Ultima consultazione: 22 ottobre 2015.

[112] Cfr. (Simpson 2013).

[113] Elliott / Williams (2006: 258).

[114] <http://public.oed.com/subscriber-services/sru-service/>. Ultima consultazione: 22 ottobre 2015.

un'urgenza, anche per una serie di invalicabili incompatibilità con l'hardware più recente che nel frattempo si erano manifestate. Tra le caratteristiche richieste al nuovo sistema, un posto spettava anche alla pubblicazione di aggiornamenti regolari distribuiti via web. Il nuovo sistema fu battezzato Pasadena e lo sviluppo fu affidato alla *software house* francese IDM[115], oggi con filiali anche negli Stati Uniti e in Cina.
Il sistema oggi si occupa della riduzione degli errori, della standardizzazione delle referenze bibliografiche, e della gestione delle citazioni.

Dai criteri di progettazione di Pasadena è possibile rilevare alcuni elementi necessari alla progettazione di un ambiente di *editing* lessicografico:

1. L'automazione di tutti i processi meccanici del lavoro lessicografico.

2. La flessibilità di adattarsi a eventuali modifiche, come richiesto da un progetto editoriale a lungo termine.

3. L'integrazione di tutte le attività coinvolte nella redazione del dizionario[116].

---

[115] <http://www.idmgroup.com>. Ultima consultazione: 22 ottobre 2015.

[116] «To describe the problem: for good reasons at the time, the 1993 editing system had been restricted to editing dictionary entries one by one, and to searching, separately, work-in-progress and the various kinds of ancillary electronic material collected over by *OED* over fifteen years. The mark-up underlying the *OED*'s electronic text still reflected the philosophy of the original digitisation project, designed to retain every feature of the original print publication and in an idiosyncratic mark-up style. Quotations collected electronically by the OED's reading programmes were kept in separate databases with separate editing systems. Administrative workflow systems had been established for various types of work chasing and monitoring, but these were operating separately from the main editorial system. Automated validation of the text was limited. The text was full of comments, administrative and editorial, which were useful in situ, but needed to be searched round or stripped

In tempi recenti il vocabolario si è aperto alla dimensione collaborativa di internet, il cosiddetto *crowdsourcing*[117], un nuovo abito 2.0 sotto il quale si ritrova una vecchia pratica che può essere fatta risalire al celebre appello di Murray (1879).

La ricchezza di soluzioni innovative proposte dall'*Oxford English Dictionary* ne conferma l'immagine di colosso lessicografico di primo rilievo in tutto il panorama mondiale. L'*OED* ha alle spalle una lunga e prestigiosa storia ma ha saputo sfruttare la sua posizione consolidata per mantenere alti livelli qualitativi anche nell'innovazione. La tradizione importante non ha mai frenato la disponibilità a valutare soluzioni nuove, anche quando si tratta di intervenire su temi delicati[118]. Prova ne sia che, con buone probabilità, la prossima edizione dell'*OED* vedrà la luce solo su internet[119].

---

out for some purposes. The *OED*'s bibliography was scarcely computerised at all. It was definitely time to look at the possibilities for a more integrated approach». Elliott / Williams (2006: 258).

[117] Simpson (2003: 49).

[118] Un esempio su tutti, la direzione dell'*Oxford English Dictionary* non ha nascosto di valutare la possibilità di un radicale cambio di politica editoriale consentendo l'accesso libero e gratuito ai servizi on-line (Simpson 2013: 50).

[119] Granger (2012: 2). Si vedano anche gli indirizzi <http://www.telegraph.co.uk/ culture/culturenews/10777079/RIP-for-OED-as-worlds -finest-dictionary-goes-out-of-print.html>, ultima consultazione: 18 ottobre 2015. <http://www.telegraph.co.uk/ culture/books/booknews/7970391/Oxford-English-Dictionary-will-not-be-printed-again.html>, ultima consultazione: 18 ottobre 2015.

2.5 L'*ANGLO-NORMAN DICTIONARY* (AND)

I lavori per l'*Anglo-Norman Dictionary*[120] sono stati avviati nel lontano 1947 con il modesto obiettivo di produrre un glossario per agevolare la lettura dei testi scritti in anglo-normanno. A dispetto degli obiettivi semplici, i primi 15 anni di lavorazione, tuttavia, non produssero grandi risultati[121]; il progetto riprese vitalità negli anni '60[122] e dal 1977 al 1992 furono pubblicati i 7 fascicoli che compongono la prima edizione del dizionario, l'*AND1*[123]. I limiti erano evidenti soprattutto nella prima parte dell'opera[124].

La tecnica di redazione era stata, fino a questo momento, particolarmente laboriosa[125]: lo spoglio delle fonti, a partire da manoscritti o da edizioni a stampa, prevedeva la registrazione dei

---

[120] <www.anglo-norman.net>. Ultima consultazione: 10 novembre 2015.

[121] *Making the Anglo-Norman Dictionary: <http://www.anglo-norman.net/dissem/ data/page2.htm>*. Ultima consultazione: 11 novembre 2015.

[122] Il piano di lavoro fu rifondato e le finalità furono rinnovate nel 1962 da William Rothwell, che decise di andare oltre la semplice lista di parole corredate da significati, anche in considerazione delle due nuove importanti collezioni di dati da cui divenne possibile attingere nuovi materiali: la collezione del vocabolario tecnico e amministrativo raccolta da J.P Collars e il *Dictionary of Law French* di Elsie Shank per la Selden Society. Queste due opere non furono mai completate ma i dati raccolti confluirono nell'AND.

[123] I fascicoli permettevano la pubblicazione in corso prima che fosse tecnicamente possibile una lessicografia evolutiva (§3.4).

[124] «The early AND was (despite an apparently impressive list of texts) overwhelmingly biased towards early and literary materials, and concerned above all to act as a guide to the formal idiosyncracies held to be the distinguishing feature of Anglo-Norman», Trotter (2013: 665). La seconda parte dimostra invece una migliore copertura sulle fonti non letterarie.

[125] Cfr. §3.1.

dati su piccole schede di carta; si procedeva quindi con l'integrazione delle informazioni provenienti dalle altre due fonti (come le collezioni di Collas e Shank). Le schede venivano quindi controllate manualmente, ordinate alfabeticamente, organizzate sotto i lemmi, e poi battute a macchina. Dalla versione dattiloscritta consegnata in tipografia sarebbero poi tornate delle bozze, che oltre ad essere corrette dai frequenti errori tipici del procedimento tipografico, potevano subire una rettifica dei contenuti o un ulteriore arricchimento con nuove citazioni. La tecnica di stampa impiegata da Maney's, il tipografo dell'*AND*, prevedeva però l'uso dei tradizionali caratteri mobili metallici, con la frustrante conseguenza che qualunque tipo di intervento sulle bozze finali doveva rigorosamente rispettare l'esatto numero di caratteri previsto per le righe interessate.

La tecnologia digitale entrò nel progetto subito dopo la diffusione del *personal computer* e dopo l'adesione al progetto da parte di Stewart Gregory e David Trotter. Considerate le premesse, l'innovazione dei *word processor* portò un considerevole aumento della produttività. Nello stesso anno, il gruppo di ricerca si arricchì della consulenza tecnologica di Andrew Rothwell — dell'University of Exeter dove era già stata avviata l'attività pioneristica nell'informatica umanistica del Pallas Project[126] — e Michael Beddow dal Kings College di Londra — altro prestigioso centro di innovazione.

L'informatizzazione fu dunque portata avanti con la riconversione in formato testo delle fonti a stampa con strumenti di OCR su un Kurtzweil Data Entry Machine presso il Kings College. Gli spogli vennero quindi automatizzati con la concordanza automatizzata, «this made possible for the first time for the editors

[126] Dal 2004 CMIT (Creative Media & Information Technology) nel 2004. Le attività del centro sono state definitivamente interrotte nel 2010.

to identify, and investigate sistematically lexical patterns wicht might otherwise have gone unnoticed[127]». Inizialmente il compito fu affidato a TACT (Text Analysis Computing Tools), nota applicazione per concordanze sviluppata presso l'università di Toronto[128], poi su un programma analogo realizzato da Rob Ward. In un secondo tempo si abbandonò il software di terze parti in favore dello sviluppo interno di un'applicazione dedicata per la consultazione on-line e l'integrazione con il vocabolario. Con la digitalizzazione della base di dati e l'automazione degli spogli l'*AND* aveva completato l'informatizzazione delle procedure di redazione.

Il vero cambio di paradigma si verificò con la ripresa dei lavori per la seconda edizione. La ripresa dell'attività redazionale fu contraddistinta da un concorso di circostanze — in cui il fattore tecnologico è perfettamente integrato anche se non sempre facile da isolare — che hanno reso l'impresa un innovativo e inedito modello di ricerca lessicografica.

L'ultimo fascicolo della prima edizione fu pubblicato nel 1992, ma l'esigenza di rinnovare l'opera era avvertita da tempo e i lavori per la realizzazione di una seconda edizione erano cominciati già nel 1989. La sezione A-E si presentava quasi quattro volte più grande rispetto a quella della prima edizione (1100 pagine contro le precedenti 289) e il nuovo progetto lasciava già intravedere una fisionomia completamente rinnovata: «the title of this second edition reserves the old name purely in order to mantain continuity whit the first edition […] Altough the title of the new Dictionary

[127] Rothwell, A. (2005).

[128] <http://projects.chass.utoronto.ca/tact/>. Ultima consultazione: 06 novembre 2015. Cfr. McEnery / Wilson (2001: 2010).

remains the same as before, its contents are very different from those of earlier work[129]».

La pubblicazione on-line, che in una prima fase doveva solo affiancare la pubblicazione cartacea, fu eletta a modalità di distribuzione esclusiva. Per implementare la sezione già prodotta su pagine web si è resa necessaria una riconversione dei file redatti in Microsoft Word in XML e la creazione di una piattaforma web. L'operazione è stata finanziata dall'Arts and Humanities Research Board del Regno Unito (AHRC). Il successo della prima fase di riconversione e pubblicazione in rete produsse, da un lato, un nuovo interesse internazionale verso il progetto e dall'altro un'impressione favorevole sull'AHRC che finanziò la revisione e la pubblicazione della sezione A-H e di quelle successive.

Il circolo virtuoso mutò il modo di intendere il finanziamento della ricerca e il modo di pianificarla. Prima di allora l'AND era stato sostenuto, come molti altri progetti umanistici di lunga durata[130], dalle università presso le quali gli *editors* erano impiegati. La riconfigurazione del dizionario si è direttamente ripercossa sulla natura e sulle forme della ricerca stessa.

> But the AND now has shifted to a founding basis more akin to that of projects in the Sciences (and indeed in continental european lexicography), drawing on *competitively–awarded time–limited* and *target–oriented* [131] research grants, employing full-time research assistants, retaining a tecnical consultant, 'buying' the time spent by editors on overseeing the project from their emploing institutions, and reimboursing

[129] Rothwell, W. (2005).

[130] Zgusta (1971: 348).

[131] Corsivo mio.

those institutions for the resources, office space and other material costs entailed in hosting work on the Dictionary[132].

Oggi l'*AND* è collegato a una base di dati di 78 testi (di cui due non indicizzati per la concordanza e disponibili in PDF). L'accesso libero al testo integrale riconfigura la base di dati come biblioteca digitale[133] dal valore autonomo, per quanto il sistema di consultazione non raggiunga i livelli di raffinatezza di una piattaforma dedicata. L'accesso alla base di dati in modalità *full-text*, peraltro non eccessivamente oneroso dal punto di vista tecnico, potrebbe lecitamente aspirare a diventare una caratteristica diffusa e auspicabile dei vocabolari del terzo millennio, se non fosse per i noti impedimenti relativi al diritto d'autore[134].

Gli articoli possono includere lunghi commenti espandibili tramite un pulsante (per es. la trattazione etimologica s.v. *jacerant*).
Il doppio click su una qualsiasi parola della citazione apre il lemma corrispondente. Se la parola non viene trovata nei lemmi o tra le varianti (*tag* <variant> nello *snippet* riportato più avanti) apre la ricerca nei testi, ma limitata alle forme citate nelle voci. Se la ricerca trova più lemmi corrispondenti, questi vengono visualizzati in sequenza.

Molto interessante la possibilità di salvare la sessione di ricerca nei preferiti del browser (l'identificativo della sessione è una variabile nella *query-string*[135]). La funzione andrebbe però

---

[132] *Same aim, new basis*: <http://www.anglo-norman.net/dissem/data/page2.htm>. Ultima consultazione: 11 novembre 2015.

[133] Per la biblioteca digitale si veda Salarelli / Tammaro (2006);

[134] Per un'introduzione generale all'*open access* si veda De Robbio (2006). Le basi documentarie dei dizionari vengono spesso fornite come corpus interrogabile ma senza l'accesso in lettura al testo completo (come per il DMF e il TLIO); su corpora e diritto d'autore cfr. Barbera (2013: 19-20).

[135] L'URL di una pagina web, oltre a contenere l'indirizzo del sito e l'identificativo della pagina (es. "www.sicilianoantico.it/vocabolario.php") può

perfezionata: le voci visualizzate non vengono memorizzate se a queste si accede da un'entrata del lemmario costituita da un rinvio (distinguibile dal carattere in giallo) o con un doppio click su una parola di una citazione; se si usa il campo di ricerca in luogo del lemmario, la cronologia viene totalmente cancellata; l'elenco è consultabile solo in ordine alfabetico e non permette di rimuovere alcune forme; per azzerare la cronologia si deve rimuovere manualmente la variabile di sessione dall'URL.

L'*AND* è implementato su una piattaforma web in XML/TEI. Un codice di esempio è riportato in Trotter (2011: §16).

```
<entry type="main" key="janglure" id="AND-201-7A6EC4E5-
B1205CE-838CC89F-2AA9BAA2"status="61" lead="gdw">
      <head>
            <form type="lemma">
                  <orth>janglure</orth>
            </form>
            <form type="variant_g1">
                  <orth>janglur</orth>
            </form>
            <form type="deviant">
                  <orth>janlur</orth>
                  <cit>
                        <bibl siglum="TLL" loc="jj59">
                        <I>TLL</I> II 59
                        </bibl>
                  </cit>
            </form>
      </head>

      <gramGrp>
            <pos>s.</pos>
      </gramGrp>
```

---

contenere una parte riservata ad alcune variabili separata dalla parte precedente da un punto interrogativo (es. "www.sicilianoantico.it/vocabolario.php? lemma=blancu"). La querystring può contenere più variabili separate da un *ampersand* (es. "www.sicilianoantico.it/vocabolario.php?lemma=blancu& modo=concordanza").

```
      <!--…-->
</entry>
```

2.6 IL *DICTIONNAIRE DU MOYEN FRANÇAIS* (DMF)

Diretto attualmente da Sylvie Bazin-Tacchella e sviluppato all'interno dell'ATILF[136], il *Dictionnaire du Moyen Français* si occupa del periodo fra il 1330 e il 1500 e, preso atto dei limiti del Godefroy, rappresenta il collegamento tra il *Tobler-Lommatzsch* per l'antico francese e l'*Huguet* per la lingua del Rinascimento[137]. Il *TLF*[138] è però il modello di riferimento[139]:

> Une des finalités essentielles du DMF est de mettre le dictionnaire de l'ancienne langue au niveau des dictionnaires modernes, tout particulièrement dans le traitement des mots sémantiquement complexes[140].

---

[136] L'ATILF (Analyse et Traitement Informatique de la Langue Française) è un attivo laboratorio di ricerca del CNRS (Centre National de la Recherche Scientifique) e dell'Università della Lorena con sede a Nancy. Per una descrizione delle attività si vedano Buchi (2013) e Pierrel / Buchi (2013).

[137] Buchi (2013: 5).

[138] Si ricordi che l'INaLF, Insitut National de la Langue Française, che ha curato la pubblicazione del TLF dal 1971 al 1994, è stato tra le istituizioni fondatrici dell'ATILF, il quale — a seguito dell'informatizzazione del TLF e dei progetti che ne sviluppano i contenuti come il TFL-Étym, ne raccoglie l'eredità ideale. Buchi (2013: 4).

[139] Martin (2012: 2).

[140] Ibidem.

Dopo la pubblicazione di un volume saggio che copriva la sezione A-AH, il progetto imboccò con decisione la strada della distribuzione elettronica, accantonando la pubblicazione su volumi cartacei. Il vocabolario prese la forma di un database lessicale codificato in XML agganciato a un corpus contenente, allo stato attuale, 242 testi per quasi sette milioni di occorrenze, accessibile attraverso le voci del vocabolario o, autonomamente, dalla voce *Les Textes*[141].

Il *Dictionnaire* è oggi una delle opere più interessanti per l'ampio ventaglio di soluzioni tecniche ma la sua caratteristica specifica è di essere un esempio formidabile di impresa lessicografica metodologicamente modellata sulle potenzialità delle tecnologie digitali. L'opera è stata concepita da Robert Martin, direttore del *Dictionnaire* tra il 1982 e il 2000, che ne fece una realizzazione concreta del suo concetto di *lessicografia evolutiva* nell'era digitale[142].

> L'idée centrale qui guide le projet du *DMF* est que l'informatique autorise désormais une lexicographie évolutive: il ne s'agit plus de rédiger le dictionnaire lettre par lettre, ce qui le laisserait dans l'inachèvement aussi longtemps que la lettre ultime n'est pas atteinte, mais plutôt de procéder par une suite d'étapes dont chacune possède sa propre clôture tout en restant ouverte à tous les développements ultérieurs. […]
> L'option choisie pour le *DMF* s'appuie fortement sur l'idée

---

141 <http://atilf.atilf.fr/gsouvay/scripts/dmfX.exe?INIT_SESSION;CRITERE=MENU_RECHERCHE_TEXTES;ISIS=isis_dmf2012.txt;OUVRIR_MENU=4;LANGUE=FR>

142 L'orientamento è stato accolto in seguito a una tavola rotonda che nel 2001 riunì numerosi specialisti a Nancy (Gerner 2005: 156).

> que les dictionnaires d'aujourd'hui, non pas commerciaux mais scientifiques, ne devraient plus être des produits figés que seules peuvent modifier d'hypothétiques rééditions, inévitablement coûteuses et elles-mêmes figées pour longtemps, mais au contraire des bases informatisées, faciles d'accès et ouvertes à peu de frais à tous les enrichissements et à toutes les améliorations que l'on peut estimer souhaitables.

Nel passo riportato il vero elemento di novità non è nella seconda parte, dove si rimarca la risaputa apertura del *medium*. Del resto, anche la versione elettronica della prima impressione del *Vocabolario degli Accademici della Crusca* è emendabile a piacimento, ma qualunque serie di modifiche non ne renderebbe l'impianto diverso da quello del 1612 se non a patto di riscrivere completamente l'opera (e anche questo non basterebbe a renderlo evolutivo). L'apertura del medium è solo il punto di partenza. Quando della modifica e della ristutturazione si fa programmaticamente una pratica fondante e si tocca la pianificazione dell'opera dal principio, allora l'essenza stessa del vocabolario si rivela nella serie di tappe di cui parla Martin, che nell'universo computazionale possono essere, allo stesso tempo e senza che vi sia contraddizione, aperte e chiuse, informativamente compiute e finite ma passibili di sviluppo ulteriore e infinito.

L'aggregazione dei lessici operata dal *Dictionnaire du Moyen Français* (*DMF*) è un interessante insegnamento di redazione per successive fasi di sviluppo, perfetto esempio di lessicografia evolutiva, perfettamente leggibile nella storia del dizionario[143]. La prima versione del vocabolario, il DMF1 del 2002, raccoglieva sotto un lemma comune 26.500 entrate raccolte su 31 lessici (i *Lexiques préalables*) redatti da alcuni collaboratori

[143]Martin (2012: 3-6).

dell'ATILF. Le informazioni provenienti dai lessici erano semplicemente raggruppate ma non sintetizzate, così come nel DMF2 del 2007. Il DMF 2009, oltre all'aggiunta di quattro nuovi lessici, ha avviato una nuova fase "evolutiva" con l'inizio dei lavori di sintesi: con l'etichetta *Synthèse* vengono indicati gli articoli che prendono in considerazione l'intera documentazione (anche non digitale) di cui dispone il DMF; con l'etichetta *Synthèse de Lexiques* vengono invece indicate quelle voci che includono il ricorso ai lessici e alla sola documentazione elettronica. Il processo è stato ulteriormente portato avanti con il DMF 2010 con 500 nuovi articoli di sintesi e nel 2012, con un'estensione quasi completa dello spoglio della documentazione non informatizzata.

Il dizionario è codificato in XML, dopo una conversione da precedenti file in formato Microsoft Word [144] . Tra la documentazione vengono presentati due esempi[145].

Per la voce normale:

```
<ART>
      <VED>AATIR</VED>
      <CODE>verbe</CODE>
      <LEM>aatir</LEM>
      <P>
            <DISC>
                  <IND>Empl. pronom.</IND>
                  <DEF>
                        "S'attaquer à qqn, diriger une
                        attaque contre qqn"
                  </DEF>
            </DISC>
            <EXE>
                  <TEXTE>
```

[144] Gerner (2005: 158).

[145] <http://atilf.atilf.fr/gsouvay/scripts/dmfX.exe?CRITERE=BALISAGE;MENU=menu_balisage;OUVRIR_MENU=MENU_BALISES_XML;ISIS=isis_dmf2012.txt;MENU=menu_accueil;OUVRIR_MENU=1;s=s005a0db8;LANGUE=FR;ISIS=isis_dmf2012.txt>. Ultima consultazione: 18 ottobre 2015.

```
                Et pour ce que trop fort mespris,
                Quant a dame de si haut pris M'osay
                nullement <OCC>aastir</OCC> De plait
                encontre li bastir
                </TEXTE>
                <REF>MACH., J. R. Nav., 1349,
                282</REF>
            </EXE>
        </P>
</ART>
```

Per la voce di rinvio:

```
<ART>
        <VED>AAGE</VED>
        <CIBLE>EAGE</CIBLE>
</ART>
```

La marcatura non standard, ragione per la quale merita qualche precisazione supplementare, si compone di un numero ridotto di tag. Non si rilevano attributi; per segnare il numero di paragrafo, ad esempio, si usa un marcatore [146] . La radice microstrutturale è <ART>.

Si nota subito una distinzione, apparentemente ridondante, tra lemma <LEM> e forma vedetta <VED>. È una piccola traccia lasciata dalla concezione evolutiva per risolvere nel DMF1 l'assenza di normalizzazione sulle entrate dei lessici. La distinzione tra lemma ed entrata consentiva di raggruppare sotto un lemma comune le entrate dei *Lexiques préalables*. La combinazione dei due marcatori è rimasta indispensabile fino a quando i lessici sono stati aggregati cumulativamente (DMF2, DMF 2009, DMF 2010); dal DMF 2012 il processo di affinamento e ripresa delle voci avrebbe potuto tecnicamente mantenere solo il lemma, dove ormai

[146] Tra la documentazione si ritrova il *tag* <NUM>. <http://atilf.atilf.fr/gsouvay/scripts/dmfX.exe?CRITERE=BALISES_XML_P;MENU=menu_balisage;OUVRIR_MENU=MENU_BALISES_XML;BACK;ISIS=isis_dmf2012.txt;MENU=menu_balisage;OUVRIR_MENU=2;s=s003227b8;LANGUE=FR;ISIS=isis_dmf2012.txt>. Ultima consultazione: 18 ottobre 2015.

sono raccolti tutti gli articoli. La distinzione è stata mantenuta perché, per motivi di comodità di consultazione, è stato scelto di usare per il lemma delle grafie modernizzate. Sotto il marcatore <VED> invece, sono state mantenute le grafie più conformi all'uso medievale (si ritornerà sul caso specifico nel paragrafo 06.7.3).

Il marcatore <CODE> contiene l'informazione grammaticale e il suo contenuto deve essere coerente con uno dei valori previsti in una tabella di codici grammaticali previsti dal dizionario[147].

<P> risponde a una funzione simile a quella del marcatore <p> di HTML[148] o di TEI, cioè di contenere un paragrafo di testo (in prosa). Qui però lo si trova inaspettatamente in una posizione gerarchicamente alta. In questo caso contiene un elemento <DISC> («un élément de discours») e un elemento <EXE> («une série d'exemples»); può contenere, inoltre un elemento <REML> («une remarque locale facultative»)[149].

Questi elementi rendono un'idea sufficientemente esaustiva dello stile di codifica del *DMF*; per i restanti punti dello *snippet*, si rimanda a alla sezione *Balisage* del sito, dove sono disponibili per il download anche lo schema e il foglio di stile.

---

[147] Codes Grammaticaux <http://atilf.atilf.fr/gsouvay/scripts/dmfX.exe?CRITERE=CODES_GRAMMATICAUX;MENU=menu_balisage;OUVRIR_MENU=MENU_BALISES_XML;ISIS=isis_dmf2012.txt;MENU=menu_balisage;OUVRIR_MENU=2;s=s003227b8;LANGUE=FR;ISIS=isis_dmf2012.txt>. Ultima consultazione: 18 ottobre 2015.

[148] *HyperText Markup Language*, il più diffuso linguaggio di marcatura utilizzato per la realizzazione delle pagine web. <http://www.w3.org/html>, ultima consultazione 26 agosto 2015.

[149] Balisage d'un paragraphe <http://atilf.atilf.fr/gsouvay/scripts/dmfX.exe?CRITERE=BALISES_XML_P;MENU=menu_balisage;OUVRIR_MENU=MENU_BALISES_XML;BACK;ISIS=isis_dmf2012.txt;MENU=menu_balisage;OUVRIR_MENU=2;s=s003227b8;LANGUE=FR;ISIS=isis_dmf2012.txt>. Ultima consultazione: 17 settembre 2015.

Il *Dictionnaire* fa largo uso del lemmatizzatore LGeRM (Lemmes, Graphies lemmatisées et Règles Morphologiques), sviluppato da Gilles Souvay. LGeRM, oltre a ricondurre a lemma tutte le forme di interi testi del medio francese, è capace di costruire glossari in modalità semiautomatica, elaborare indici lemmatizzati, e gestisce una delle opzioni di ricerca del vocabolario, attiva per *default*, al fine di agevolare la consultazione del vocabolario per l'utente non specialista.

LGeRM è stato sviluppato per il *DMF*, ma, viste le potenzialità, in seguito è stato riadattato con successo anche per funzionare su domini linguistici differenti, cambiando il lessico morfologico alla base: *LGeRM médiéval* è il lessico sviluppato per il periodo 1300-1550; *LGeRM XVIe-XVIIe* per il francese dei secoli XVI XVII; *BGV* concerne le forme verbali dall'antico francese al francese del Rinascimento.

LGeRM viene anche utilizzato come interfaccia per indirizzare eventuali chiamate esterne al dizionario, attraverso link provenienti da siti esterni. Il lemmatizzatore intercetta la richiesta – accettando la chiave di ricerca come parametro nell'URL – la elabora e la reindirizza sul dizionario:

> Le DMF peut être appellé par le biais d'un lien depuis n'importe quelle application sur la toile. LGeRM sert d'interface pour décoder la demande qui est formulée:
>
> > Exemple pour afficher le lemme *amer*:
> > http://atilf.atilf.fr/gsouvay/scripts/dmfX.exe?LIEN_DMF;LEMME=amer
> > Exemple pour afficher l'analyse de la forme *amer*:
> > http://atilf.atilf.fr/gsouvay/scripts/dmfX.exe?LIEN_DMF;FORME=amer[150]

[150] Martin / Souvay (2008: 2019).

A onore del vero, le chiamate dell'URL passano tutte, indistintamente, attraverso l'applicazione dmfX.exe, anche quelle delle pagine di presentazione dell'opera; ne risultano *querystring* molto complesse ed estremamente lunghe, come per gli indirizzi riportati nelle note di questo paragrafo.

Il *DMF* è inoltre inserito in un sistema di risorse informative che lo rendono uno strumento complesso e ricco:

> Ma c'è di più: attraverso gli anni, il *DMF* si è trasformato da semplice dizionario in un portale di ricerca e di redazione con quattro livelli di consultazione collegati in modo da permetterne la navigazione: il dizionario stesso, una ventina di lessici specializzati (come il lessico dell'autore Andrieu de la Vigne o il lessico della letteratura didattica), la banca dati testuale e uno strumento di assistenza all'edizione testuale. Così il *DMF* già rappresenta un dispositivo abbastanza complesso in sé[151].

A questo si dovrà aggiungere la rete di collegamenti che lega il *DMF* agli altri vocabolari. Il *DMF 2012*, come le precedenti edizioni, è collegato direttamente al *Godefroy* e al *TLF*, ma la versione 2015 estenderà le funzioni di collegamento all'*AND2*, al *DEAF*, al *DÉCT* (*Dictionnaire Électronique de Chrétien de Troyes*) e alla *Base des mots fantôme*[152] dell'ATILF[153]. L'operazione rileva un altro punto saliente della lessicografia computazionale:

---

[151] Buchi (2013: 5).

[152] <http://www.atilf.fr/MotsFantomes/>.

[153] Vers un DMF 2015, < http://atilf.atilf.fr/scripts/dmfX.exe?CRITERE=VERS_DMF_ 2015_3;ISIS=isis_dmf2012.txt;OUVRIR_MENU=1;OO1=-1;s=s0c3a01b8;LANGUE= FR;ISIS=isis_dmf2012.txt>. Ultima consultazione: 19 ottobre 2015.

l'integrazione delle risorse e la "complementarità interna[154]", cioè quando il complesso delle risorse a disposizione, in questo caso l'insieme dei progetti dell'ATILF, riesce a sviluppare relazioni sistematiche tra i suoi costituenti tali da diventare un unico «dispositivo lessicografico (e metalessicografico) ragionato, dove ogni risorsa ha il suo ruolo[155]».

Riassumendo, il *DMF* è un vocabolario elettronico, pubblicato solo on-line, che delle tecnologie digitali fa un uso esteso sia al versante della redazione e che a quello del risultato pubblicato, ma la caratteristica esemplare di quest'opera rimane sicuramente l'approccio evolutivo. Tuttavia il dizionario affida la codifica delle voci al linguaggio di marcatura XML tramite uno schema non *standard*, e si affida spesso a un software dedicato proprietario, il lemmatizzatore LGeRM.

2.7 IL *TESORO DELLA LINGUA ITALIANA DELLE ORIGINI* (*TLIO*)

Dal Regio Decreto 735 che interruppe i lavori per la quinta impressione del *Vocabolario degli Accademici della Crusca*, all'attuale Istituto dell'Opera del Vocabolario[156], che si accinge a celebrare l'attività trentennale con 30.000 voci accessibili in rete, intercorrono oltre novant'anni in cui i diversi fili che ricollegano i due momenti si sovrappongono e si intrecciano in una gestazione non sempre lineare. Ritroviamo il capo di uno di questi nel laboratorio di Gallarate di padre Roberto Busa, da dove ha preso

[154] Buchi (2013: 5).

[155] Ibidem.

[156] <www.ovi.cnr.it>.

avvio il nostro discorso, nel corso della visita compiuta da Aldo Duro e Giovanni Nencioni per sperimentare in prima persona le potenzialità delle tecnologie della IBM. Questo filo si dipana attraverso tutti i maggiori centri europei di ricerca lessicografica nella prima metà degli anni '60, grazie a una serie di viaggi di Duro sulle tracce dello stato dell'arte della lessicografia europea dell'epoca[157]. Nonostante le migliori intenzioni e la meritevole volontà di collocarsi da subito sul fronte tecnologicamente più avanzato, non sono mancate false partenze, promesse mancate, e tentativi costellati di «problemi e di drammi, a volte angosciosi[158]». Nulla di strano[159] per un'epoca in cui la schedatura xerografica — che, a voler banalizzare, sta agli spogli elettronici come la posta pneumatica sta all'e-mail — è un processo considerato all'avanguardia. Nessuno di questi fili manca di valore storico-culturale di interesse metodologico, ma motivi di brevità rendono lecito riallacciarsi direttamente al momento in cui il progetto «passa dal mondo delle intenzioni a quello degli strumenti esistenti, dal piano dei sogni a quello della realtà[160]». Del resto, «che il *TLIO* sia il nuovo *Vocabolario della Crusca* si può dire vero solo in senso 'etimologico'[161]».

Prima sezione del vocabolario storico italiano, il *Tesoro della Lingua Italiana delle Origini* è un dizionario storico redatto e pubblicato dall'Opera del Vocabolario Italiano (OVI), istituto del Consiglio Nazionale delle Ricerche. A partire dall'*Indovinello*

---

[157] Cfr. Duro (1973: 17) e Duro (1985: 434-436).

[158] Duro (1973: 18).

[159] L'informatizzazione dei progetti lessicografici prima degli anni '90 presentava alti rischi di fallimento (Gleßgen 2006: 17-18).

[160] Vaccaro (2013: 14) a cui si rimanda per una sintesi sul periodo antecedente alla direzione Beltrami.

[161] Squillacioti (2002: 503).

Veronese e fino al 1375 (anno della morte di Boccaccio e termine per il quale si rispetta l'obbligo di completezza ma che non esclude sconfinamenti[162]), il *TLIO* documenta l'italiano dei primi secoli come un insieme di varietà trattate in modo unitario, configurandosi come «un vocabolario della lingua italiana delle origini nella sua pluralità, piuttosto che delle origini della lingua italiana, intese come la fase antica di quella che sarà poi codificata come varietà nazionale[163]». I redattori sono invitati a restituire un'immagine in cui tutte le varietà siano rappresentate nel modo migliore possibile, rappresentando l'intero dominio italoromanzo. L'unico antecedente che può avvicinarsi a una concezione simile è il *GAVI* di Giorgio Colussi[164], che tra l'altro, quando nel 1998 l'OVI cominciò a pubblicare il vocabolario, riprese in esame tutte le entrate per le quali esisteva una voce corrispondente sul *TLIO*, tenendone conto nella descrizione[165].

Sfruttando le potenzialità della rete, il vocabolario viene oggi pubblicato ad accesso libero sul sito dell'Istituto man mano che le voci vengono redatte, senza aspettare il completamento dell'intera opera e senza dover rigorosamente rispettare l'ordine alfabetico. La pubblicazione in fase di elaborazione del *TLIO* è vicina all'approccio evolutivo, considerato che «la struttura modulare della voce e il metodo di pubblicazione elettronica in rete consentono successivi sviluppi sistematici senza imporre la modifica delle parti non interessate»[166], concentrando la redazione

---

[162] Cfr. anche Vaccaro (2013: 5-6).

[163] Beltrami (2009: 34).

[164] Squillacioti (2010: 2)

[165] Beltrami (2009: 3). Il Battaglia riporta citazioni di testi non toscani ma, a differenza del *TLIO*, senza interesse linguistico (Beltrami 2011: 377) e rimane «sostanzialmente un vocabolario delle *origini dell'italiano*» (Squillacioti 2010: 2).

[166] Beltrami (2013a: 1).

su un circoscritto insieme di punti, «mentre altri tipi di informazioni, come per es. quelle relative agli usi sintattici, possono essere introdotte in modo 'leggero' entro uno schema modulare che potrà facilmente essere integrato in futuro»[167].

Il *TLIO* conduce programmaticamente una lessicografia di prima mano, deducendo cioè l'informazione quasi interamente dalla banca dati elaborata dall'Istituto stesso, il *Corpus TLIO*[168]. Lessici e strumenti lessicografici rientrano nella redazione solo come sussidi di documentazione e controllo. Occasionalmente alcune parole non documentate nei testi del *Corpus* possono legittimamente essere inserite nel vocabolario, ma in questi casi vengono esplicitamente segnalate come voci "fuori corpus". L'integrazione tra vocabolario e corpus è quindi molto stretta:

> The *TLIO* is an Internet-based dictionary which lies balanced, or suspended, somewhere between a language dictionary and a corpus dictionary; in some way it is a corpus dictionary, a glossary of ancient Italian texts, with a philological bias, but with a method which seeks to describe ancient Italian lexis other than simply by text commentary[169].

La banca dati è gestita da GATTO (Gestione degli Archivi Testuali del Tesoro delle Origini), software ideato e sviluppato da Domenico Iorio-Fili presso l'Istituto Opera del Vocabolario Italiano. Nato come strumento finalizzato alla costruzione, gestione

---

[167] Ibidem.

[168] I testi non ancora lemmatizzati vengono invece momentaneamente raccolti nel *Corpus TLIO aggiuntivo*; l'insieme delle due banche dati costituisce il *Corpus OVI dell'Italiano antico*.

[169] Beltrami / Fornara (2004: 373).

e interrogazione del corpus *TLIO*[170], GATTO si è progressivamente arricchito di funzionalità e strumenti.

GATTO si fonda su dati strutturati in un database[171] relazionale, e ne adotta la logica, com'è immediatamente visibile anche da un'analisi superficiale del database di un corpus qualunque (la tabella *For2*, per esempio, registra il formario completo, *Ind* che conserva i contatori per ogni testo e le associazioni con note e traduzioni). Se il programma sembra allontanarsi per un momento da questa logica nella registrazione delle singole occorrenze, non è per ripiegare su una ricerca *full-text* ma solo per sollevare il database dai compiti che richiederebbero una scalabilità che Access non poteva offrire negli anni in cui il programma è stato sviluppato[172]. La concezione robusta sembra

---

[170] Per la verità il primo corpus implementato in GATTO era composto da un sottoinsieme di 21 testi fiorentini duecenteschi estrapolato dal *Corpus TLIO* per il progetto Italant — dal quale è derivata la *Grammatica dell'Italiano Antico* (Salvi / Renzi 2010) — e da cui poi è stato sviluppato anche il *Corpus Taurinense* (cfr. Beltrami 2010: 332 e Barbera 2009). La versione 2 del programma, prima di entrare in uso per la redazione del *TLIO* è stata profondamente revisionata per adeguarsi alla diversa dimensione della base di dati.

[171] Nella storia del programma i dati sono stati affidati a tre diversi *DBMS* (Database *Management System*): Microsoft Access, database storico di GATTO, limitato in termini di prestazioni e di scalabilità ma versatile; Microsoft SQL Server, impiegato per GATTOWEB, potente e completo ma poco pratico per l'interrogazione in locale; SQLite, minimale ma potente, candidato a divenire il motore delle future versioni di GATTO, dalla 4 in poi.

[172] I motivi che potrebbero aver incoraggiato l'iniziale utilizzo di Access sono facilmente ipotizzabili e condivisibili: è stato uno dei primi DBMS disponibile per questa fascia di applicazioni, immagazzina le basi di dati su un unico file, e, soprattutto, ha sempre vantato un notevole livello di integrazione con l'ecosistema di sviluppo Microsoft, col quale GATTO ha avuto un legame che ha costituito uno dei suoi punti di forza (Iorio-Fili 2007: 365).

rispondere perfettamente all'affidabilità richiesta al programma[173] ma l'autore ne ha anche rilevato la maggiore velocità in risposta sulle interrogazioni[174].

Grazie a GattoWeb — una controparte *server-side* che permette di pubblicare in rete i corpora creati in GATTO — l'intera base empirica del vocabolario diventa liberamente e completamente accessibile on-line. L'interfaccia web dispone di quasi tutte le funzioni di interrogazione della piattaforma utilizzata dai redattori, creando una piattaforma preziosissima per la combinazione tra ricchezza della base di dati (oltre 20 milioni di occorrenze) e raffinatezza degli strumenti di interrogazione. Il pacchetto di installazione GATTO può essere liberamente scaricato, riprodotto e usato per la ricerca senza fini di lucro, con l'unico obbligo di citare il software in tutte le pubblicazioni derivate, proponendosi quindi come un programma di gestione di corpora testuali che si rivolge a tutti gli studiosi interessati alla costituzione e alla manipolazione di corpora testuali di diversa natura[175].

Il *Progetto Artesia* ha scelto di implementare il proprio *Corpus* su GATTO[176] per disporre di uno strumento efficace e ben

---

[173] Iorio-Fili (2007: 367).

[174] Ivi, pp. 372-373.

[175]Il download è disponibile alla pagina <http://www.ovi.cnr.it/index.php?page=scaricare-gatto> (ultima consultazione: 14 ottobre 2015). Nelle distribuzioni del *Corpus Artesia*, GATTO viene distribuito in allegato nella versione più recente disponibile al momento della pubblicazione e sulla quale è stata verificata la compatibilità del corpus pubblicato.

[176] Non si tratta certo di una novità metodologica, si veda a questo proposito il *Vocabolario del Pavano* (Paccagnella 2012: LIII) o il progetto per un dizionario storico del napoletano (De Blasi / Montuori 2006), di cui però non si hanno più notizie (la pagina del progetto non segnala aggiornamenti recenti: <http://www.unior.it/ateneo/3586/1/dizionario-storico-del-napoletano.html>; ultima consultazione: 9 dicembre 2015).

collaudato e per limitare lo sviluppo immotivato di soluzioni personalizzate in mancanza di concrete esigenze di personalizzazione (Pagano 2011: 315). I benefici si sono riscontrati soprattutto nell'evitare la primissima fase di progettazione e analisi del sistema, dove l'esplicitazione di requisiti e dominio e la sistemazione di una miriade di dettagli tecnici coinvolge anche i nodi teorici sui quali fondare il corpus (la tokenizzazione, il sistema di marcatura, i sistemi di datazione).

Il *VSM*, così come stabilito da Pagano (2011) e Pagano (2012), adotterà deliberatamente il modello del *TLIO* allo scopo di conformarsi a un illustre e autorevole modello lessicografico condiviso. La modularità della voce del *TLIO* cui si è fatto cenno, descritta in dettaglio nel *Normario* di Beltrami (2013), si sostiene su una microstruttura che ripartisce rigorosamente l'informazione. Questa caratteristica, come si vedrà più avanti, permette di tentarne il riadattamento su progetti lessicografici diversi, come se si trattasse di uno *standard* o — in combinazione con le indicazioni del *Normario* e con gli strumenti di interrogazione — di un '*framework* lessicografico'. Per la centralità che il *TLIO* occuperà nei capitoli successivi, ulteriori aspetti del *Tesoro* saranno trattati successivamente.

3 I COSTITUENTI DEL VOCABOLARIO ELETTRONICO

3.1 IL SISTEMA INFORMATIVO

Un primo fondamentale motivo di affinità tra dizionario e l'informatica si basa su un principio basilare: ciascuna operazione lessicografica presuppone un 'sistema informativo'[177] che raccoglie, organizza e gestisce le informazioni necessarie per perseguire gli scopi propri dell'impresa. Il sistema informativo non va confuso col sistema informatico e con la sua specifica realizzazione, il database, per quanto oggi nel concreto vengano spesso identificati; gli aspetti relativi all'automazione e al supporto tecnologico non sono consustanziali al sistema informativo e neppure condizioni necessarie al suo allestimento.

Il problema della gestione e della catalogazione dell'informazione testuale era avvertito nella pratica lessicografica nell'epoca pre-informatica e ciascuna impresa disponeva di un proprio sistema informativo costituito da numerose schede cartacee (*citation slip*) e classificato in sterminati e onerosi archivi.

> «L'analisi della particella *et* nello schedario del *Thesaurus Linguae Latinae* a Monaco implicava, per lo studioso che avesse avuto l'opportunità di accedervi, lo spoglio di 40 cassetti di 1200 schede ciascuno».[178]

[177] Atzeni et al. (2002: 1).

[178] AA. VV. (1996) cit. in Marello (1996: 156).

La stima del volume medio dei dati implicati nell'attività è strettamente correlata alla grado atteso di esaustività e di sistematicità ma anche all'atteggiamento lessicografico nei confronti del dato empirico[179]. La memorizzazione dei dati relativi alla ricerca su nuovi supporti è delle sfide della prima metà del XX secolo, come dimostra il caso delle schede perforate[180]; Bush (1945) affidava le speranze di sistemi informativi più efficienti ai progressi della chimica e dell'ottica, ipotizzando fotoriproduzioni di 3 millimetri quadrati per il Memex, progetto visionario e pioneristico per un sistema meccanizzato per la gestione dei dati nella ricerca che rispondeva anche alla domanda insistente di nuove forme di miniaturizzazione.

Quando si giunse all'elaborazione digitale non si verificò solamente la desiderata dematerializzazione degli schedari ma mutarono le modalità e i tempi di accesso alle informazioni. La mancanza di automazione implicava una perdita di efficienza e lo spreco di una buona quantità di tempo e di impegno del lessicografo in lente e faticose cernite manuali. In ambito lessicografico questi fattori hanno sempre condizionato i principi di spoglio e selezione della tavola dei citati e la qualità della spigolatura delle fonti.

---

[179] Si veda la testimonianza di Murray (1979) citata da Rundell / Kilgarriff (2011: 258) a proposito del lavoro nell'ambito dell'*Oxford English Dictionary*.

[180] L'invenzione della scheda perforata, supporto che permise le prime sperimentazioni lessicografiche mediate dal calcolatore, precede di parecchi anni l'invenzione del computer. «Help in storing and processing of language material already came at the beginning of the 20th century. Together with special electromechanical machines, punched cards were invented by the German-American Herman Hollerith (1860-1929), who used them in an American census. The first scientist using Hollerith-Cards was father Roberto Busa, S. J., who in 1946 started the Index Thomisticus […] Busa could inspire the just founded IBM, the follower of the Hollerith Company, to fund his project» (Lenders 2013: 969-970).

Tradizionalmente i sistemi informativi venivano allestiti manualmente da soggetti che non necessariamente coincidevano con i redattori (spesso si trattava di lettori volontari). Il lettore si adoperava nella lettura dei testi e selezionava gli esempi da riportare nelle schede da consegnare ai redattori (spoglio di scelta). In questi casi la qualità della raccolta dipendeva quindi dall'affidabilità dei lettori e dall'accuratezza delle linee guida previste dal lessicografo[181] ma l'aspettativa qualitativa era più bassa rispetto a quella di uno spoglio integrale condotto in prima persona dal redattore:

> Il principio dello spoglio integrale si fonda ovviamente sulla convinzione che sono più fallaci i lettori dei testi da spogliare che i redattori degli articoli, e che pertanto la selezione degli esempi deve avvenire preferibilmente ad opera dei redattori, i quali possiedono una preparazione ed una uniformità di criteri senza subbio superiori[182].

Lo spoglio di scelta continuava ad essere largamente impiegato perché sui lenti sistemi informativi tradizionali migliorava il rendimento dei redattori e ottimizzava le risorse:

> Il principio dello spoglio di scelta si fonda invece, oltre che — per certi periodi — sulla materiale impossibilità della schedatura meccanica, su un calcolo economico: non conviene sommergere i redattori sotto una valanga di materiale schedato, che essi dovranno sceverare a prezzo di tempo e fatica, ma piuttosto affidare fiduciosamente a buoni lettori tale compito[183];

Nella lessicografia quindi qualità ed economicità sono stati

---

[181] Simpson (2009: 56).

[182] Nencioni (1955: 133).

[183] Ibidem.

termini inversamente proporzionali e il loro bilanciamento spesso implicava un compromesso. L'interrogazione dei corpora mediata dal computer ha però reso estremamente più rapido, e quindi economico, lo spoglio integrale delle fonti da parte dei redattori.

Il passaggio non fu immediato. Alle origini degli spogli elettronici la mole di informazione prodotta gravava ancora sul redattore sotto forma di tabulati. A questo livello di sviluppo tecnologico la fase di lemmatizzazione poteva incaricarsi anche di scremare i contesti da esaminare; «la selezione delle occorrenze da lemmatizzare corrispondeva dunque, entro il metodo dello spoglio elettronico, alle letture e agli spogli manuali con cui tradizionalmente si ottenevano i materiali per i dizionari storici[184]».

Solo l'accesso diretto e l'interrogazione dinamica dei corpora elettronici hanno reso lo spoglio esaustivo realmente efficiente, incentivando quindi il ricorso allo spoglio integrale. Certamente una preselezione delle citazioni riesce ancora oggi a risparmiare tempo al lessicografo, riportando, tra più occorrenze di un termine, quelle più adatte al vocabolario. Per alcuni, ad esempio Simpson (2009: 63), i due metodi potrebbero continuare a convivere e integrarsi proficuamente, perché l'esperienza e la competenza linguistica del lettore contribuiscono a rilevare regolarità e irregolarità nel linguaggio e la lettura esperta rimane sempre maggiormente sensibile per ciò che non è coperto dal vocabolario o per i fenomeni rari. Ma lo spoglio tradizionale riveste oggi un ruolo sempre più marginale, e la collaborazione di lettori esterni assume forme diverse come il *crowd-sourcing* dell'*OED*[185].

Quantificare lo scarto qualitativo tra l'epoca pre-elettronica e la nostra è ovviamente impossibile, ma il tentativo di valutare statisticamente l'affidabilità delle informazioni nell'*OED* tentato da

[184] Beltrami (2008: 35).

[185] Cfr. §2.4.

Schäfer (1980) ha ricavato una serie di valutazioni significative sulla discrepanza tra i spogli manuali e spogli automatizzati, confermando lo scarto tra le due epoche [186]. In molti casi di schedatura tradizionale, i lettori si sono lasciati sfuggire alcune retrodatazioni di parole o sensi, in altri casi hanno tralasciato termini che non sono entrati a far parte del lemmario del dizionario [187]. Del resto, verso la metà degli anni '90, con la pubblicazione delle prime basi di dati informatizzate on-line, le ripercussioni sono state avvertite chiaramente all'interno dell'*Oxford*:

> The emergence of such databases has had a profound effect on the revision of the *Oxford English Dictionary* […] We started the revision and update in earnest in the mid nineties, when the firs small historical databases were beginning to appear, in fits and starts, in the internet. Previously our access to machine readable texts had been primarly through computer concordances — some of wich had been available in one form or another since the late 1960s.
>
> The results from these emergent resources caused us to review the early draft revisions we had made to the *OED* in the 1990s and […] to re-edit entries in the light of this new historical materials. The results were astounding[188].

---

[186] «The aim of the endeavour is to function as a pilot project investigating the dictionary's reliability as a source of documentation; in other terms, to develop criteria of evaluation allowing present-day scholars to use the dictionary's data with due discernment and awareness of the discrepancy between the pre-computer age methods employed in the original compilation and those available nowadays, and thus to correct such distortions as may result from the difference» (Van Noppen 1983: 707).

[187] Simpson (2009: 56).

[188] Ivi, p.58.

Dematerializzazione e velocità sono fenomeni che si rilevano nel passaggio dall'analogico al digitale spinti dalle caratteristiche e dalle *performance* dei nuovi sistemi e nella lessicografia possono aiutarci a rendere efficienti ed economiche le metodologie più onerose in fase di preparazione dell'opera. La tecnica trascina con sé anche forme alternative di organizzazione razionale dei sistemi informativi, le quali non si limitano a incrementarne l'efficienza ma agiscono sul modo di rappresentare e costruire i dati. La forma tipica che questi sistemi hanno assunto negli ultimi anni è quella dei database, fondati sul paradigma relazionale.

3.2 LA LOGICA RELAZIONALE

Tra la pluralità di scelte che si sono profilate per rappresentare qualsiasi tipo di informazione in ambiente digitale si rintraccia una particolare conformità — o, se vogliamo, profonda affinità — tra lessicografia e database[189]. Se consideriamo il vocabolario nella sua anatomia interna, l'identificazione con la base di dati può considerarsi plausibile: «it's sometimes useful to think of the dictionary in 'database' terms, as a *set* of components (such as definitions, etymologies, and pronunciation) that can be dealt with discretely[190]».

---

[189] Un classico manuale di riferimento è quello di Atzeni et al. (2002). Per una sintetica ma efficace introduzione pratica è possibile consultare la pagina sulle *Nozioni fondamentali sulla progettazione di* database, sul sito di supporto di Microsoft: <https://support.office.com/it-it/article/Nozioni-fondamentali-sulla-progettazione-di-database-eb2159cf-1e30-401a-8084-bd4f9c9ca1f5>. Ultima consultazione: 14 novembre 2015. Un'introduzione ai database per l'umanista, con alcuni esempi di classificazione documentaria, è Scotti (2003).

[190] Atkins / Rundell (2008: 22).

La corrispondenza si sostiene sulla specifica natura informativa dei vocabolari [191] i quali, in via di principio, sono costituiti da elementi testuali (o, se vogliamo, stringhe alfanumeriche) di lunghezza contenuta, organizzate in blocchi separati che si comportano come unità discrete incasellate all'interno di una struttura relativamente rigida e strutturata ma senza che vi sia conflitto con la dinamicità dei dati: «come forma culturale il database rappresenta il mondo come un elenco di voci non ordinate che si rifiuta di ordinare[192]», o, più realisticamente, che si ripropone di riordinare e aggregare in modi diversi a ogni interrogazione.

Su un vocabolario, anche sulla fissa e lineare pagina stampata, la conoscenza contenuta è parcellizzata in blocchi di informazioni che non intrattengono relazioni tra loro se non con quelli immediatamente correlati nella gerarchia, la quale — per la stabile convenzialità che la caratterizza — è altamente predicibile[193] (motivo per il quale il lettore può decodificarla senza fatica anche su un dizionario mai visto in precedenza).

Si potrebbe giustamente notare che all'interno di un vocabolario l'insieme delle voci costituisce un sistema organico, un ecosistema coerente, ma un'osservazione simile si colloca su un livello di descrizione diverso. In questo momento, per la linearità della trattazione si possono lecitamente tralasciare considerazioni di questo tipo in favore della configurazione astratta delle informazioni; considerazioni analoghe ci permettono di trascurare la presenza di sovrapposizioni parziali nella gerarchia idiosincratica degli elementi di alcuni vocabolari.

---

[191] «The increased focus on information in present-day society has made it clear that lexicography is an information discipline *par excellence*», Tarp (2012: 56).

[192] Manovich (2002: 281).

[193] Trotter (2011: §12).

La parcellizzazione dell'informazione in unità autosufficienti spiega il motivo per il quale le opere di consultazione [194] sono estranee alle reticenze che molti lettori mostrano per il consumo di prodotti di editoria elettronica. Con i dizionari (così come per i glossari, le enciclopedie, i manuali, ecc.) si avverte meno l'abbandono della fisicità del volume cartaceo e viene maggiormente percepita e apprezzata la riduzione dei tempi di accesso all'informazione[195].

> Certes, elle ne fait pas oublier l'attrait de la page imprimée, encore moins le plaisir de feuilleter le livre. L'artifice de l'écran, sa relative exiguïté, les aléas qui demeurent de la connexion, tout cela peut faire naître la nostalgie de la lecture traditionnelle. Cela dit, un dictionnaire ne se lit pas, mais se consulte sur tel ou tel point particulier. Ce sont les ouvrages de consultation qui souffrent le moins des faiblesses de l'écran[196].

Nel bilancio tra costi e benefici questi ultimi risaltano con evidenza: velocità di accesso, lettura non lineare, suggerimenti

---

[194] Le opere di consultazione si caratterizzano secondo Tarp (2012: 56) per la volontà di rispondere a bisogni informativi specifici (*specific information needs*), in opposizione ai bisogni informativi generali (*global information needs*) correlati a una conoscenza più profonda di un soggetto di interesse. Vale però la pena di ricordare un passo dell'*Introduzione* di Salvatore Battaglia al *GDLI*: «un dizionario non si legge, si consulta appena. […] Anche le definizioni dei significati e delle proprietà verbali confinano il dizionario tra le opere di rapida e occasionale consultazione. E tuttavia le citazioni, per quanto siano di necessità frammentarie e discontinue, riconducono il dizionario nell'ambito della letteratura e della vita» (*GDLI* p.V).

[195] Se non bastasse il riferimento a una percezione comune facilmente riscontrabile a fugare possibili obiezioni derivanti dai risultati di Bergenholtz (2011), basti qui circoscrivere l'affermazione ai tempi di accesso del medium a prescindere dall'effettiva struttura e dai contenuti lessicografici.

[196] Martin (2012).

nella ricerca, collegamenti ipertestuali. Su questi converrà insistere per fornire al lettore la migliore esperienza di consultazione possibile.

Sull'ipertestualità il vocabolario dimostra ancora una volta una particolare attinenza con l'informatica. L'ipertesto è un concetto più anziano del calcolatore, tanto che le prime teorizzazioni, da parte di studiosi come Vannevar Bush[197], Doug Englebart e Ted Nelson, precedono le implementazioni concrete in contesti digitali[198].

Ma il rinnovamento profondo sollecitato dai database riguarda l'approccio alla gestione dei dati: se un semplice file testuale fornisce meccanismi di accesso e condivisione semplici, i database gestiscono i dati in modo integrato e flessibile, corredandoli con strumenti più potenti e con la capacità di gestire la struttura propria del 'modello di dati'.

> Un modello dei dati è un insieme di concetti utilizzati per organizzare i dati di interesse e descriverne la struttura in modo che essa risulti comprensibile a un elaboratore. Ogni modello dei dati fornisce meccanismi di strutturazione, analoghi ai costruttori di tipo dei linguaggi di programmazione, che permettono di definire nuovi tipi sulla base di tipi (elementari) predefiniti e costruttori di tipo. […][199]

Nella progettazione relazionale i dati sono raccolti in tabelle; un database può essere composto da una o più tabelle. Le righe delle tabelle sono chiamate *record*, e ogni colonna *campo*. La

[197] Bush (1945).

[198] Di Tonto (2003).

[199] Ivi, p.3.

strutturazione impone l'identificazione dei tipi di *entità* [200] da registrare, essenziale per la corretta suddivisione dei dati sulle tabelle e, all'interno di ciascuna tabella, sui campi. Ciascun dato andrà scomposto nelle sue componenti minime (per esempio il nome di un individuo andrebbe scomposto in nome e cognome per consentire ricerche mirate); il principio, valido in generale, andrà comunque di volta in volta misurato con la valutazione del livello di dettaglio richiesto dal progetto (per esempio, per la scheda anagrafica di un autore antico, a differenza del nominativo di una prenotazione aerea, non sarà necessario distinguere tra nome, cognome, titolo e iniziali).

Il modello relazionale pone l'accento e forza la corretta interpretazione dei dati al fine di eliminare la ridondanza, evidenziarne le relazioni (e di conseguenza garantirne l'integrità). Una progettazione corretta richiede che siano rispettati almeno 4 imperativi:

- Suddividere le informazioni in tabelle sulla base di categorie generali che rendono conto delle analogie tra le entità rappresentate.
- Identificare le informazioni sulle quali il database può unire e collegare i dati contenuti su tabelle diverse.
- Garantire l'accuratezza e l'integrità referenziale tra le informazioni.
- Tenere conto delle esigenze di elaborazioni dei dati e della creazione di *report*.

Il *design* efficace di una base di dati è responsabile dei trattamenti computazionali possibili, dello spettro delle funzioni di ricerca che possono essere implementate sulle informazioni e

[200] Cfr. Atzeni et al. (2002).

l'interoperabilità con banche dati simili. Per evitare di creare futuri ostacoli, per l'ottimizzazione del sistema e per non porre vincoli alle possibilità di analisi delle informazioni di un vocabolario elettronico, i principi della logica relazionale non dovranno mai essere persi di vista.

3.3 AUTOMAZIONE

L'informatica ci spinge a ripensare preventivamente l'intero processo lessicografico in termini di flusso di elaborazione dell'informazione e di manipolazione algoritmica automatizzata. Si consideri la "filiera" dei principali momenti lessicografici, così come analizzato da Rundell / Kilgarriff (2011: 261) in termini d'automazione computazionale:

- corpus creation
- headword list development
- analysis of the corpus:
    - to discover word senses and other lexical units (fixed phrases, phrasal verbs, compounds, etc.)
    - to identify the salient features of each of these lexical units
        1) their syntactic behaviour
        2) the collocations they participate in
        3) their colligational preferences
        4) any preferences they have for particular text-types or domains
- providing definitions (or translations) at relevant points

- exemplifying relevant features with material gleaned from the corpus
- editing compiled text in order to control quality and ensure consistent adherence to agreed style policies

La promessa delle nuove tecnologie di rendere quasi istantanee tutte le operazioni ripetitive e meccaniche ha incentivato lo sviluppo di soluzioni automatizzate per ciascuno dei punti elencati. Purtroppo non tutte le soluzioni sono adatte a tutti i tipi di dizionari e conformi con le scelte lessicografiche a cui obbediscono. Per questo motivo un progetto prudente dovrà preventivamente individuare tutti quei punti della filiera caratterizzati da incoerenza endemica, zone di resistenza che si annidano in tutte quelle attività in cui le istanze lessicografiche si rivelano, totalmente o in parte, irriducibili a procedimenti algoritmici. Ai nostri fini se ne affronteranno due sulla base delle attività già considerate dal *Progetto Artesia*: la costituzione del corpus e la lemmatizzazione.

Rundell / Kilgarriff (2011: 262) muovono dal concetto di "web corpus", per arrivare a considerare «things of the past» problemi come la scarsità dei dati. Il web costituisce senza ombra di dubbio una banca dati sterminata; il ricorso alle ricerche su Google è diventato una pratica comune nella redazione dei dizionari che spesso si affianca alle interrogazioni delle basi di dati[201]. Anche tralasciando i problemi di finitezza e definizione[202], il

[201] Si veda, per es., il caso dello Zingarelli, dove la sensibilità per il neologismo trova nelle ricerche di Google «il terreno privilegiato» per una «verifica sul campo» lessicografica (Cannella, 2013: 154).

[202] «Il WWW è sempre in movimento, non si può considerare né definito (almeno non nel senso di consentire la ripetibilità degli esperimenti) né finito (nel senso di costituire un insieme numericamente dato, su cui si possano fare operazioni statistiche deterministiche)», (Barbera 2013: 21).

sogno di creare automaticamente un corpus attraverso la rete è, almeno per il momento, irrealizzabile per un dizionario storico come il *VSM*, sebbene in certi casi si possano trovare in rete indicazioni utili per le fasi della lingua più recenti[203]. E non solo perché la rete non è in grado attualmente di fornire in numero congruo i materiali di cui si ha bisogno e gli strumenti adeguati per isolarli, quanto perché le dovute cautele filologiche richieste a un dizionario storico vietano l'acquisizione di massa[204]. Pertanto, relativamente al primo punto dello schema (*corpus creation*), vengono meno i presupposti per i quali i tempi di allestimento dei corpora dovrebbero essere ridotti «from years to weeks, and for a small corpus in a specialized domain, from months to minutes[205]». Nel nostro caso, l'allestimento del *Corpus Artesia* ha infatti richiesto alcuni anni di lavoro e una serie, non ancora conclusa, di aggiornamenti periodici descritta in Pagano / Arcidiacono (2013), a cui va aggiunto l'ultimo aggiornamento presentato nei prossimi capitoli.

Al secondo punto troviamo l'identificazione delle forme vedetta e la lemmatizzazione del corpus. In questo settore la precisione delle procedure automatiche non ha mai raggiunto l'eccellenza, come peraltro rilevato dagli stessi Rundell / Kilgarriff (2011: 264), con complicazioni di una certa entità in presenza di testi antichi. Il nostro caso specifico non è comunque privo di elementi positivi: il *VSM* eredita dal *TLIO* un *set* di descrittori grammaticali semplificato[206], con l'ulteriore vantaggio che il *Corpus Artesia* è composto da materiali riconducibili a un unico

[203] Si veda, per es., Pascual (2013: 26).

[204] Si veda Martin (1998: 965), Pagano (2009), De Robertis (1985).

[205] Rundell / Kilgarriff (2011: 262).

[206] Esperti (1979), ma con alcune modifiche. Si vedano anche Beltrami (2013) e Beltrami (2005: 338 e sgg.).

volgare italo-romanzo.

Sebbene non sia possibile consegnare l'intera fase al calcolatore, gli eccellenti risultati ottenuti da Iorio-Fili (2010) circoscrivono il carico di lavoro manuale previsto alla lemmatizzazione del corpus di addestramento[207], all'integrazione delle coppie forma-lemma non presenti nel corpus di addestramento e alla revisione finale. Questi motivi hanno spinto a posticipare l'avvio della lemmatizzazione a un momento successivo al rilascio di GATTO 4, ma la riduzione dei tempi non significa un annullamento del lavoro manuale, a patto di non accettare consistenti percentuali di errore.

Tralasciando per il momento stime e previsioni per i punti successivi dello schema, nulla lascia comunque presagire che per i punti successivi il tempo possa diventare una variabile trascurabile, e ciò dovrebbe rappresentare un invito a considerare la velocità del digitale non in termini assoluti ma in relazione ai cicli operativi dell'attività lessicografica. Nelle scienze "dure" generazioni tecnologiche e cicli di scoperte tendono ad avvicendarsi seguendo ritmi che combaciano: nuove apparecchiature, migliorate capacità di calcolo e nuovi ambienti di simulazione catalizzano l'emergere di ricerche e risultati, così come ogni avanzamento nelle scienze applicate porta a nuovi strumenti. Nel nostro caso questo rapporto tende ad assestarsi su intervalli asincroni, con ritardi che rendono la ricerca vulnerabile all'obsolescenza, minacciando la trasportabilità dei dati.

Del resto i dizionari hanno sempre mancato gli appuntamenti con il pubblico. Zgusta (1971: 348) dichiarava di non aver mai trovato un dizionario che avesse rispettato i tempi preventivati, e

[207] Tuttavia, viste le dimensioni ridotte del *Corpus Artesia*, non è detto che la dimensione più adeguata del corpus di addestramento sia così limitata rispetto al totale.

stimava un aumento dei tempi medio del 100-150%[208]. La prima *Crusca*, sebbene sia stata completata in appena vent'anni, si fece attendere con ansia nonostante le diverse dichiarazioni che davano la pubblicazione per imminente[209]; nelle previsioni iniziali la redazione dell'*OED* avrebbe richiesto 10 anni, mentre poi ne prese 49[210]. Nicola Zingarelli, che firmò un contratto dove si impegnava a consegnare l'opera «entro nove mesi al massimo», impiegò 10 anni[211]. Negli anni '60 per il vocabolario storico dell'italiano furono previsti cinquant'anni di lavorazione (10 per il *Tesoro* dei primi due secoli e 40 per la restante parte)[212].

Secondo Zgusta (1971: 349) una stima precisa è possibile solo a circa metà del completamento. L'informatica può e deve aiutare a limitare i rischi. In primo luogo, come evidenziato da (Trotter 2011: 6), la digitalizzazione in sé impone ordine e consistenza sui progetti basati su larghe scale di tempo e quindi soggetti a una variabilità ampia dettata dai fattori umani. In seconda istanza lo sviluppo può essere pianificato per tappe autonome più facili da gestire e destinate a integrarsi in un'opera complessivamente coerente.

---

[208] <http://public.oed.com/history-of-the-oed/>, ultima consultazione: 30 novembre 2015. Zgusta (1971: 349) riporta una stima di 13 anni contro 39 effettivi.

[209] Maraschio / Poggi Salani (2014: 34-35).

[210] Zgusta (1971: 349).

[211] Cannella (2013: 151).

[212] Vaccaro (2013: 5).

3.4 LESSICOGRAFIA EVOLUTIVA

«Parlare di completamento per i dizionari accademici è come parlare di arcobaleni che si allontanano man mano che ci si avvicina[213]». Il fisiologico allungamento dei tempi, le risorse e il sostegno di editori privati o dei finanziamenti pubblici[214] sono solo la parte contingente del problema. La compiutezza impossibile a priori e al lessicografo è negato il privilegio di apporre una volta per tutte la parola fine alla sua opera: l'oggetto della lessicografia riguarda una classe aperta di fenomeni, a partire dall'insieme delle parole, impossibile da circoscrivere e fissare nella sua interezza[215]. L'esaustività della descrizione è un obiettivo impossibile da raggiungere, il dizionario è condannato a una condizione ineliminabile di provvisorietà. A dispetto degli imperativi dell'editoria commerciale — che rimane comunque maggiormente interessata ai più remunerativi e circoscritti dizionari di uso pratico che alle grandi opere lessicografiche[216] — il vocabolario rimane sempre un *work in progress*[217].

---

[213] Klein (2013:14).

[214] Klein (2013: 15). Gli argomenti trattati in questo paragrafo non sono estranei al problema della stabilità delle risorse: «the capacity constantly to update requires of compilers an ongoing and seemingly endless commitment to projects which may depend on external funding, and which may have been conceived as relatively short term ventures (the current and almost universal Europe-wide insistence on funding only four- or five-year projects creates major problems for serious lexicographical projects). It is essential, too, to ensure the longevity of electronic resources. In the case of the printed book this is a matter of using acid-free paper: guaranteeing the long-term survival of an electronic dictionary is by no means as straightfoward», Trotter (2013: 663).

[215] Zgusta (1971: 16).

[216] Klein (2013:14).

[217] Atkins / Rundell (2008: 2).

Anche per De Mauro[218], l'impossibilità deriva da ragioni interne al sistema linguistico aperto e, già nel 1994, nel corso dell'Incontro di Studio *Fabula in tabula. Dal racconto degli indici alla retorica del testo elettronico* ne vede una soluzione nelle banche dati on-line:

> La lingua, non costituendo un sistema chiuso e matematico per la sua innovatività permanente e irregolarità nella formazione, non permette mai a un dizionario di essere definitivo ed esaustivo, cosicché la creazione di banche dati per via Internet offre nuova adeguatezza e possibilità di aggiornamento sconosciute alle opere cartacee.

La lessicografia di lingue antiche non fa eccezione, per quanto esonerata dal continuo confronto col cambiamento linguistico e con l'incessante nascita di nuove parole e sensi; anche il vocabolario storico è quindi un «*opus perpetuum*, che rispecchia il perpetuo moto della cultura[219]»:

> Scholarly dictionaries (even once they have been printed) are never finished. Even with a language that is long since "dead" and so has ceased to evolve in the mouths and at the hands of its original speakers, previously unknown documents are still being found, or previously known ones read in a new light (sometimes quite literally so: technological advances in ultraviolet readers, multispectral photography and xray imaging are allowing scholars to see things in medieval manuscripts that had apparently long since vanished from view)[220].

---

[218] Cit. in Buzzetti / Quaquarelli (1995: 240-241).

[219] Nencioni (1955: 120).

[220] <http://www.anglo-norman.net/dissem/data/page2.htm>. Ultima consultazione 5 aprile 2015.

Nella lessicografia computazionale lo stato provvisorio del dizionario viene assorbito e retto tramite il mezzo tecnico in una nuova forma di risorsa nella quale «non c'è completezza ma solo elaborazione [221] ». L'informatica annulla il problema dell'eleborazione 'infinita' con l'apertura dei supporti digitali e, con portata veramente significativa, della pubblicazione on-line: il lavoro rimane sempre aperto per successive correzioni, modifiche e aggiornamenti.

Compilazione e pubblicazione non possono più essere considerate fasi realmente distinte, ma devono essere viste come un sistema solidale, dove i due momenti si compenetrano in un flusso di lavoro nuovo. Il *design* dei sistemi informativi per un dizionario elettronico deve farsi carico dell'analisi del flusso di lavoro per poi proseguire con una modellizzazione computazionale delle attività, progettarne e implementarne gli strumenti, compreso il *design* delle interfacce e curare la pubblicazione dei risultati per ogni fase conclusa. Operativamente se ne ricava un'ottimizzazione delle risorse, che possono essere realisticamente commisurate agli scopi per ogni dato momento di tempo, compatibile con l'atteggiamento di «filosofia del possibile, e della lessicografia come scienza pratica» descritta da Beltrami (2005: 317).

In questo punto, che un programmatore definirebbe come 'raccolta dei requisiti' e 'analisi', si profila un'interazione tra l'ingegneria del software e le scienze umane e induce, ancora una volta, un condizionamento forte, come si è visto nella lessicografia evolutiva del *DMF*. Oltre a reimpostare la pratica di ricerca, l'approccio evolutivo contribuisce senza dubbio a definire il concetto di dizionario elettronico; inoltre, se rapportato a un complesso disciplinare allargato, pare lecito affermare che in questo modo di procedere si trovano le peculiarità di un modo di

[221] Klein (2013:18).

affrontare l'attività di studio e ricerca che, con Burdick et al. (2014: 14), possiamo indicare col nome di "umanistica generativa", ovvero «una pratica che prevede cicli rapidi di prototipazione e analisi» perché cosciente che nell'esito dei nuovi processi produttivi digitali emergeranno di problematiche nuove — nell'accezione già riscontrata anche in Avalle (1985a: 80) e Mordenti (2001: 25-28). L'Umanistica Generativa non solo accetta la possibilità del fallimento nel processo produttivo, ma lo rivaluta come catalizzatore capace di innescare di un circolo virtuoso.

Riguardo alla concreta applicazione del paradigma evolutivo, si è visto che l'incarnazione più rappresentativa del nuovo metodo è il *DMF*, ma i segnali di una consapevolezza del nuovo modo di concepire il dizionario elettronico risalgono sicuramente a parecchi anni prima. Tra le prime manifestazioni di coscienza della possibilità di una nuova pianificazione del lavoro su un'impresa di grandi proporzioni si trova sicuramente quella della Oxford University Press nel periodo della preparazione della base di dati elettronica per l'*OED*, nella seconda metà degli anni '80[222]. Se la consapevolezza non basta a fare dell'*OED* un'impresa evolutiva, occorre anche ricordare che le prime voci del *TLIO*, della cui impostazione aperta si è già avuto modo di parlare, precedono il *DMF1* di parecchi anni[223].

Per avvicinarci ancora alle nostre finalità, il Progetto Artesia non è, per forza di cose, estraneo a un simile atteggiamento e le operazioni di bilanciamento effettuate sul *Corpus* sono il risultato reso possibile da una volontà programmatica[224] di *design* ciclico, per successive approssimazioni, in accordo con Biber (1993) e

[222] Lee Berg / Gonnet / Tompa (1990-1991).

[223] La prima voce del TLIO vede la luce il 14 gennaio 1996 (Squillacioti 2011: 1).

[224] Arcidiacono (2011: 291).

Atkins / Clear / Olster (1991: 5). I principi evolutivi saranno declinati in termini progettuali per il progetto del *VSM* nella sezione 6.2, ma alcune annotazioni sulle pratiche di aggiornamento invece saranno riportate nel prossimo paragrafo.

3.5 **ACCESSIBILITÀ** ON-LINE

> Immagino che qualcuno potrebbe dire: "Perché non mi lasciate da solo? Non voglio far parte della vostra Internet, della vostra civiltà tecnologica, o della vostra società in rete! Voglio solo vivere la mia vita!" Bene, se questa è la vostra posizione, ho delle brutte notizie per voi. Se non vi occuperete delle reti, in ogni caso saranno le reti ad occuparsi di voi. Se avete intenzione di vivere nella società, in questa epoca e in questo posto, dovrete fare i conti con la società in rete. Perché viviamo nella Galassia[225] Internet[226].

Se, appena pochi anni fa, si poteva riscontrare il pieno dominio del mercato dei dizionari su CD-ROM[227] — primo supporto di successo commerciale che ha storicamente[228]

---

[225] La metafora galattica ritorna diffusamente e su varie scale sulla questione di cambio di paradigma. Tra la Galassia Gutenberg di McLuhan (1962) e la Galassia Internet di Castells (2002) va doverosamente citata la Galassia Von Neuman di Gigliozzi (1997: 150-159) e Gigliozzi (2003).

[226] Castells (2002: 262).

[227] Solo pochi anni fa, per es., Chiari (2007a: 84) poteva affermare: «Attualmente tutti i principali dizionari vengono distribuiti con una versione elettronica su CD-ROM, pochi altri forniscono ulteriori possibilità di consultazione online».

[228] In Italia il primato della pubblicazione di un dizionario commerciale su CD-ROM va alla Zanichelli, con l'edizione del CD-ROM Multilingue nel 1987 (Cfr. Marello 1996: 157-158).

consentito l'accesso alla forma elettronica a un vasto pubblico[229] — oggi si sta per completare un processo di inversione obbligato[230] in favore dei dizionari on-line, spinta anche da una tendenza generale che preme verso la progressiva scomparsa delle memorie ottiche[231].

Il vocabolario presuppone un cospicuo volume di dati anche sul prodotto finito. Nell'era pre-elettronica, a fronte di vocabolari

---

[229] Grainger (2012: 2-3).

[230] Si confronti il già citato Chiari (2007a: 84) con Chiari (2012: 98).

[231] I più recenti modelli di computer portatili, per esempio, sono privi di lettori di CD-ROM e DVD ROM. Anche nelle opere commerciali — dove la versione elettronica del vocabolario è il complemento, ormai necessario, del volume e dove è sentito di più il bisogno di un oggetto fisico per chiari motivi editoriali — il supporto ottico non può più essere l'unico veicolo di distribuzione. Per ricollegarci alla nota precedente, nello stesso anno Chiari (2007b) analizzava l'impiego di alcuni dizionari elettronici su CD-ROM in glottodidattica il cui elenco può essere utilizzato per un confronto. *ZNG* fornisce una licenza annuale on-line valida per 365 giorni dall'attivazione e una *app* per Android (venduta separatamente). *DO*, che ha significativamente modificato il titolo in *Devoto Oli-Digitale*, fornisce un codice per scaricare e installare sul proprio computer il software e 12 mesi di accesso on-line; il *DISC*, fermo all'edizione 2008 è uno dei dizionari messi a disposizione sul sito del Corriere della Sera; DZ concede due anni di accesso alla versione on-line; Il *GRADIT*, invece, opera che si colloca su una fascia editorale (e di prezzo) diversa, è stato l'unico a passare dal CD-ROM a una chiavetta USB dal design ricercato, con una *docking station* che consente di esporre il supporto nella libreria o sulla scrivania. *DMP* è fuori commercio dal 2009 ma prima di andare fuori catalogo è stato per un periodo accessibile on-line sul sito dell'editore; cercando in rete, si trovano ancora molti messaggi di protesta per l'interruzione del servizio, mentre non si trovano critiche altrettanto feroci per l'uscita dal commercio del volume (un resoconto su <http://archivio.panorama.it/cultura/libri/Kindle-e-alle-porte-Ma-il-dizionario-De-Mauro-scompare-dal-web>). Oggi il dizionario è ospitato da *Internazionale* all'indirizzo <http://dizionario.internazionale.it/>, ultima consultazione: 23 ottobre 2015.

sempre più ricchi, ma sempre più voluminosi, anche il mercato editoriale doveva in qualche modo manifestare segni di insofferenza. Così, nel 1979, la Oxford University Press tentò una goffa compressione dei 20 volumi dell'*OED* nei due tomi dell'*OED Compact* — in un formato di stampa che faceva corrispondere a nove pagine dell'originale una pagina della versione ridotta — e corredando il cofanetto con un apposito alloggiamento con una lente di ingrandimento per permetterne la lettura. Il primo macroscopico effetto dei supporti elettronici, anche qui, è stato il pieno compimento della desiderata dematerializzazione

Spesso l'abbattimento delle barriere fisiche e delle limitazioni editoriali è stato genericamente salutato come un affrancamento da qualunque costrizione, un sentiero verso l'illimitata espansione informativa dei contenuti del vocabolario. Indubbiamente l'informatica ha cancellato l'angoscia della compressione dell'informazione negli esigui spazi previsti — «one of the main frustrations of dictionary professionals[232]» — e ha smorzato l'esigenza di utilizzare numerose abbreviazioni. Ci si è presto resi conto che la dematerializzazione non può promuovere la sovrabbondanza informativa nelle voci: la quantità di informazione va commisurata ai bisogni dell'utente e non alle possibilità del mezzo. Oggi, con gli *smatphone*, l'*ubiquitous computing*, la domotica e la *wearable computing*[233], sono inoltre gli stessi dispositivi, dagli schermi sempre più piccoli, a incoraggiare un corretto e misurato dosaggio dell'informazione all'interno delle schermate[234].

Ma ancora una volta il pieno compimento del processo non si ferma alla dematerializzazione. Nell'evoluzione storica del

[232] Granger (2012: 3).

[233] Cfr. Colistra (2008).

[234] Ibidem.

dizionario elettronico la pubblicazione on-line segna un punto di svolta irreversibile e differente da quello dei supporti elettronici tradizionali. Per Fuertes-Olivera / Bergengenholtz (2011: 1) il dizionario elettronico on-line è figlio della società della conoscenza e dell'informazione e richiede un approccio differente rispetto alla lessicografia elettronica del primo periodo[235]. Altri studiosi, come Gouws (2011: 17), limitano l'applicazione del termine 'dizionario elettronico' ai soli dizionari on-line, escludendo quelli pubblicati su CD-ROM[236].

La pubblicazione on-line consente di raggiungere vaste fette di pubblico e di abbattere i costi editoriali, è aperta alla multimedialità, all'ipertestualità, al dinamismo informativo delle voci ma ha il vantaggio del pieno controllo dell'editore in qualunque momento e dell'aggiornamento continuo dei contenuti. Per questo, come già riscontrato, oggi si possono cominciare a pubblicare le prime voci disponibili (anche in forma parziale), senza attendere che l'intero vocabolario sia completato.

Sebbene la pubblicazione in tempo reale sia possibile, è pratica comune procedere per aggiornamenti a intervalli di tempo regolari che interessano interi gruppi di voci (ad es. il *TLIO* procede per aggiornamenti bimestrali, l'*OED* per aggiornamenti trimestrali[237], ma, come già detto, si sta valutando l'ipotesi di

---

[235] Fuertes-Olivera / Bergenholtz (2011: 1) portano come esempio la capacità che hanno solo i dizionari on-line di collegamento con risorse extra-lessicografiche.

[236] Accusati anche di essere, nella maggior parte dei casi, semplici riedizioni di dizionari cartacei su supporto digitale (Gouws 2011: 17-18).

[237] Le notivtà dell'ultimo aggiornamento sono disponibili all'indirizzo <http://public.oed.com/the-oed-today/recent-updates-to-the-oed/>, ultima consultazione: 19 ottobre 2015. Dallo stesso link è possibile accedere alla pagina con i precedenti aggiornamenti.

pubblicare le bozze delle voci prima che queste siano complete[238]): una soluzione a cadenze regolari è preferibile, poiché un oggetto in continuo cambiamento è una referenza "debole", una fonte mutevole da cui derivano citazioni difficilmente verificabili[239].

Per lo stesso motivo, il *DMF*, che viene aggiornato per versioni successive, mantiene un archivio[240] con tutte le precedenti edizioni [241] . L'*OED* on-line, invece, permette il controllo di eventuali versioni obsolete della voce e ne consente la visualizzazione dinamica all'interno della stessa pagina.

I primi risultati diventeranno immediatamente visibili; il rischio di versioni rese inutilizzabili per l'obsolescenza dei supporti viene eliminato o circoscritto a rare incompatibilità con i browser[242]; le novità nell'infrastruttura informatica miglioreranno l'accesso ai contenuti senza generare prodotti editoriali incompatibili coi nuovi sistemi operativi e dispositivi; i lettori disporranno sempre di una versione aggiornata e funzionale. Sembrerebbe tutto perfetto, ma occorre considerare anche un altro lato della medaglia.

> Finally, it is important to remember that an online dictionary only exists for as long as its server is running (if the server

---

[238] Simpson (2013: 37).

[239] Ad affermarlo è lo stesso Martin (2008: 1251).

[240] <http://atilf.atilf.fr/scripts/dmfX_2012.exe?INIT_SESSION;CRITERE=ACCUEIL;ISIS=isis_dmf2010.txt;OUVRIR_MENU=1;ISIS=isis_dmf.txt;OUVRIR_MENU=0;s=s100a05a8;ISIS=isis_dmf.txt>. Ultima consultazione: 18 ottobre 2015.

[241] Vengono ripresentati i contenuti delle precedenti versioni ma l'interfaccia viene sempre aggiornata.

[242] Tendenzialmente i browser tendono ad essere retrocompatibili con le vecchie soluzioni tecniche (tranne casi eclatanti come la vicenda, non ancora conclusa, della compatibilità con Macromedia/Adobe Flash).

> goes down, the dictionary disappears); threrefore taking all the above into consideration, the continued existence of the publishing institution becomes essential. (Beltrami 2013b: 630).

La codifica in XML/TEI, che verrà affrontata più avanti, è il miglior modo attualmente disponibile per garantire la portabilità tecnica dei dati nel tempo e sconfiggere l'obsolescenza del software, ma il sistema centralizzato di distribuzione basato su *web-server* lascia i dati solo nelle mani dell'editore. Tutti i progetti analizzati basati su XML non danno accesso al sorgente, eccetto il *Vocabolario degli Accademici della Crusca*, che tuttavia non rimanda al file XML ma mostra la marcatura all'interno della pagina HTML. La pubblicazione esclusiva tramite sito web, a meno che non si dia la possibilità all'utente di scaricare l'archivio come fa per esempio Wikipedia[243], è garanzia di accesso non di consegna dei dati ai lettori.

3.6 LESSICOGRAFIA STORICA *CORPUS-BASED* E *CORPUS-DRIVEN*

Un corpus è una «raccolta di dati linguistici che possono costituire la base empirica per l'analisi di una lingua naturale»[244]. La pratica di utilizzare corpora per gli studi linguistici è ben più antica del calcolatore, ma è grazie alla capacità dei computer di trattare grosse collezioni di dati che, dalla seconda metà del XX secolo, il settore ha manifestato una notevole vivacità, espressa

[243] L'enciclopedia libera rende disponibile il *dump* xml del database: <https://it.wikipedia.org/wiki/Aiuto:Download_di_Wikipedia>, ultima consultazione: 1 dicembre 2015.

[244] Beccaria (1994: 191).

anche dal costituirsi di un settore di studi dalla spiccata specificità, la linguistica dei corpora[245]. Storicamente la linguistica dei corpora si sviluppa in serrata polemica col generativismo[246] e professa il rifiuto di ricorrere all'interrogazione dei parlanti e all'introspezione.

I corpora inoltre sono essenziali nella lessicografia storico-etimologica, per almeno altri due motivi: la necessità di disporre agevolmente di grandi quantità di dati[247]; per la possibilità che offrono di poter studiare stati di lingua non più accessibili:

> even relatively recent stages of language are inaccessible to introspection or elicitation […] The inevitability of external sources and indirect evidence in the case of remote stages of language(s) makes corpus linguistics a most suitable — one would almost claim: the only suitable — methodological option for diachronic studies[248].

Allo stesso tempo queste condizioni tracciano inevitabilmente un limite oltre al quale alla lessicografia storica non è dato di andare:

---

[245] Per una trattazione introduttiva si rimanda a Spina (2001) o al più recente Barbera (2013).

[246] «La nozione di c[orpus], nei suoi attributi di concretezza e finitezza […] è stata spesso criticata dalla grammatica generativa […] in quanto considerata troppo dipendente dall'esecuzione della lingua, e pertanto incapace di renderne l'aspetto creativo» (Beccaria 1994 s.v. *corpus*). Celebre l'attacco di Chomsky (1962), cit. in McEnery / Wilson (2001: 10): «Any natural corpus will be skewed. Some sentences won't occur because they are obvious, others because they are false, still others because they are impolite. The corpus, if natural, will be so wildly skewed that the description would be no more than a mere list».

[247] Gleßgen (2006: 17).

[248] Pusch / Kabatek / Raible (2005b: 1-2).

> Sono necessariamente analisi ristrette a un c[orpus] chiuso[249] quelle relative ad una lingua morta, in cui il c[orpus] di dati è rappresentato da una collezione finita di testi, né è possibile il ricorso al parlante nativo per ampliare la base di dati o per vagliare ipotesi interpretative[250].

Il ruolo dei corpora in lessicografia si articola su un doppio livello: «as raw material which lexicographers mine and refine to produce rich lexical entries, and downstream, as an integral part of the electronic dictionary to which users have direct access and which they can mine for themselves[251]». Nel dizionario elettronico esiste poi la possibilità concreta per l'utente, più volte riscontrata nei casi di studio del capitolo precedente, di collegare l'analisi lessicografica all'osservazione della parola nei suoi contesti naturali di occorrenza, anche mediante collegamento ipertestuale diretto.

Se con i corpora elettronici è possibile un trattamento più potente ed economico, questo riduce gli oneri di un'analisi lessicale condotta da zero, senza ricorrere alla tradizione lessicografica precedente. Si può parlare di dizionari "di prima mano[252]" quando l'intera catena di analisi e operazioni messe in atto dal lessicografo muove ed è circoscritta alla base documentaria iniziale, senza ricorrere ad altri lessici. Questa scelta metodologica si avvicina ai concetti di analisi *corpus-based* e *corpus-driven*[253], stabilita con Tognini-Bonelli (2001)[254]:

---

[249] Sulla nozione di *corpus* si tornerà nel paragrafo 5.3.3.

[250] Beccaria (1994 s.v. *Corpus*).

[251] Granger (2012: 3).

[252] Cfr. per es. Beltrami (2008: 34) o Beltrami (2009: 50).

[253] Sulla lessicografia corpus-based si veda Hank (2012). Per un tentativo di studio comparativo su un insieme di parole tra approccio corpus-based e

> The term corpus-based is used to refer to a methodology that avails itself of the corpus mainly to expound, test, or exemplify theories and descriptions that were formulated before large corpora became available to inform language study[255]

Per l'approccio *corpus-driven*, invece «il linguista non può e non deve affermare altro che non sia desunto da un corpus»[256]:

> In a corpus-driven approach the commitment of the linguist is to the integrity of the data as a whole, and descriptions aim to be comprehensive with respect to corpus evidence. The corpus, therefore, is seen as more than a repository of examples to back pre-existing theories or a probabilistic extension to an already well defined system. The theoretical statements are fully consistent with, and reflect directly, the evidence provided by the corpus.[257]

Nel *TLIO*, sebbene i lessici vengano usati per verifica o per documentazione, le voci che non trovano fondamento sulla base di dati vengono rigorosamente segnalate come "voci fuori corpus". Fatta salva quindi la giusta flessibilità di questo tipo di entrata, l'obbligo di segnalazione avvicina il *TLIO* al paradigma *corpus-*

---

intuition-based si veda Verlinde / Thierry (2001). Sulla distinzione tra studi linguistici corpus-based e corpus-driven si veda Tognini-Bonelli 2001 e il precedente Tognini / Bonelli (1996).

[254] Anticipata nel precedente Tognini-Bonelli (1996). Barbera (2013: 15) retrodata l'avvio effettivo della linea *corpus-based* con Fillmore (1992).

[255] Tognini-Bonelli (1992: 65).

[256] Barbera (2013: 16).

[257] Ivi, p.84.

driven[258]. Mentre, per quanto l'*AND* condivida la sfiducia nelle fonti di seconda mano, la sua metodologia non arriva a questi livelli di rigore: «for a variety of reasons, not least a well-founded skepticism about the reliability of glossaries, the *AND* has not usually made direct use of glossaries, and certainly does not rely on them». L'avverbio «usually» rimanda a un atteggiamento *corpus based.*

Un dizionario di prima mano è, almeno in via di principio, immune da tutti quegli errori che si trascinano per inerzia da un vocabolario all'altro. L'errore lessicografico, del resto, ha modi e forme di propagazione proprie che lo trasmettono da dizionario a dizionario, fino a inquinare tutta la conoscenza e gli studi che attingono alle informazioni delle voci scorrette. Il fisico sa che l'errore iniziale è un elemento ineliminabile, già nelle diverse forme (sistematiche o accidentali) relative alle misurazioni che forniscono le informazioni di partenza per la sperimentazione[259]. Nella lessicografia storica, la presenza dell'errore iniziale potrebbe essere considerata altrettanto ineliminabile, salvo benaccette ma improbabili dimostrazioni del contrario.

---

[258] «Lemmi documentati non dal corpus, ma da altre fonti (lessici, strumenti lessicografici e altra bibliografia), si redigono voci 'fuori corpus' in cui questo fatto è sempre dichiarato ed eventualmente discusso», Beltrami (2013a: 2).

[259] La scienza dell'ultimo mezzo secolo ha dimostrato un interesse crescente nell'individuare l'entità e gli effetti imprevedibili che l'errore iniziale può indurre, soprattutto in tutti quei sistemi caratterizzati dalla cosiddetta «dipendenza sensibile dalle condizioni iniziali», più nota, grazie a Edward Lorenz, nell'immaginario collettivo come "effetto farfalla". Per una rapidissima introduzione si rimanda Bonavita (2008) o, per un volume dal taglio più divulgativo a Gleick (1987). Un informatico non dovrebbe dimenticare che la questione era stata anticipata da Turing (1950).

Limitandoci quindi all'errore iniziale, in fase di pre-redazione si dovrà prestare particolare attenzione a tre attività fondamentali:

— Arginare la propagazione degli errori da fonti secondarie.

— Adottare un approccio filologico per le fonti primarie. Più in generale una buona dose di cautela e attenzione dovrebbe sempre precedere e assistere l'allestimento del corpus[260].

— Evitare che il trattamento automatico dei dati testuali generi anomalie, distorsioni o amplificazioni dell'errore.

Per il primo punto, come si è detto, il vocabolario di prima mano evita che l'errore si riproponga ereditariamente.

Sul punto due, il caso, diventato emblematico con Orlandi (2004), della pessima edizione elettronica sul portale Icon[261] del *Jaufré Rudel* del Carducci avverte come il problema potrebbe spostarsi, nelle stesse forme, dal vocabolario alla banca dati. Il testo è stato digitalizzato originariamente da due grandi imprese: la *LIZ*, prima in ordine di tempo, e *Biblioteca digitale della letteratura italiana* [262] , che raccoglie l'eredità della *LIE*; successivamente gli errori sono stati riproposti in numerosi archivi (tra cui Biblioteca Italiana[263] e Liber Liber[264]) che hanno attinto dalle collezioni di testi delle due iniziative.

---

[260] Il principio interessa tutte le discipline che fanno uso di corpora elettronici e testi codificati. I richiami in letteratura si sprecano, così come le segnalazioni di progetti che conducono le attività trascurando precauzioni che dovrebbero essere basilari. Sul rapporto tra filologia e base documentaria, sul quale conviene sorvolare per motivi di spazio, si vendano almeno Nencioni (1961), Varvaro (1998), Beltrami (2010a e 2010b), Formentin (2014).

[261] <http://www.italicon.it/>. Ultima consultazione: 2 dicembre 2015.

[262] <http://www.letteraturaitaliana.net/>. Ultima consultazione: 2 dicembre 2015.

[263] <http://www.bibliotecaitaliana.it/>. Ultima consultazione: 2 dicembre 2015.

Per l'ultimo punto, oltre a demandare l'introduzione di distorsioni accidentali a una buona programmazione, ci si dovrebbe chiedere se l'aspetto quantitativo della tecnologia sia, di per sé, un fattore di rischio come per Formentin (2013: 195):

> nell'era dell'acquisizione dei testi in grandi banche dati interrogabili in rete, il nostro fantasma diventa subito patrimonio comune, sicché l'errore si moltiplica per *n* volte e rischia di ripercuotersi a diversi livelli della ricerca scientifica.

Il principio sembra dipendere (e in modo proporzionale) dal numero di consultazioni più che dalla tecnologia digitale, per quanto sia intrinseco ai nuovi mezzi di comunicazione l'aumento e l'accelerazione delle connessioni e il riuso e reimpiego[265] delle risorse testuali. Inoltre, le banche dati elettroniche, come si precisa anche nella citazione, si contraddistinguono per le dimensioni generose che possono influire almeno in tre modi:

- richiedono tecniche di digitalizzazione di massa (OCR, lemmatizzazioni automatiche, ecc.), le quali presentano una percentuale sistematica di errore;

- maggiori dimensioni significano minor controllo perché l'accesso a grandi quantitativi di informazione spesso è mediato dalla disgregazione e ricombinazione dei dati (dagli indici alle analisi quantitative alle tecniche di *data mining*) che rendono invisibile l'errore nell'*output*;

---

[264] <http://www.liberliber.it/>. Ultima consultazione: 2 dicembre 2015.

[265] Nel corso di rapide ricognizioni sulle opere più rilevanti non è stato rilevato un uso particolare di strumenti appositi, tranne alcune piattaforme con *widget* di condivisione come per *Il Nuovo De Mauro* <http://dizionario.internazionale.it>.

- le grandi dimensioni tendono a centralizzare e riutilizzare le fonti documentarie su cui vengono sviluppati più progetti.

Formentin (2014: 197) intravede un rimedio nell'apertura del mezzo, già ampiamente trattata al §3.4, la quale però andrebbe accompagnata da una maggior sensibilità da parte di chi gestisce le banche dati. Nel caso del Jaufré Rudel — tralasciando la poca chiarezza sull'edizione di riferimento riprodotta dagli archivi e perdonando gli errori alle pioneristiche *LIZ* e *LIE* — è ingiustificabile è che, ad oggi[266] nonostante l'intervento di Orlandi, nessuna delle banche dati citate abbia rivisto la propria edizione.

L'antidoto tecnico è inutile se ci si rifiuta di assumerlo.

[266] 2 giugno 2015.

4 LA CODIFICA DEL DIZIONARIO

4.1 IL PROBLEMA DELLA CODIFICA

Trasportare la testualità sul calcolatore significa rappresentarlo in formato digitale, o *Machine Readable Form* (*MRF*), «termine ormai desueto ma sempre valido nella sua connotazione concettuale»[267]. La codifica, passaggio obbligato per qualunque applicazione informatica nell'universo delle lettere, è il «momento iniziale (ma centrale) di qualsiasi indagine»[268]; non procedura tecnica ma «tematica filologica e semiotica, come lo sono l'elaborazione della scrittura o quella della stampa»[269], ha alimentato negli anni una tradizione di studi nutrita e ha maturato una sua autonomia:

> Questa operazione, preliminare a qualsiasi trattamento, può assumere una sua dignità autonoma se la si considera nel suo aspetto conservativo o qualora si decida di produrre un considerevole numero di esemplari dei materiali messi in memoria[270].

La codifica iniziale determina gli esiti e le possibilità nella costituzione del dizionario e del corpus su almeno due assi:

---

[267] Orlandi (2010: 3).

[268] Gigliozzi (1987: 67).

[269] Gleßgen (2006: 15).

[270] Gigliozzi (2003: 63).

- Seleziona le informazioni che vengono ricodificate in digitale e definisce l'insieme delle future operazioni di interrogazione possibili.
- Favorisce l'applicazione di trattamenti algoritmici diversi, e in un contesto applicativo, influenza il *design* del software. In altre parole, «il modo in cui sono rappresentate le informazioni ha un impatto sulla complessità e sull'efficienza delle applicazioni che devono gestirle»[271].
- La peculiarità dei dati contenuti in un vocabolario, come visto, è di avere una strutturazione forte. La corretta progettazione diventa prerequisito per una rappresentazione valida e per un accesso ottimale ai dati.

Prima di convertire qualunque documento si dovrà però definire lo sfuggente concetto di testo e individuarne le caratteristiche da rappresentare, perché attraverso queste caratteristiche — che non sempre possono essere ridotte alla semplice successione di lettere, segni di punteggiatura e diacritici — si determina l'informazione da rappresentare. Senza voler affrontare una trattazione dettagliata[272], ci si limiterà a ricordare che l'individuazione dell'oggetto testo non è un processo oggettivo e meccanico ma frutto di un processo di interpretazione, tuttavia nel nostro caso il compito è semplificato dall'articolazione evidente della microstruttura dei vocabolari e dai modelli di riferimento adottati.

Quando si parla di codifica ci troviamo di fronte a due livelli di astrazione sovrapposti gerarchicamente: una codifica di base, che crea una corrispondenza tra caratteri e codice binario, e una codifica di alto livello che arricchisce il testo codificato al livello

[271] Braga / Campi / Ceri (2005: 45).

[272] Si vedano almeno Segre (1981) e Segre (1985: 28-29), Orlandi (2010: 3-25), Ciotti (1995: 149-158),

zero con informazione alla struttura del testo e alla sua formattazione.

4.1.1 La codifica di basso livello (*plain text)*

I dispositivi digitali, come ben noto, memorizzano e trattano i dati scomponendoli in unità minime di informazione, i *bit*, che possono assumere solo due valori o configurazioni: 0 e 1. Al livello più basso, o livello zero della codifica, la rappresentazione di un linguaggio alfabetico ha quindi come fine di trasformare le sequenze di caratteri del testo in sequenze di *bit*.

La messa in codice si realizza istituendo una corrispondenza biunivoca tra i simboli di un determinato sistema di scrittura (il *character set*) con una sequenza di valori binari (il *code set*). In questo contesto un carattere è inteso come entità astratta, distinta dalla possibile rappresentazione grafica (glifo)[273]. I due *set* e le relazioni che li collegano definiscono un *coded charachter set* — tabella di codici o, informalmente, codice — un ponte ideale che collega l'universo delle lettere ai segnali elettrici che corrono dentro i componenti elettronici del calcolatore.

> Le tabelle di codici — qualunque esse siano — che ci permettono di compiere quella che possiamo definire l'operazione "nucleare" per l'utilizzazione dell'informatica nel settore umanistico (ma ovviamente anche in tutti gli altri campi), e cioè la scrittura, rappresentano l'elemento minimo, ma non sufficiente per la costruzione "modello del testo" che vogliamo creare e al quale rivolgiamo le nostre domande per avere risposte sul testo stesso.[274]

[273] Lenci / Montemagni / Pirrelli (2005: 57).

[274] Gigliozzi (1997: 103).

C'è una sproporzione aritmetica tra l'unità minima del testo e quella del digitale: con un *bit* possiamo rappresentare solo 2 valori (0 e 1), mentre un singolo carattere può assumere 26 valori, a voler considerare solamente i grafi dell'alfabeto latino di base — ai quali andrebbero aggiunte le maiuscole, le lettere accentate, i numeri, lettere con diacritici particolari e così via — in un insieme che può aumentare indefinitamente con l'introduzione di caratteri di altri sistemi di scrittura.

Per sopperire a questo dislivello, si devono impiegare più cifre binarie per codificare un solo grafema. Per semplificare, si immagini di dover associare a ognuna delle 26 lettere dell'alfabeto un simbolo arbitrario preso da un insieme di altri 26 elementi: in questo caso a ogni lettera corrisponderà un solo simbolo. Si provi adesso ad associare le stesse lettere a un numero, partendo da [spazio] = 0, *a*=1, *b*=2, ecc. Arrivati alla *j* si userà istintivamente il 10, composto stavolta da due cifre decimali, e questo perché una cifra decimale riesce a rappresentare un massimo di 10 valori (da 0 a 9), meno di quelli richiesti dal primo insieme[275].

Come apparirà chiaramente nella tabella ASCII parzialmente riportata nel prossimo paragrafo, più valori si useranno nel *code set*, più la tabella conterrà punti (*code points*) per rappresentare i caratteri da codificare.

Oltre a configurarsi come il momento primo e fondamentale per la costruzione del testo elettronico, la codifica di basso livello è quella su cui si riscontra il massimo livello di *portabilità*, proprietà del testo codificato che ne definisce il grado di compatibilità nel trasporto tra ambienti hardware e software diversi (portabilità orizzontale) e trasportabilità garantita nel tempo (portabilità verticale). Come risaputo, i più noti formati di uso quotidiano per i documenti di testo, come il .doc e .docx di Microsoft Word o il

[275] Cosa rappresenterà questo valore? *J* o [spazio] + *a*?

PDF sviluppato da Adobe, non rispondono a questi requisiti: la codifica del testo è combinata con le informazioni sulla formattazione e la composizione tipografica, rendendo impraticabile la diretta e libera elaborazione algoritmica del testo; si tratta inoltre di formati proprietari, la cui permanenza nel tempo è pericolosamente minacciata dai destini che interesseranno le aziende che ne detengono o diritti[276].

4.1.2 Dalla tabella ASCII allo *standard* ISO-8859

La tabella che ha avuto più successo nella storia, è il celeberrimo codice ASCII (American *standard* Code for Information Interchange), «primo codice *standard* per la rappresentazione binaria dei caratteri[277]». Nel codice ASCII a ogni simbolo alfabetico è riservato un *byte*, composto da 8 *bit* di cui l'ultimo riservato per verificare la correttezza della trasmissione[278] (per un totale di 128 combinazioni diverse). Considerando che i primi caratteri minuscoli dell'alfabeto cominciano ad apparire dalla posizione 97[279] e che il numero della posizione corrisponde al valore binario del *code set,* utilizzando questa tabella si avrà:

---

[276] «Con tecnologia "proprietaria" si intende una tecnologia posseduta in esclusiva da un soggetto (ad esempio una ditta o casa produttrice) che ne detiene il controllo mantenendo segreto il funzionamento e che può modificarla a sua discrezione senza doverne rendere conto. Con formato "proprietario" ci si riferisce a un formato le cui specifiche non sono diffuse pubblicamente in modo esplicito e che la casa produttrice si riserva il diritto di cambiare a sua discrezione», (Lenci / Montemagni / Pirrelli 2005: 68).

[277] Lenci / Montemagni / Pirrelli (2005: 58).

[278] In termini tecnici, *bit* di parità.

[279] Le prime 33 posizioni sono riservati a caratteri di controllo (compreso lo spazio, l'accapo e la tabulazione), poi seguono alcuni segni di punteggiatura, quindi i numeri e le lettere maiuscole.

97	01100001	a	109	01101101	m
98	01100010	b	110	01101110	n
99	01100011	c	111	01101111	o
100	01100100	d	112	01110000	p
101	01100101	e	113	01110001	q
102	01100110	f	114	01110010	r
103	01100111	g	115	01110011	s
104	01101000	h	116	01110100	t
105	01101001	i	117	01110101	u
106	01101010	j	118	01110110	v
107	01101011	k	119	01110111	w
108	01101100	l	120	01111000	x

Dall'esempio è possibile apprezzare la natura prettamente meccanica e convenzionale della codifica. Concatenando la sequenza di *bit* di ciascuna parola si otterranno sequenze di lettere, parole, frasi, ecc.

Per motivi storici e commerciali questo codice ha un *character set* basato sull'alfabeto delle lingue anglosassoni, ed è quindi privo dei caratteri con diacritici, a partire dalle lettere accentate. Le sue potenzialità rappresentative sono dunque gravemente pregiudicate anche sull'insieme degli alfabeti latini e, ovviamente, totalmente inadeguate per gli alfabeti non latini. Per ovviare a questo problema il codice ASCII è stato progressivamente esteso aumentando la lunghezza delle stringhe binarie del *code set*.

Le prime estensioni dell'ASCII portarono la codifica da 128 a 256 punti di codice, mantenendo inalterate le prime posizioni della tabella. In altre parole, tutti i codici derivati da ASCII ne includono le prime 127 posizioni. Per i rimanenti punti, invece, sono state formulate diverse associazioni convenzionali che a volte cambiano anche a seconda del sistema operativo utilizzato. Con la

"famiglia" ISO-8859 sono state create diverse specifiche per favorire la standardizzazione anche al di fuori dell'Europa occidentale. Ogni tabella della famiglia coincide ed è compatibile con ASCII e con le altre famiglie nei primi 128 punti e riserva le restanti posizioni per i caratteri dei diversi sistemi grafici.

Il sistema ISO-8859 si scontra, però, con due problemi fondamentali:

1. Copertura ristretta: lo *standard* non prevede una copertura per i paesi dell'Asia dell'est (Corea, Giappone, Cina). La scrittura ideografica richiede molti più punti rispetto a quelli previsti dal formato di codifica.
2. Le diverse famiglie costituiscono sistemi *mutuamente esclusivi*[280]: sui punti da 128 in poi, i diversi *standard* codificano glifi distinti in presenza dello stesso valore binario. Lo stesso valore, in sostanza, verrà interpretato diversamente a seconda dello *standard* impostato al momento. Un file, inoltre, dovrà usare lo stesso *standard* per tutto il suo contenuto, quindi, ad esempio, non si potrà passare dai caratteri latini al greco moderno all'interno dello stesso documento.

4.1.3 Unicode

La soluzione ai limiti della specifica ISO-8859 è stata offerta da Unicode[281], un recente *standard* universale per la codifica dei caratteri che si prefigge l'ambizioso obiettivo di rappresentare i caratteri di qualunque scrittura conosciuta. L'estensione del *coded character set*, nell'ultima versione, prevede 1.114.112 punti, la maggior parte dei quali è destinata a codificare caratteri. I glifi usati nelle principali lingue del mondo si collocano nelle prime

[280] Lenci / Montemagni / Pirrelli (2005: 60).

[281] <http://www.unicode.org>.

65.536 posizioni, chiamate anche *Basic Multilingual Plane* (BMP). Una dimensione di oltre un milione di punti permette anche la compresenza di tutti i sistemi grafici all'interno della stessa tabella, e risolve quindi anche i problemi di incompatibilità.

L'estensione di Unicode candida questo *standard* a diventare il mezzo per una codifica universale:

> The overall capacity for more than 1 million characters is more than sufficient for all known character encoding requirements, including full coverage of all minority and historic scripts of the world[282].

L'attenzione alle lingue antiche ha aperto nuovi scenari per gli studi filologici e linguistici e ha subito attirato l'attenzione attiva dei medievisti. Ne sono nati progetti come la *Medieval Unicode Font Initiative*[283] e strumenti come Junicode[284], sviluppato da Peter Baker dell'Università della Virgina, un *font* che renderizza i caratteri usati nei manoscritti medievali europei codificati nella tabella Unicode.

Mentre ASCII associava a ogni carattere un numero, al quale corrispondeva esattamente una stringa di *bit*, Unicode associa un numero di tabella che astrae dalla rappresentazione binaria e non dice nulla su come questo debba essere codificato fisicamente sul disco (questa proprietà di Unicode viene spesso omessa). Per passare dal codice Unicode alla stringa binaria da scrivere sul disco o da caricare in memoria bisogna eseguire una trasformazione. Il componente incaricato a svolgere questo compito è chiamato *codec*. UTF-8 (Unicode Trasformation Format) uno dei più diffusi *codec* a

---

[282] Unicode *standard* V.8.0, p. 2.

[283] <http://folk.uib.no/hnooh/mufi/>

[284] Attualmente alla versione 0.7.7.

lunghezza variabile[285] — la stringa binaria che codifica un carattere non è fissa e un glifo può essere rappresentato con un minimo di un *byte* fino ad un massimo di quattro *byte* — riesce a rappresentare rappresentare qualunque carattere dello *standard* Unicode. Nella codifica del testo UTF-8 è importante per almeno due motivi:

- permette di rappresentare l'intero *set* dei caratteri Unicode (non tutti i *codec* sono capaci di dominarla per intero);
- come si vedrà meglio nel paragrafo 4.2, è usato come *codec* predefinito per i file XML e, di conseguenza, per i documenti codificati con lo *standard* della Text Encoding Initiative.

4.2 XML/TEI

Con Unicode si è raggiunto l'obiettivo della completa rappresentazione dei glifi ma la valenza informativa del testo — è una nozione immediatamente intuibile — non si limita alla sequenza di caratteri. Nel contenuto informativo di un testo dobbiamo annoverare tutti gli aspetti che si riferiscono alla presentazione visiva del documento (*font*, corsivi, grassetti, ecc.) e quelli concernenti l'organizzazione del testo (capitoli, paragrafi, titoli, ecc.). La codifica di basso livello riesce a rappresentare una sequenza continua di caratteri ma esclude le informazioni di ordine superiore. Come già anticipato, i comuni programmi di videoscrittura riescono a padroneggiare informazioni di livello più

[285] Se il primo *bit* del *byte* comincia per 0 allora si tratta di un carattere semplice che coincide con la tabella ASCII, in caso diverso si tratta di *byte* concatenati per creare caratteri meno frequenti. Se *byte* comincia per 110 o 1110 allora si tratta del primo *byte* di una sequenza; i *byte* successivi (da uno a tre) saranno riconoscibili perché cominceranno per 10.

alto ma le mischiano con il contenuto testuale, impediscono l'elaborazione automatica e comportano dei seri rischi per la permanenza dei dati nel futuro. Se ne ricava un'equazione per la quale a un grado elevato di portabilità corrisponde un grado ridotto di espressività[286].

Un linguaggio dichiarativo permette di risolvere il conflitto tra portabilità e rappresentatività, specificando, per ogni stringa di caratteri codificata, la collocazione dei segmenti testuali nella struttura del testo[287]. Per ottenere questo scopo si servono di un insieme di marcatori o *tag* e di un insieme di regole che ne definiscono il comportamento e le proprietà di combinazione.

Le preoccupazioni concernenti la portabilità sono superate costruendo gli stessi marcatori con speciali sequenze di caratteri codificati in *plain text*. In questo modo, a livello puramente formale, non c'è differenza tra un file Unicode e un file codificato attraverso un linguaggio dichiarativo. Un corollario diretto è che un file costituito da testo e marcatori codificati in stringhe testuali è leggibile sia dai software — e, cosa più importante, da applicazioni anche tecnologicamente distanti — sia dall'occhio umano.

---

[286] Lenci / Montemagni / Pirrelli (2005: 68).

[287] Prima di codificare le informazioni più complesse, si dovrà però scegliere quali caratteristiche rappresentare e definire il livello di dettaglio che si intende mantenere nella riproduzione testuale e linguistica. All'individuazione e alla marcatura degli elementi microstrutturali del vocabolario sarà dedicato il cap. 6.3. Per una trattazione generale si rimanda a Lenci / Montemagni / Pirrelli (2005: 65-66) o Gigliozzi (2003: 67- 73). Un ottimo punto di partenza rimane l'*Introduzione* di Fabio Ciotti a Burnard / Sperberg-McQueen (2005), anche se il manuale nel suo complesso è ormai obsoleto perché descrive una vecchia versione dello schema (TEI P4).

La tecnologia oggi più diffusa per la marcatura dichiarativa è l'Extensible Markup Language[288] — più noto come XML — un sistema per la rappresentazione di informazioni testuali semi-strutturate.
Sebbene a volte ci si riferisca a XML parlando di linguaggio, si tratta, più propriamente, di un metalinguaggio, ovvero di un insieme di regole di base attraverso le quali è possibile definire infiniti linguaggi per la codifica di informazioni semistrutturate.

Nato nel 1998 come evoluzione di SGML[289], ha avuto un notevole successo, soprattutto nello sviluppo delle tecnologie legate al web[290].
Un file XML è essenzialmente un file di testo semplice, per *default* codificato in UTF-8 a lunghezza variabile, nell'ambito quindi della tabella Unicode[291]. Il formato ha un buon grado di compatibilità con la codifica ASCII di cui è un *superset*[292], per cui i file contenenti puro testo ASCII sono anche documenti UTF-8. Per la sua generalità XML è adatto all'impiego in più contesti, non a caso è stato sviluppato per favorire la condivisione di dati attraverso la rete, e ha trovato applicazione in moltissimi contesti differenti. La

[288] Per una guida di riferimento si veda Harold / Scott (2001: 26). Per informazioni più aggiornate si faccia riferimento alla pagina ufficiale: < http://www.w3.org/XML/>. Ultima consultazione: 20 ottobre 2015.

[289] Per un'introduzione alla codifica SGML/TEI, la cui utilità si riduce oggi alla documentazione si una fase superata nella storia della disciplina, si veda Gigliozzi (1997: 109-116).

[290] XML è uno *standard* del W3C (www.w3.org);

[291] Cfr. §4.1.3.

[292] Uno schema più ampio che mantiene al suo interno le posizioni ASCII.

sua semplicità favorisce lo sviluppo di applicazioni che riescono ad analizzarne i file derivati e a manipolarli[293].

L'adozione di Unicode ne incoraggia l'impiego in applicazioni internazionalizzate che richiedono sistemi grafici separati e il trasferimento di informazioni su sistemi operativi diversi.
La preferenza di XML per Unicode non vincola il metalinguaggio alla codifica di basso livello: XML può gestire un gran numero di *coded character set* diversi. Qualora si desideri utilizzare un *set* di caratteri diverso da UTF-8, basterà semplicemente specificare la codifica scelta all'inizio del documento[294].

Il documento XML è in sostanza un file all'interno del quale il testo è strutturato e annotato attraverso *tag*, chiamati anche marcatori, costituiti a loro volta da stringhe di caratteri alfanumerici chiusi all'interno di parentesi angolari. La marcatura di un paragrafo sarà quindi rappresentata per mezzo della stringa " " per l'apertura della sezione e " " per la chiusura, dove lo *slash* premesso al nome del *tag* indica la chiusura della porzione di testo interessata.

I *tag* possono essere annidati e devono essere inseriti correttamente l'uno dentro l'altro. Ogni *tag* può contenere anche degli attributi con un valore che va inserito all'interno di virgolette. Per esempio, se volessimo marcare un forestierismo con un *tag* <foreign> potremmo specificare la lingua con l'attributo *lang*:

```
La sua <foreign xml:lang="en"> cupcake</foreign>
alla vaniglia era buonissima.
```

[293] Molti linguaggi di programmazione offrono, in aggiunta, supporto alla gestione di XML attraverso funzioni native. Relativamente a PHP, per es., < http://php.net/manual /it/refs.xml.php>. Ultima consultazione 20 ottobre 2015.

[294] Harold / Scott (2001: 26).

Una delle prerogative di XML è quella di occuparsi del contenuto dell'informazione e non della sua rappresentazione grafica, che verrà affidata a uno o più fogli di stile[295]. Affinché due applicazioni possano scambiarsi dei dati in formato XML è necessario però che queste conoscano come viene strutturata l'informazione all'interno del file, ovvero quali sono e come sono chiamati i *tag* e gli attributi che costituiscono il file XML. Per questo scopo sono state sviluppate alcune tecnologie quali XML Namespace, Dtd e XML Schema.

Per evitare la proliferazione incontrollata degli schemi di codifica e favorire la convergenza verso uno *standard* condiviso, dalla fine degli anni '80 è stato avviato un progetto per la definizione di un sistema di riferimento per la marcatura del testo nelle scienze umanistiche. Nacque così — anche grazie all'appoggio della Association for Computers and the Humanities (ACH), della Association for Computational Linguistics (ACL) e della Association for Literary and Linguistic Computing (ALLC) — la *Text Encoding Initiative* (TEI[296]). Divenuta consorzio permanente

---

[295] XSL (eXtensible Stylesheet Language). Il sistema di compone di due parti:

- XSLT (eXtensible Stylesheet Language transformation): serve a trasformare un documento XML in un altro documento XML o in un documento HTML, XHTML o SVG. La trasformazione viene realizzata da un XSLT Processor (es. Xalan) che riceve come *input* il file XML da trasformare e il file XSL con la definizione dello stile e produce come *output* il file trasformato.
- XSL-FO (eXtensible Stylesheet Language – Formatting Objects). Permette di definire parametri di impaginazione e produrre documenti PDF, RTF o PS.

[296] <http://www.tei-c.org/>. Ultima consultazione: 30 ottobre 2015. Si veda anche il Wiki <http://wiki.tei-c.org/>. Il Consortium si occupa anche di organizzare l'annuale *TEI Conference* e di pubblicare il *Journal of the Text Encoding Initiative* <http://journal.tei-c.org/>.

nel 2000, la TEI sviluppa e pubblica una serie di linee guida (le *Guidelines TEI*) che descrivono uno schema XML modulare di riferimento. Giunto alla 5ª versione (TEI P5), lo schema TEI è ormai uno *standard* riconosciuto a livello internazionale per chiunque si occupi di informatica umanistica.

4.3 LEXICAL MARKUP FRAMEWORK

Il Lexical Markup Framework (LMF) è lo *standard* ISO (International Organization for Standardization) per la rappresentazione di dizionari processabili automaticamente. LMF è stato rilasciato nel 2008, con la specifica ISO - 24613:2008, «after a 5-year study and series of meetings gathering 60 lexicon managers and linguists coming from various cultures and languages»[297].

Attraverso questo sistema si mira a fornire un quadro di riferimento comune per la creazione e l'utilizzo di risorse lessicali per lo scambio di dati e per l'integrazione tra queste risorse. LMF si basa sulla codifica Unicode e usa UML (Unified Modeling Language, Linguaggio di Modellazione Unificato), un linguaggio di modellazione e specifica basato sul paradigma orientato agli oggetti. Le costanti linguistiche — come /feminine/ o /transitive/ — non sono definite nel *framework* ma sono specificate in un Data Category Registry (DCR)[298].

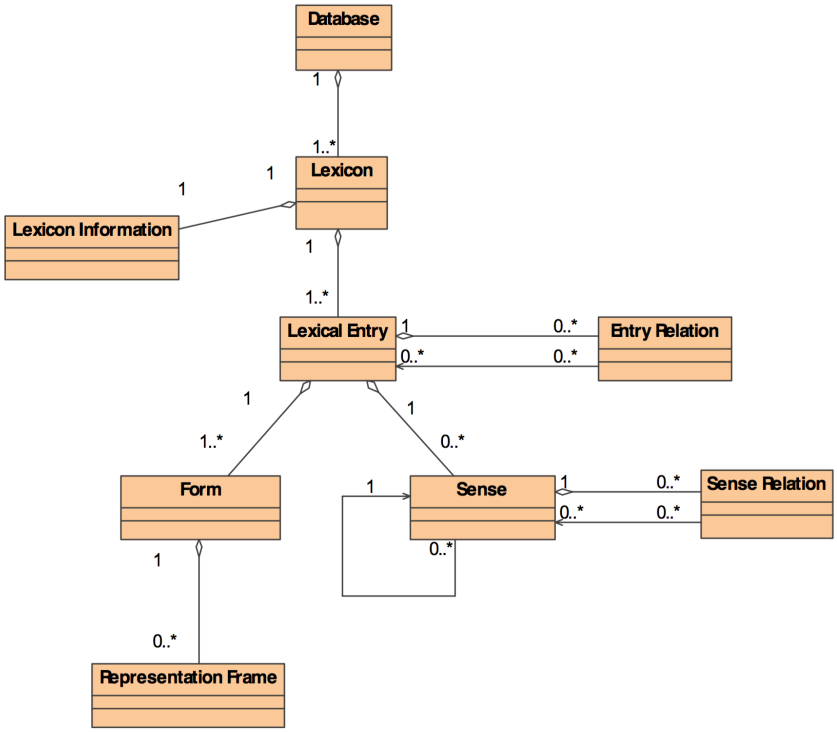
LMF è composto da due componenti principali:

3. Il *core package*, che descrive la struttura gerarchica delle informazioni. L'intera risorsa è inclusa all'interno

[297] Francopulo / Huang (2014).

[298] Francopulo et al. (2006).

della classe Database, che può contenere uno o più classi *Lexicon*. A ogni *Lexicon* è associata una classe *Lexicon Information* che contiene le informazioni amministrative e altri attributi generali relativi al *Lexicon*. Ogni *Lexicon* contiene da una a più entrate lessicali (*Lexical Entry*), connesse a una o più forme (classe *form*) e uno o più sensi (classe *sense*). Ogni entrata gestisce anche eventuali relazioni con altre entrate e i loro sensi. Il *core* può essere raffigurato con il seguente diagramma[299]:



4. Le estensioni del *core package*. Una specifica impresa lessicografica può scegliere un insieme di estensioni utili per i propri fini (ad esempio per la morfologia o la sintassi).

La descrizione strutturale è conforme ai principi di modellazione di UML (Unified Modeling Language) e può essere

[299] Francopulo et al. (2006).

espressa anche in XML. Si riporta un esempio tratto dalla pagina ufficiale[300]:

```
<?xml version="1.0" encoding="UTF-8"?>

<!--<!DOCTYPE LexicalResource SYSTEM
"DTD_LMF_REV_16.dtd">-->
<LexicalResource dtdVersion="16">
      <GlobalInformation>
            <feat att="label" val="Simple English LMF
            test suites"/>
            <feat att="languageCoding" val="ISO 639-3"/>
      </GlobalInformation>

      <Lexicon>
            <feat att="language" val="eng"/>
            <LexicalEntry>
                  <!--Inflected forms of clergyman are
                  clergyman and clergymen-->
                  <feat att="partOfSpeech"
                  val="commonNoun"/>
                  <Lemma>
                        <feat att="writtenForm"
                        val="clergyman"/>
                  </Lemma>
                  <WordForm>
                        <feat att="writtenForm"
                        val="clergyman"/>
                        <feat att="grammaticalNumber"
                        val="singular"/>
                  </WordForm>
                  <WordForm>
                        <feat att="writtenForm"
                        val="clergymen"/>
                        <feat att="grammaticalNumber"
                        val="plural"/>
                  </WordForm>
            </LexicalEntry>
      </Lexicon>
```

[300] <http://www.lexicalmarkupframework.org/>. Ultima consultazione: 14 ottobre 2015.

```
</LexicalResource>
```

La struttura di XML, forse più familiare nel nostro dominio di ricerca, ripropone e traduce la controparte UML. Le regole per effettuare la trasposizione sono semplici:

> Every UML class is rendered as an XML element with the same name and associated with a very small *set* of mandatory attributes like the DTD version.
>
> All XML elements may be adorned with a pair combining a DCR (data category registry) constant (like "grammatical Number") with a value which may be another DCR constant (like "singular") or a free string value (like "clergymen") [301].

4.4 SULLA CODIFICA DEI FILE PER GATTO

Merita un'ultima annotazione il formato ANSI, utilizzato da GATTO per la codifica dei testi da indicizzare nei corpora. Con questa sigla — nota all'utente medio come una delle prime opzioni di salvataggio dell'applicazione Notepad di Windows — si indica un'estensione del *set* di carattere ASCII, con l'aggiunta di 128 codici supplementari, per un totale, dopo l'ingresso del simbolo dell'euro, di 218 posizioni [302] (la codifica ANSI usa 8 *bit* per carattere contro i 7 *bit* dell'ASCII). Questa tabella di codifica è stata sviluppata da Microsoft ed è nota anche come Windows-1252. I caratteri ANSI da 32 a 127 corrispondono a quelli del codice ASCII a 7 *bit* (*Basic Latin Unicode Character Range*), mentre i

[301] Francopulo / Huang (2014: 43).

[302] Una tabella è disponibile all'indirizzo < http://www.alanwood.net/demos/ansi.html>, ultima consultazione: 10 novembre 2015.

caratteri da 160 a 255 corrispondono al *Latin-1 Supplement Unicode*.

Le caratteristiche del sistema di annotazione di GATTO ne fanno un formato proprietario non *standard*. Il supporto a XML/TEI e Unicode, corredato da alcune utility per la conversione, dovrebbe essere presente già nella versione 4 del programma al fine di garantire l'interoperabilità e l'aderenza agli *standard*. Pur aspettando questa interessante quanto necessaria caratteristica, fino a questo momento il sistema di GATTO si è rivelato di grande versatilità: isolando un insieme minimale di caratteristiche testuali da codificare, si configura come strumento agile, sintetico e caratterizzato da una curva di apprendimento estremamente rapida. Inoltre i documenti sono di fatto dei file ANSI e, pur non essendo dotato di una specifica esplicita, il sistema di marcatura è descritto accuratamente ed esaustivamente nel manuale di GATTO[303]. In questo modo l'accessibilità e la permanenza dei dati dei corpora codificati con questo sistema non sono compromessi.

[303] Iorio-Fili (2012).

5 LA RACCOLTA DEI DATI: *CORPUS ARTESIA 2015*

5.1 IL PROGETTO *ARTESIA*

Artesia (*Archivio Testuale del Siciliano Antico*) [304] è un progetto, diretto da Mario Pagano e Margherita Spampinato, condiviso tra l'Università di Catania e il Centro di Studi Filologici e Linguistici Siciliani di Palermo[305]. *Artesia* vuole dare conto, da una prospettiva romanza, della produzione testuale in volgare siciliano dalle prime attestazioni del XIV secolo sino alla prima metà del XVI, periodo in cui il siciliano è sostituito dal toscano come lingua dell'amministrazione.

Artesia sviluppa una serie di strumenti e risorse per lo studio del siciliano medievale che si raccolgono attorno al portale[306], dove, attraverso un database relazionale, i nodi della rete di risorse

[304] Per una presentazione del progetto si vedano anche Pagano (2009) e Pagano / Arcidiacono (2013).

[305] Il progetto è stato sviluppato negli anni, anche grazie alla partecipazione a Progetti di ricerca di Rilevante Interesse Nazionale (PRIN 2007, «Studio, Archivio e Lessico dei Volgarizzamenti Italiani (SALVIt); <www.salvit.org>, ultima consultazione: 02 dicembre 2015), con il fruttuoso contatto con il progetto TLIoN (Tradizione della Letteratura Italiana on the Net), ma soprattutto con l'OVI. Due finanziamenti (nel 2005 e nel 2006) dall'Assessorato ai Beni Culturali ed Ambientali della Regione Sicilia ne hanno permesso l'avvio.

[306] <http://www.artesia.unict.it>, attualmente in manutenzione per l'implementazione di alcuni aggiornamenti sviluppati anche nell'ambito della presente ricerca.

vengono articolati in un unico sistema documentario organico. La base di dati, implementata in Microsoft SQL Server, integra e struttura i contenuti generando dinamicamente alcuni percorsi ipertestuali che raccolgono l'informazione disponibile in quadri chiari ed esaustivi.

Selettori e filtri permettono all'utente di selezionare dinamicamente l'informazione desiderata e un motore di ricerca interno, basato su Lucene [307], provvede alla ricerca *full-text* all'interno di tutto l'archivio. Più che un insieme di pagine statiche, il *Portale Artesia* aspira quindi a costruire uno spazio dinamico e flessibile: l'utente può accedere alla rete di risorse da un punto qualsiasi del sistema (un autore, un'opera, un manoscritto ecc.) e navigare attraverso tutte le risorse pertinenti disponibili selezionate in automatico dal sito.

La *Tradizione testuale* fornisce un archivio in grado di garantire un'informazione complessiva su *Autori*, *Opere*, *Incipit/Explicit*, *Manoscritti*, *Bibliografia*. Le schede sugli *Autori*, oltre a riportare una breve presentazione per ciascuno di questi, mettono in relazione l'autore con tutte le sue opere e selezionano dinamicamente tutti i riferimenti pertinenti presenti nel sito. Un'icona segnala l'eventuale indicizzazione all'interno del *Corpus Artesia* e permette il reindirizzamento alla banca dati interrogabile con GattoWeb.

Nella sezione *Manoscritti* vengono brevemente descritti i codici[308], alcuni dei quali visualizzabili in formato immagine. In qualche caso, la galleria di immagini è corredata anche da un restauro virtuale creato appositamente per *Artesia*. Lo *Scaffale di*

[307] <https://lucene.apache.org/core/>, ultima consultazione: 2 dicembre 2015.

[308] L'elenco dei testi in essi contenuti è ordinabile sia per carte, sia alfabeticamente.

Artesia è una piccola biblioteca digitale destinata a contenere edizioni e studi, alcuni dei quali anche di non facile reperimento.

5.2 IL *CORPUS ARTESIA*

Tra le risorse sviluppate e messe a disposizione dal progetto *Artesia*, un ruolo di primo piano spetta al *Corpus*.

Il *Corpus Artesia* è un corpus storico di testi in volgare siciliano selezionati su un criterio principalmente diacronico. Il limite di inclusione inferiore è rappresentato dall'effettiva disponibilità degli scritti: i testi in siciliano medievale più antichi risalgono ai primi decenni del XIV secolo e il primo documento a riportare una datazione certa sono i *Capitula super cassia* del 1320 [309]. A quest'altezza cronologica il siciliano comincia ad emergere all'interno di una tradizione di scritture in volgare che è essenzialmente espressione della strategia politico-dinastica della corte di Federico III d'Aragona e dell'emergere di «una precisa identità culturale e linguistica»[310], la cui manifestazione più alta si raggiunge già pochi anni dopo con i tre grandi volgarizzamenti del XIV secolo (DialaguXIVS; EneasXIVF; ValMaxXIVU).

Il limite superiore si situa intorno alla prima metà del XVI secolo, quando i numerosi fattori che hanno contribuito a riconfigurare lo spazio culturale e linguistico in direzione extraisolana trovano un punto di arrivo ideale nella sostituzione del toscano al siciliano negli impieghi pubblici. Già dai primi anni del XVI secolo, i due sistemi linguistici entrano in «competizione

[309] Rinaldi/2005 (1).

[310] Bruni (1980: 208).

silenziosa[311]»; il primo documento in forma compiutamente toscana redatto in Sicilia porta la data del 3 aprile del 1524. L'incertezza e le oscillazioni, che producono un'alternanza linguistica tra siciliano toscanizzato e toscano infarcito di sicilianismi, verrà formalmente risolta l'8 ottobre del 1652 con la promulgazione della *Prammatica vicereale* la quale stabiliva che «di qua innanzi, tutti gl'Atti e Contratti che si faranno, s'habbiano di publicare in lingua volgare[312]».

L'unico ulteriore requisito della banca dati riguarda la disponibilità e la qualità dell'edizione utilizzabile, da considerare nella sua affidabilità filologica al fine di garantire adeguati *standard* di affidabilità dei dati. La valutazione può avere come esito la piena accettazione, l'allestimento di un'edizione contenente emendamenti (previa compilazione di un file di GATTO associato di tipo 'note'), oppure il totale rifiuto.

In accordo con le linee guida del progetto, il *Corpus* è composto da:

- Testi pubblicati in edizione affidabile.

  Uno degli obiettivi di questo segmento è stata la copertura completa dei testi editi nella "Collezione dei testi siciliani dei secoli XIV e XV" del Centro di studi filologici e linguistici siciliani.

- Testi editi grazie all'attività di ricerca svolta nell'Università di Catania;

- Testi inediti le cui edizioni elettroniche sono state espressamente approntate per Artesia.

---

[311] Lo Piparo (1987: 735).

[312] Prammatica vicereale dell'8 ottobre 1652, in Muta (1773: 178-187), cit. in Sardo (2008: 9).

L'edizione del testo viene inserita all'interno del *Corpus* integralmente, includendo, dove opportuno, la marcatura selettiva degli apparati. Eventuali operazioni di normalizzazione, ma soprattutto gli interventi filologici mirati a garantire l'attendibilità linguistica, vengono sempre segnalati in bibliografia o descritti in apposite note di GATTO. Ogni intervento sulla punteggiatura, sulla paragrafematica, sulla segmentazione delle parole e sulle lezioni è sempre registrato all'interno del *Corpus* e immediatamente visibile in fase di consultazione. In una prospettiva orientata, oltre che sui testi, anche sui testimoni, per le opere pervenute in più di un manoscritto, tutti i testimoni più significativi ed editi adeguatamente troveranno posto nel *Corpus*.

L'obiettivo complessivo è quindi di raccogliere e rendere accessibili, oltre ad una ricca e significativa selezione di documenti, la totalità dei testi letterari e paraletterari in volgare siciliano disponibili in edizioni affidabili. Lo scopo è stato in buona parte raggiunto attraverso le 4 pubblicazioni degli ultimi anni:

Corpus Artesia 2008:	66 opere	33 documenti.
Corpus Artesia 2009:	68 opere	171 documenti.
Corpus Artesia 2010:	72 opere	171 documenti.
Corpus Artesia 2011:	76 opere	184 documenti.

La prima versione del *Corpus Artesia*, pubblicata anche su CD-ROM, è stata rilasciata il 24 luglio del 2008, con un totale di 66 opere e 37 documenti. L'aggiornamento del 2009 ha visto l'importante inclusione dei 154 documenti trecenteschi editi da Rinaldi (2005); al sostanziale incremento dei documenti marcati è corrisposto un miglioramento del bilanciamento tra le due partizioni individuate in fase di *corpus design* [313] . Con l'aggiornamento del 18 marzo 2010, il *Corpus* ha raggiunto

[313] Si veda oltre.

l'obiettivo di indicizzare tutti i testi editi nella «Collezione di testi siciliani dei secoli XIV e XV» del Centro di studi filologici e linguistici siciliani.

Come già accennato, il *Corpus Artesia*, già preziosa fonte di evidenza linguistica per il siciliano medievale, si candida programmaticamente a costituire la base documentaria per la redazione del *Vocabolario del Siciliano Medievale*. Sebbene sia stato già impiegato più di una volta in progetti e studi lessicografici alcuni dei quali, come il *VSES*, non privi di prestigio, l'ultima versione del *Corpus*, aggiornata oltre quattro anni fa, lasciava ancora aperte alcune questioni riferibili ad almeno due ordini di interventi:

1. Incremento della rappresentatività lessicografica del campione di testi relativo alla scopertura sui testi d'archivio per il secolo XV.
2. Lemmatizzazione del *Corpus*.

Come già accennato, è stato scelto di rimandare la lemmatizzazione sistematica del *Corpus* dopo la futura riconversione in GATTO 4, per il quale si aspetta il rilascio ufficiale del programma e della relativa documentazione. Il primo punto verrà invece affrontato nei prossimi paragrafi, con la proposta una riflessione teorica sulla rappresentatività lessicale e, a seguire, di una nuova versione del *Corpus*, sviluppata nell'ambito del presente lavoro e del quale fa parte a tutti gli effetti come necessario momento applicativo.

5.3 CAMPIONAMENTO E DEFINIZIONE DELLA POPOLAZIONE

5.3.1 Rappresentatività

«È evidente che, mirando all'analisi induttiva di dati linguistici autentici per risalire a conclusioni valide ad un livello più ampio e generalizzato dello studio linguistico, la base empirica debba necessariamente aderire a criteri di rappresentatività»[314].

L'assunto è inattaccabile: non è data induzione valida senza rappresentatività. La rappresentatività si definisce come la capacità di un "campione" (*sample*) di riprodurre adeguatamente la variabilità della "popolazione" di riferimento (*population*) da cui il campione è estratto. In altre parole, data una popolazione, definita come «la totalità dell'universo fenomenico» [315] sulla quale si vogliono osservare fatti, regolarità e dinamiche, si tratta isolare un campione, «sottoinsieme osservabile» [316], capace di includere lo spettro completo della variabilità propria della totalità dal quale viene prelevato[317].

Una base di dati che non sia rappresentativa è, nella migliore delle ipotesi, inutile, nella peggiore pericolosa e fuorviante. Il *Corpus Artesia* dal momento in cui si propone di sostenere la redazione del *VSM* è perciò chiamato a esibire preliminarmente una ricchezza che azzeri il pericolo di dimostrarsi inadeguato a

---

[314] Barbera / Corino / Onesti (2007: 49).

3. Cosa è un corpus? Per una definizione più rigorosa di corpus, token, markup. 25-88. Cit a pag. 49

[315] Beccaria *(1994 s.v. Linguistica statistica)*.

[316] Ibidem.

[317] La terminologia suggerisce la tradizionale vicinanza con le scienze sociali (Lenci / Montemagni / Pirrelli 2005:123).

riprodurre in scala tutti i fenomeni degni di interesse lessicografico: la rivelazione dell'insufficienza rischierebbe di manifestarsi a lavoro avviato, quando potrebbe essere troppo tardi per tornare indietro.

Il primo parametro nella valutazione del *Corpus* che farà da base al vocabolario è quello di stabilire in primo luogo se è stata raggiunta la grandezza ottimale per rappresentare adeguatamente il siciliano medievale. La stessa definizione di corpus presuppone, anche intuitivamente, che la collezione di testi abbia dimensioni generose. Come prescrivono le linee guida dell'Expert Advisory Group on Language Engineering Standards (EAGLES), «the default value of Quantity is large. A corpus is assumed to contain a large number of words. The whole point of assembling a corpus is to gather data in quantity»[318].

Come va interpretato quel «large»? Possiamo tenere in conto le generiche esperienze di indagine sul *Corpus Artesia* effettuate negli ultimi anni o dobbiamo aspettarci criteri particolari mirati sullo scopo lessicografico? La scienza che si occupa della rappresentatività dei campioni è la statistica; nel rispetto delle giuste competenze disciplinari — e a dispetto delle naturali difficoltà per l'umanista nel dominare il terreno del formalismo matematico — nelle sue leggi si dovranno cercare i fondamenti elementari che regolano i comportamenti dei campioni e ne definiscono le proprietà[319].

5.3.2 La statistica monovariata e rappresentatività

Un caso esemplare di applicazione delle teorie della statistica descrittiva e delle tecniche monovariate è descritto in

---

[318] *EAGLES Guidelines* alla pagina <http://www.ilc.cnr.it/EAGLES96/typology/node11.html#SECTION00045100000000000000>.

[319] Atkins et al. 6 e 7.

Biber (1993) — sulla base anche sui precedenti lavori (Biber 1986, Biber 1988 e Biber 1990). La sua analisi muove dal concetto di deviazione *standard*, argomento che potrebbe sembrare elementare a uno specialista, ma che non è detto che lo sia per il filologo o il linguista; la distanza disciplinare giustificherà l'integrazione preliminare di pochi sintetici appunti sulle nozioni di statistica campionaria che partecipano alla definizione dell'errore *standard*.

Dato un campione di (N) elementi, possiamo calcolare su di esso un insieme di statistiche campionarie. Con μ si indica la media, che, come noto, si calcola mediante la formula:

$$\mu = \Sigma X/N$$

Con σ si indica la deviazione *standard*, chiamata anche scarto quadratico medio, che rappresenta la distribuzione degli elementi della popolazione o del campione che si vuole osservare, «una misura statistica che ci consente di "allargare" la nostra inquadratura sul modo in cui i dati si dispongono intorno alla media[320]»:

$$\sigma = \text{sqrt } [(\Sigma((X-\mu)^2))/(N)]$$ [321]

Dividendo la deviazione *standard* per la radice quadrata di N (numero di elementi del campione) si può stimare la distanza tra la media del campione e la media della popolazione; questa operazione è, appunto, il calcolo dell'errore *standard* della media:

$$\text{Errore } standard = \sigma/\text{sqrt}(N)$$

L'errore *standard*, come si può cogliere anche intuitivamente, diminuisce con il diminuire della deviazione *standard* della variabile (σ) e con l'aumentare delle dimensioni del

[320] Lenci / Montemagni / Pirrelli (2015: 130).

[321] $\sigma = \text{sqrt } [(\Sigma((X-\mu)^2))/(N-1)]$ se si misura sul campione.

campione. A partire da questa equazione prende avvio l'applicazione ai corpora testuali di Biber (1993: 248):

> «if the sample size is greater than 30, then the distribution of sample means has a roughly normal distribution, so that 95% of the samples taken from a population will have means that fall in the interval of plus or minus 1.96 times the *standard* error. The smaller this interval is, the more confidence a researcher can have that she is accurately representing the population mean».

In altre parole, se la grandezza del campione supera un certo valore soglia, la media dei campioni, nel 95% dei casi, non si allontanerà da quella della popolazione oltre un dato limite (1,96 volte). Dalla formula si deduce inoltre che l'influenza delle dimensioni del campione sulla rappresentatività è costante e che per dimezzare l'errore *standard* si deve aumentare la dimensione del campione di quattro volte. Per esemplificare, sul calcolo dei nomi contenuti in un testo si potrà affermare quanto segue:

> if the sample standard deviation for the number of nouns in a text was 30, the sample mean score was 100, and the sample size was nine texts, then the *standard* error would be equal to 10:
>
> *Standard* error = 30/V(9) = 30/3 = 10
>
> This value indicates that there is a 95% probability that the true population mean for the number of nouns per text falls within the range of 80.4 to 119.6 (i.e. the sample mean of 100 ± 1.96 times the *standard* error of 10).
>
> To reduce this confidence interval by cutting the *standard* error in half, the sample size must be increased four times to 36 texts; i.e.
>
> *Standard* error = 30/V(36) = 30/6 = 5

A dispetto dell'autorità e apparente perfezione della formula, sulla specifica applicazione al dominio testuale emergono tre difficoltà peculiari:

1. La grandezza del campione (N) deriva da una determinazione preliminare dell'intervallo di affidabilità del corpus. L'incertezza tollerabile deve quindi essere determinata a priori.
2. L'equazione dipende dalla deviazione *standard* di una particolare variabile. La legge rivela, insomma, che la dimensione ottimale di un corpus dipende da quella variabile, cioè dal tipo di fenomeno linguistico indagato (e questa è una prima risposta ai nostri interrogativi).
3. Il problema più importante è quello della circolarità. Prima di applicare la formula bisogna aver calcolato la deviazione *standard* di un campione (che dobbiamo presupporre rappresentativo).

Alle considerazioni di Biber (1993: 248) — senza voler sminuire l'importanza e l'eleganza della descrizione — andranno anche accostate le perplessità di Atkins / Clear / Olster (1991: 6-7):

> Unfortunately the *standard* approaches to statistical sampling are hardly applicable to building a language corpus. First, often it is very difficult to delimit the total population in any rigorous way. Textbooks on statistical methods almost always focus on clearly defined populations. Second, there is no obvious unit of language which is to be sampled and which can be used to define the population. We may sample words or sentences or 'texts' among other things. Third, because of the sheer size of the population and given current and foreseeable resources, it will always be possible to demonstrate that some feature of the population is not adequately represented in the sample.

Le leggi di base della statistica campionaria hanno confermato e descritto numericamente la nozione intuitiva concernente l'influenza della dimensione sulla rappresentatività del corpus (per dimezzare l'errore *standard* si deve quadruplicare la dimensione del campione). Su questa base sarà possibile stimare genericamente i vantaggi attesi per gli aggiornamenti. La deviazione *standard* ha inoltre formalizzato la stretta dipendenza del concetto di rappresentatività con la distribuzione dell'elemento che si vuole rappresentare.

A ben vedere ricaviamo anche alcune indicazioni supplementari. Nel rapporto tra la tolleranza all'errore che siamo disposti ad accettare e la frequenza del fenomeno che vogliamo indagare, Biber (1993: 253) quantifica tolleranze abbondantemente più basse per i fenomeni rari. Più che domandarci quanto vogliamo spingerci nel territorio del lessico raro, l'osservazione richiede prima che si risponda a l'altro tipo di quesiti: la parola "media" che ci si aspetterebbe di trovare in un dizionario quanto è rara? Mediamente, quanto possono dirsi rare le parole che vogliamo studiare?

5.3.3 Lessicografia e fenomeni infrequenti

Non va nascosta una certa reticenza, se non impossibilità oggettiva, nel descrivere attraverso una sola variabile numerica le informazioni lessicograficamente e filologicamente rilevanti. Forse si stanno fraintendendo gli assunti statistici, ma il concetto di popolazione si mostra per molti versi sfuggente.

Vanno inoltre tenute in considerazione le osservazioni già proposte sui corpora storici da considerare a tutti gli effetti corpora chiusi[322]. Le strozzate vie d'accesso al dato linguistico per i testi medievali possono giungere a mettere in discussione l'esaustività

[322] Cfr. § 3.6.

dei vocabolari delle lingue medievali e la stessa definizione di corpus:

> Whilst dictionaries of modern languages can reasonably aspire to exhaustiveness, in particular by the use of vast electronic corpora, this is probably never achievable for medieval languages. In the first instance, it is very difficult to imagine successfully creating a properly *representative* corpus; and clearly, it is also in practice unlikely that we will ever digitize the entirety of the surviving documents in even as relatively circumscribed a language as Anglo-Norman. It would be more correct to describe even the most comprehensive dictionaries of medieval languages as making use of textual data banks or what I have called here *text-bases*, rather than corpora *stricto sensu*[323].

Si dovrebbe guardare alla tradizione di scritture in siciliano medievale di cui siamo in possesso, nel suo complesso con atteggiamento simile a quello di Raffaele (2009: 16):

> Ci si dovrebbe chiedere chi fossero gli autori e a quali lettori (quanto meno idealmente) essi si rivolgessero; di quale livello di istruzione disponessero sia gli uni che gli altri; quali fossero i rispettivi ambiti sociali di appartenenza e per quali circuiti di circolazione queste opere fossero state concepite. E ci si chiede, soprattutto, […] quale sia stata la loro reale consistenza quantitativa, rispetto sia al resto della produzione culturale isolana sia a quanto è sopravvissuto fino ad oggi. Sono interrogativi, questi, ai quali allo stato attuale degli studi non è agevole dare precise risposte.

La carenza più sentita in questi casi è nel mancato accesso alla lingua parlata, che non si dà mai nella piena compiutezza, per quanto alcuni testi possano avvicinarsi alla resa accurata, spesso

[323] Trotter (2011: §8).

nella dimensione testimoniale di alcuni di essi. I casi sono rari e, malgrado i noti limiti, quanto mai preziosi.

Si prendano, per esempio, le Testimonianze di ingiurie rivolte dal notaio Giovanni Cesareo, maestro di sciurta, al cavaliere Giovanni Aiello[324] indicizzate nel *Corpus Artesia*. La resa del dialogo, mediata dal ricordo dei testimoni (e per questo ancora più interessante nelle differenze e in alcune aggiunte colorite nella rievocazione nella memoria del parlante) è quanto mai vivace e realistica. In virtù del vivace repertorio di ingiurie riportate dal documento, il caso ci riporta anche dell'infuocato attacco di Chomsky (1962), emblema dell'opposizione generativista alla linguistica dei corpora che determinò «il blocco pressoché completo dei finanziamenti ai progetti computazionali di tutta una generazione[325]»:

> Any natural corpus will be skewed. Some sentences won't occur because they are obvious, others because they are false, still others because they are impolite. The corpus, if natural, will be so wildly skewed that the description would be no more than a mere list.

Si consideri qualche esempio di lessemi «*impolite*», quantificandone anche le ulteriori attestazioni:

raballusu, agg. *cornuto* (*detto come ingiuria*). Con questo stesso significato presente nel dialetto odierno, cfr. *VS* s.v. *rravagghjúsu*.

[324] Rinaldi/2005 (61).

[325] Barbera (2013: 12);

Il lessema non è un *hapax*, *ruvaglusu* e *ravaglusu* sono attestati anche anche in DeclarusXIVM.

Bagassa, s.f. Donna di malaffare (con significato fortemente dispregiativo [326]). Con questo stesso significato presente nel dialetto moderno, Cfr. *VS* s.v. *bbagàscia*.

Bagassa, che doveva essere di uso altrettanto comune se un confessionale riporta «hai iniuriatu ad alcunu e dictuli 'cani', oi 'bagaxa'?[327]» è meno frequente ma sufficientemente attestato. Il termine inoltre è glossato anche nel *VSL* (che sarà indicizzato nell'aggiornamento del 2016). In quanto a *cani*, come agg. qualificativo di disprezzo, le attestazioni si rilevano con una certa frequenza in tutto il *Corpus*[328].

Spicca poi, in una letteratura dominata da testi «per lo più di argomento devoto, edificanti, diretti alla formazione del buon cristiano[329]», la locuzione ingiuriosa *fillu di previti*, per il quale si rimanda a Spampinato (2013: 687).

Gli esempi non mirano, come ovvio, a smentire i passi prima citati di Trotter e Chomsky, ma ci danno l'opportunità di trarre osservazioni concrete sulla probabilità statistica di lessemi reali e sulla diversa probabilità di occorrenza dei fenomeni lessicali: ci riportano realisticamente alle tolleranze di errore da accettare, alle preoccupazioni sulla rappresentatività lessicografica di forme

[326] Cfr. anche *TLIO* e *GDLI* s.v. *Bagascia*.

[327] Conf3XVB, 173, 15.

[328] Cfr. *TLIO* s.v. *cane* (4).

[329] Pagano (2013: 793).

comuni nella lingua ma penalizzate sui corpora, e anche alle accezioni particolari e rare.

Il fenomeno raro in lessicografia è uno degli elementi con i quali si avrà spesso a che fare perché rappresenta in qualche modo la normalità. La descrizione di Meyer (2002: 14) della composizione del *LOB Corpus*[330] fotografa la condizione classica della distribuzione statistica delle occorrenze lessicali:

> In the *LOB Corpus*, the five most frequent lexical items are the function words the, of, and, to and a. The five least frequent lexical items are not five single words but rather hundreds of different words that occur from ten to fifteen times each in the corpus. These words include numerous proper nouns as well as miscellaneous content words such as alloy, beef, and bout.

Compilato negli anni '70, il *LOB Corpus* è stato progettato con lo scopo di realizzare una controparte del *Brown Corpus*[331] per l'inglese britannico e, per questo motivo, ne ha ricalcato la composizione, raccogliendo 500 campioni di circa 2000 parole per un totale di 1 milione di occorrenze.

Siamo quindi vicini, in termini di occorrenze, alle dimensioni del *Corpus Artesia*. Le rilevazioni sono equiparabili: nella versione del 2011, le cinque forme più frequenti sono *di* (54.496 occorrenze), *et* (51.661 occorrenze), *lu* (40.270

---

[330] *Lancaster-Oslo/Bergen Corpus* (*LOB*), <http://www.essex.ac.uk/linguistics/external/clmt/w3c/corpus_ling/content/corpora/list/private/LOB/lob.html>, ultima consultazione: 6 dicembre 2015.

[331] *Brown Corpus of American Written English*, allestito nel 1964 da Winthrop Nelson Francis ed Henry Kučera alla Brown University. < http://www.essex.ac.uk/linguistics/external/clmt/w3c/corpus_ling/content/corpora/list/private/brown/brown.html>, ultima consultazione: 6 dicembre 2015.

occorrenze), *la* (34.990 occorrenze) e *li* (33.566 occorrenze). Con 214983 occorrenze complessive, queste cinque forme (che da sole rappresentano lo 0,0074% delle forme indicizzate in *Artesia*) coprono quasi il 20% delle occorrenze totali. Come ampiamente prevedibile, anche in questo caso siamo in presenza di parole dal significato grammaticale, dette anche 'parole vuote', 'parole funzionali' o '*function words'* [332]. Questo genere di parole occorre con una probabilità molto più alta rispetto alle parole dal significato grammaticale ('parole piene' o '*content words*'), tanto che, nelle tradizionali concordanze a stampa, dove lo spazio era prezioso, subivano trattamento diversificato[333].

Le ultime cinque posizioni dell'indice di frequenza (parole che si ritrovano con 1, 2, 3, 4 o 5 occorrenze totali), invece, comprendono 56.312 forme. Il *Corpus Artesia* è costituito per l'83% da queste parole e percentuali molto simili si potrebbero lecitamente riscontrare su corpora dalle dimensioni simili, come del resto ci conferma il *LOB*. In altre parole, la stragrande maggioranza delle forme del corpus può essere considerata rara, se non molto rara. Su questa moltitudine di elementi che occorrono con una bassa probabilità andrà stimata la rappresentatività del *Corpus Artesia*. Il fenomeno raro è una condizione normale.

La similarità tra *Artesia* e il *LOB* non può derivare, come ovvio, dal sistema linguistico, dalla collocazione diacronica, o dalle differenti tipologie di testi su cui i corpora sono stati bilanciati[334]; anche la dimensione simile, raggiunta una certa massa critica, può essere in parte trascurata. Se è possibile accostare con tanta facilità

---

[332] Cfr. Adorno (2003).

[333] Marello (1996: 175).

[334] Per il *LOB Corpus* si veda l'indirizzo <http://www.essex.ac.uk/linguistics/external/clmt/w3c/corpus_ling/content/corpora/list/private/LOB/lob.html> (ultima consultazione 12/09/2015).

la distribuzione del lessico in corpora così distanti — un corpus di inglese scritto ed edito nel 1961 e in un corpus di testi in un volgare medievale — è in virtù di una regolarità nella distribuzione delle parole nei linguaggi dalla portata più ampia, probabilmente universale, che ci assicura che questo tipo di regolarità tende a presentarsi con uniformità costante attraverso lingue e tempi diversi[335]. Il fenomeno, giustificato dal principio di economia di Martinet, è dimostrato dalla *legge armonica di Zipf-Estoup*, dalla *legge canonica di Mandelbrot* e dalla legge di Zipf sul numero di parole di eguale frequenza. Per quest'ultima si ha:

$$f_n = \frac{p}{n}$$

Dove:

- n indica il rango della parola, la sua posizione nell'indice di frequenza.
- f_n indica la frequenza.
- p è una costante che regola il rapporto tra rango e frequenza.

Il rapporto tra rango e frequenza è quindi specificato da una costante p. L'equazione può essere applicata sul *Corpus Artesia* e osservata per mezzo di un'analisi della distribuzione.

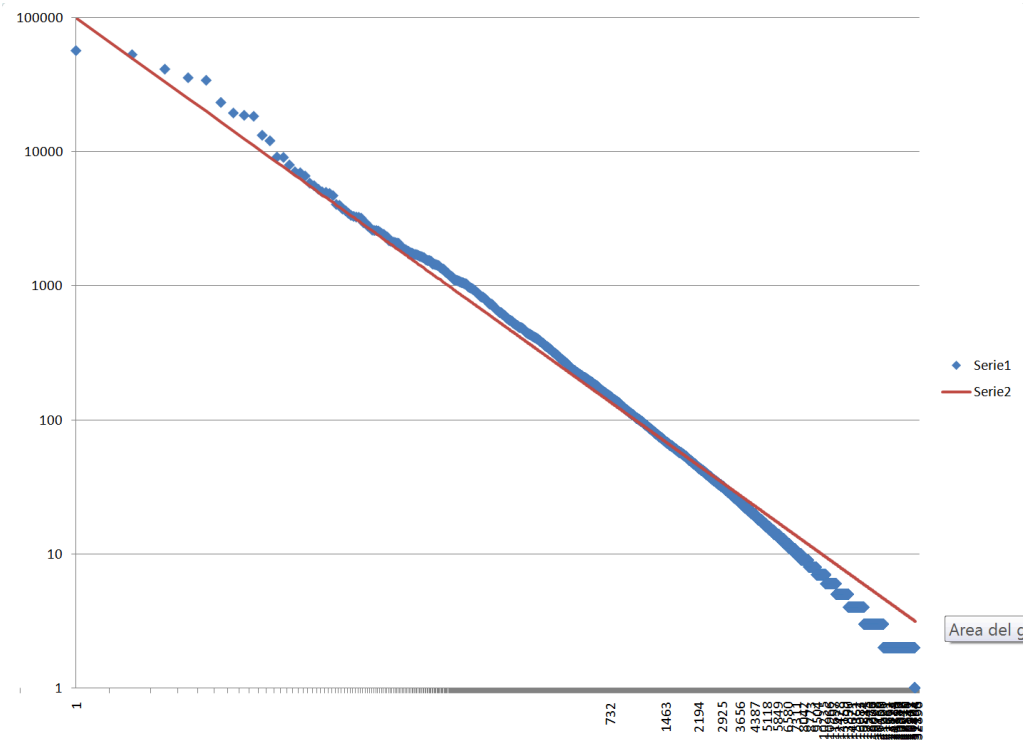
La preparazione dei dati richiede 4 fasi:

1. La creazione dell'indice di frequenza a partire da GATTO.
2. L'esportazione dei risultati in un file di testo contenente campi separati da tabulatori e la successiva importazione dell'*output* di GATTO in Excel.
3. La manipolazione dei dati.

[335] Ed è stata spesso applicata con successo anche e soprattutto al di fuori della linguistica statistica.

4. La creazione di un grafico a dispersione.

Aggiungendo una colonna a indicare il rango, alle singole forme può essere agevolmente applicata la legge di Zipf. I risultati della costante hanno un picco massimo di 172.535 e un picco minimo di 31.896. Verrebbe da dire, con Mandelbrot (1954)[336], «lorsque Zipf essayait de representer tout par cette loi, il essayait d'habiller tout le monde avec des vetements d'une seule taille». Basta però una rappresentazione grafica per riconoscere l'approssimarsi della distribuzione degli elementi alla linea descritta dalla legge. Rappresentando il grafico su scala logaritmica, ecco come si presenta il lessico del *Corpus Artesia 2015*:

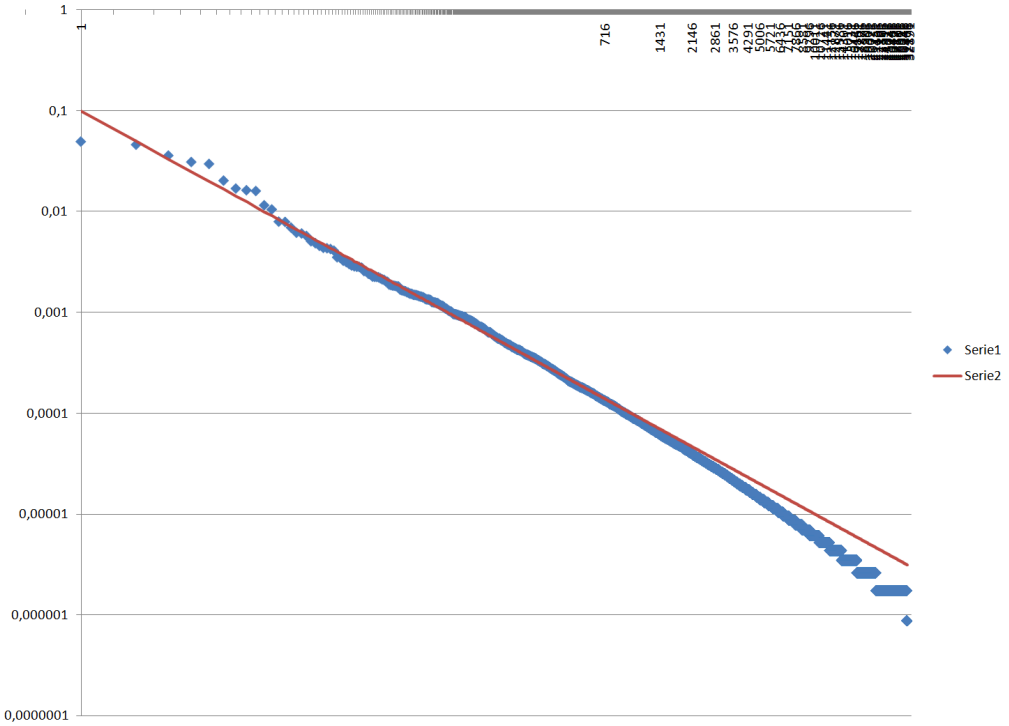


[336] Cit. in Manning-Schütze (1999: 26).

La linea rossa rappresenta le frequenze attese calcolate con la formula di Zipf; in blu sono invece rappresentate le distribuzioni del formario; sull'asse delle y sono riportate le frequenze in scala logaritmica e in quello delle x la scala logaritmica dei ranghi.

Nella parte in alto a sinistra, dove la linea diventa meno densa e si sgrana in corrispondenza delle parole con il rango più alto, si riescono a distinguere anche le forme *di*, *et*, *lu*, *la* e *li* citate prima.

Creando una colonna sulla quale sono state calcolate le frequenze relative (che si ottengono rapportando le frequenze assolute al totale delle occorrenze del *Corpus*) la distribuzione si approssima ancora di più alla linea delle frequenze attese secondo la legge di Zipf:



5.3.4 Il rapporto *type*/*token*

Il rapporto *type*/*token* [337] (*T/t*), o *tipo/unità*, è un semplicissimo quanto tradizionale indice di ricchezza lessicale di un testo o di un corpus. Il rapporto e le sue implicazioni con la rappresentatività sono state messe in luce per la prima volta[338] da Sinclair (1987a)[339]. Sinclair notò che se tutte le parole di un corpus avessero la stessa frequenza di occorrenza, ci sarebbe spazio solo per 5-10 occorrenze di ciascuna di esse. Come invece abbiamo appena visto, alcune sono ripetute molte volte mentre altre lo sono molto meno. Nel *Corpus Artesia 2011* la forma *di*, *così* come tutte le altre, avrebbe ben 16 occorrenze contro le 54.496 effettive. Questo dovrebbe segnalarci che nel *Corpus Artesia* le parole sono maggiormente ripetute rispetto all'esempio di *Sinclair* e che quindi il *Corpus Artesia* è meno ricco: rapporto T/t è dello 0,06[340]. Il *Corpus TLIO* invece sarebbe poverissimo, attestandosi allo 0,02. Il fatto è che l'indice ricavato con il rapporto tra *type* e *token* tende comprensibilmente ad abbassarsi aumentando la lunghezza del testo o ampliando le dimensioni del corpus su cui è calcolato. La matematica ci consegna molti strumenti di controllo ma nessuno perfettamente soddisfacente. La natura sfuggente del linguaggio è un fattore intrinseco di ambiguità. In prospettiva filologica e di

---

[337] Lenci / Montemagni / Pirrelli (2005: 133).

[338] Hanks (2012: 403).

[339] Come Sinclair fa notare, un paio di attestazioni non sono sufficienti a fornire tutti gli schemi di co-occorrenza, e su questa base stimava che un corpus da un milione di parole sarebbe insufficiente per fornire un'informazione lessicografica completa. Sinclair reputava che il più grande corpus esistente all'epoca, da 7.3 milioni di occorrenze, avrebbe consentito una piena rappresentatività solo per le parole più frequenti, lasciando parzialmente scoperte le parole con frequenza più bassa.

[340] Il rapporto T/t può assumere un valore minimo di 0,01 e un valore massimo di 1.

lessicografia storica come poi assegnare un valore a quella singola occorrenza che spesso può valere come preziosissima attestazione? Come si possono poi interpretare per i fini lessicografici delle analisi che vengono ricondotte unicamente al numero di forme indicizzate e ignorano entità di ordine superiore come i lemmi, i significati, o le poliremathiche? In relazione all'applicazione meccanica dei metodi "duri" in alcuni casi si sono tentate ridefinizioni dello stesso concetto di rappresentatività, reindirizzando l'attività di *corpus design* verso approcci "pragmatici":

> This is not perhaps the statistician's definition of the word […] What we mean by *representative* is covering what we judge to be typical and central aspects of the language, and providing enough occurrences of words and phrases for the lexicographers, and other student of language, to believe that they have sufficient evidence from the corpus to make accurate statements about lexical behaviour.[341]

La rappresentatività si ridefinisce nell'atto del giudizio, una strategia per riportare al centro della costituzione del corpus la competenza linguistica dell'esperto che lo progetta. L'esperto, in quanto tale, avrà sicuramente cognizione delle regole statistiche alle quali obbediscono le rappresentazioni dei fenomeni linguistici su grandi archivi testuali ma le sue sue stime saranno integrate anche da competenze diverse.

5.4 IL PROBLEMA DEI TESTI D'ARCHIVIO

[341] Summers (1993: 186-187).

Si riconsiderino due limitazioni già evidenziate nelle scorse pagine:

- Nella composizione di corpora storica non è data piena libertà di movimento, giacché per gli stati passati di lingua ci si trova a lavorare con collezioni di dati limitate. Le condizioni dettate dalla tradizione testuale tendono a limitare le possibilità di *design* del corpus nei confini indeformabili dei corpora 'chiusi' [342], con la conseguente difficoltà ad allargare e a modellare a piacere il campione.
- Le dovute preoccupazioni filologiche impongono una valutazione caso per caso dei materiali da inserire al fine di evitare dati inquinati da lezioni fuorvianti. Il pungolo dell'espansione della base documentaria, reso talvolta più acuto dall'esiguità dei dati, va frenato dalla inderogabile necessità della cernita e della valutazione minuziosa dei testi.

Il campionamento[343] dei testi da includere nel *Corpus Artesia* è stato articolato sulla distinzione primaria tra opere e documenti. Se per la prima partizione è possibile e opportuna la completa inclusione di tutti i testi letterari e paraletterari in volgare siciliano, fino alla maggiore approssimazione possibile al limite indicato nel

---

[342] «Si è soliti distinguere tra c[orpus] chiuso, costituito da un insieme linguistico di dati finiti (parole, frasi, ecc.), e c[orpus] aperto, non finito, bensí (sic) continuamente estendibile. Sono necessariamente analisi ristrette a un c[orpus] chiuso quelle relative ad una lingua morta, in cui il c[orpus] di dati è rappresentato da una collezione finita di testi», Beccaria (1994 s.v. *corpus*).

[343] Il campionamento dei testi può essere distinto in due modalità principali: il campionamento probabilistico (probability sampling) e il campionamento non probabilistico (non-probability sampling) (Cfr. Meyer 2002: 43). La differenza tra i due metodi è l'utilizzo di metodi statistici per compiere un'attenta selezione della popolazione da analizzare per mezzo di tecniche statistiche. Nel nostro caso il campionamento probabilistico non è una strada percorribile.

primo punto, il secondo segmento si rivela filologicamente problematico e richiede di usare come criterio primario la prudenza ribadita dal secondo punto. Sono infatti due i fattori che scoraggiarono il perseguimento dell'esaustività per i documenti: la mancanza di un repertorio dei documenti adatto allo scopo — in altre parole, un *sampling frame*[344] problematico; gli oscillanti e incerti *standard* filologici per le edizioni dei testi d'archivio[345].

Il *Corpus* è quindi idealmente partizionato su queste due tipologie testuali fondamentali, trattate separatamente con due distinte tecniche di campionamento: per la prima parte si è imposto un campionamento esaustivo (sull'edito), per i documenti è stato necessario effettuare una selezione basata sulla valutazione caso per caso.

> Per quanto riguarda i documenti, per es., è necessario indicizzarne quanti più possibile, specie quelli che possono fornire informazioni sulla lingua della vita quotidiana. Ne sono un esempio tredici inventari di dote e di eredità, risalenti agli anni 1430-1456, ottimamente editi da Bresc (1995), che sono stati inseriti nell'aggiornamento del 7 dicembre 2011; essi rappresentano un davvero interessante «tableau [...] de l'objet, du vêtement, de la maison [Bresc (1995: 109)][346].

Sebbene sia possibile ipotizzare che «il volgare compare nelle cancellerie in misura proporzionale a quello che affiora nella

---

[344] Sulla nozione di *sampling frame* si rimanda a Biber (1993: 243-244).

[345] Per la relativa attendibilità delle edizioni dei documenti cfr. Varvaro (1995). Sull'edizione delle fonti documentarie cfr. Pratesi (1957).

[346] Pagano (2012: 120). Si tenga conto che prima di Rinaldi (2005) il testo di riferimento per i documenti trecenteschi è stato Li Gotti (1951)

pratica letteraria[347]» allo stato attuale è impossibile bilanciare equamente le due sezioni del *Corpus*.

I 154 documenti editi da Rinaldi (2005), «certamente uno dei maggiori contributi, negli ultimi anni, alla conoscenza del siciliano medievale[348]», sono stati indicizzati e inseriti integralmente nel *Corpus* a partire dal secondo aggiornamento (versione del 31 gennaio 2009). L'apporto del volume contribuisce sensibilmente alla rappresentatività del *Corpus* con un ricco gruppo di testi di natura non letteraria, caratterizzati da un alto grado di variabilità linguistica[349] e ben diversificati per varietà tipologica; nella classificazione dell'editrice: gabelle, calmieri, capitoli, formule di giuramento; ordinanze e lettere pubbliche; cedole, obbligazioni, stime, testamenti; testimonianze; lettere di cambio; lettere private; inventari, conti e appunti[350].

Dei documenti contenuti nella raccolta, gli inediti si riducono a una ventina, ma la nuova edizione vanta un ricontrollo completo dei testi, effettuato ripartendo dai manoscritti) e mostra un livello qualitativo che ne giustifica la scelta come edizione di

---

[347] Rinaldi (2005: XV).

[348] Pagano (2011: 253).

[349] Musso (2006: 352).

[350] La parte più cospicua dei testi proviene dall'Archivio di Stato di Palermo (fondi della Real Cancelleria, del Protonotaro, della Corte Pretoriana, dei Notai defunti, dei monasteri di san Martino, santa Maria del Bosco, santa Margherita di Polizzi e della Cattedrale di Palermo) e dall'Archivio Storico del Comune di Palermo (Atti del Senato). Sono anche presenti, in misura minore, testi provenienti dall'Archivio de la Corona de Aragón dall'Archivio della Città del Vaticano, dall'Archivio del Monastero di Montecassino, dalla Biblioteca Nazionale Universitaria di Cagliari, nell'Archivio Arcivescovile di Patti, nella Biblioteca Ursino-Recupero, nei tabulari dei monasteri di san Benedetto e di san Nicolò l'Arena a Catania.

riferimento per il *Corpus Artesia* per la totalità dei documenti che contiene. Per questo motivo, nell'allestimento del *Corpus Artesia 2015*, per i documenti già editi in Rinaldi (2005) ma poi riproposti in Bresc-Bautier / Bresc (2014) non è stata giudicata necessaria la sostituzione dell'edizione, mentre le nuove versioni hanno sostituito i precedenti documenti di Bresc (1995).

I documenti di Rinaldi (2005) si distribuiscono nel periodo relativo al *regnum Siciliae*, dai *Capitula super cassia propter guerram*[351] del 1320, testo più antico in siciliano medievale con datazione certa, fino alla fine del XIV secolo[352] (con una riduzione del numero di documenti nell'ultimo scorcio del secolo[353]). La raccolta quindi non copre che i 2/5 del periodo di riferimento del *Corpus Artesia*. La restante parte è stata rappresentata principalmente dalle insufficienti raccolte di Curti (1972), Sardina (2006) e Bresc (1995). La continuità della sezione è quindi stata fortemente sbilanciato in favore del XIV secolo, come più volte segnalato[354].

5.5 GLI INVENTARI DI BRESC-BAUTIER / BRESC (2014)

Il recentissimo e ponderoso lavoro di Bresc-Bautier / Bresc (2014) ci consegna un alto numero di testi per colmare il dislivello tra i documenti indicizzati. Da questa raccolta si possono selezionare e inserire 41 testi, per complessive 28.055 occorrenze e

[351] Rinaldi/2005(1).

[352] Rinaldi/2005 (19) e (75).

[353] Rilevato anche in Musso (2006: 349).

[354] Arcidiacono (2011: 286), Pagano (2012: 120), Pagano / Arcidiacono (2013).

4.345 forme (rapporto *type*/*token* del 15%, non altissimo per un campione di questa dimensione).

La raccolta non registra testi per il primo decennio ma questo periodo era però già abbondantemente rappresentato, per la sezione documenti, già dalle precedenti versioni del *Corpus* con le 14 lettere edite da Curti (1972) e con Sardina/2006 (8). I nuovi inventari ci forniscono solo due brevi[355] testi per il secondo[356] decennio del secolo: *Inventario del bagaglio del magnifico messere Giovanni Valguarnera trasportato dal castello alla casa di Giovanni di Paci* del notaio Pittacolis e L'*Inventario dei beni del fu Federico di Luparello divisi tra i suoi figli Riccardo e Petruzzo*[357] del notaio Bonafede.

Compatta la sezione dagli anni venti fino alla fine della prima metà del secolo: 9 testi nel terzo decennio con 1829 occorrenze per 485 forme; un corposo insieme di 9 testi nel quarto decennio con 9050 occorrenze per 1986 forme; ancora 9 nel quarto con 3816 occorrenze e 934 forme; ben 13 documenti negli anni cinquanta con un totale di 12733 occorrenze per 2430 forme. L'ultimo testo della raccolta è datato 27 agosto 1461 e con 1331 occorrenze per 483 forme è l'unico documento a garantire la continuità per questo decennio.

Per arginare la mancata copertura da quest'altezza cronologica in poi, sono stati indicizzati i testi editi da Biondi (2000). I sette inventari della raccolta riescono a raggiungere lo scopo solo in parte e permane un'estesa lacuna dalla metà dell'8° decennio fino alla fine degli anni '90 (ma col merito di attestarsi appena oltre la soglia del XVI secolo con due testi).

---

[355] 272 occorrenze per 111 forme complessive.

[356] Bresc/2014 (201).

[357] Bresc/2014 (247).

L'indicizzazione *dell'Inventario degli armamenti del castello della città di Malta acquistati da messere Gutierrez de Nava da donna Costanza, vedova di messere Gonsalvo de Monroy*[358] e *l'Inventario degli armamenti del Castellammare di Malta (oggi Forte Sant'Angelo a Birgu) elencati dal castellano messere Gutierrez de Nava*[359] apre la collezione dei testi di *Artesia* ai documenti maltesi e viene aggiunta una nuova variabile categoriale diatopica, rappresentata nelle schede bibliografiche dal valore "mal." nel campo *area generica*. L'arricchimento del *Corpus* con testi maltesi è uno degli obiettivi che il progetto *Artesia* si prefigge di perseguire in futuro[360].

Il ferreo stile diplomatico nel medioevo è imposto solo ai documenti emanati dalle cancellerie ecclesiastiche o civili, mentre gli atti notarili presentano una grande varietà di forme[361] e oscillazione linguistica. A fronte di una piccola minoranza di inventari nei quali «bona vero sunt in vulgari descripta»[362], la grande maggioranza dei testi editi nei sei tomi è in latino, sebbene in un contesto di diglossia a base latina, l'aspetto linguistico dell'atto notarile non si dà sempre in forma pura e varia anche in relazione a fattori extralinguistici come i supporti[363] e le modalità di scrittura: alcuni giovani di studio dei notai redigevano in siciliano e traducevano in latino in un secondo momento, come mostra l'esempio di Bresc/2014 (315): «comme le montre l'exemple de Charonus Taguil (cccxv); le tabellion n'a traduit que quatre lignes

[358] Bresc/2014 (293).

[359] Bresc/2014 (294).

[360] Pagano (2012: 120);

[361] De Lasala / Rabikauskas (2003: 15).

[362] Bresc/2014 (298).

[363] In alcuni casi il documento ci perviene assieme ai fogli di brutta inseriti nella legatura del registro. Cfr. Bresc-Bautier / Bresc (2014: 1580).

et a finalement préféré recopier le brouillon en vulgaire[364]». Una soluzione largamente adottata era di redigere l'atto in latino ma riservare la parte centrale dell'inventario, quella più 'tecnica', al volgare siciliano. Per render conto dell'ibridismo è stato adottato un criterio di inclusione linguistico accomodante, anche a seguito di indicazioni informali dei redattori del *TLIO*, includendo anche testi in cui il siciliano emergeva in frammenti contenuti come in Bresc/2014 (293) [365] . L'accettazione di inventari in cui l'infiltrazione del volgare su un testo a base latina crea un tessuto testuale intricato, come in Bresc/2014 (442) o, con una componente volgare più marcata, Bresc/2014 (447) è stata incoraggiata dal rilevamento, anche a una prima lettura superficiale, di lessemi non altrimenti attestati e non privi di interesse:

Arripizatu, agg. Rappezzato, raccomodato. Con questo stesso significato presente nel dialetto moderno, Cfr. *VS* s.v. *Arripizzatu*.

Buccatura, s.f. Imboccatura del fodero, att. solo in Bresc/2014 (442).

O come alcuni alterati che, in accordo con le norme del *TLIO*[366], danno vita a un'entrata autonoma:

Buttunellu, s.m. Piccolo bottone. Att. solo in Bresc/2014 (447).

Nello studio di Bresc-Bautier / Bresc (2014) gli spogli e le edizioni dei testi d'archivio sono il complemento testuale di un'indagine più ampia ed estensiva sugli oggetti nella Sicilia medievale. Il primo tomo, che precede quelli dedicati all'edizione,

[364] Bresc-Bautier / Bresc (2014: XXV).

[365] «Solution, largement adoptée dans la pratique notariale, est de rédiger l'acte en latin et de réserver une place centrale pour les parties techniques, toujours en sicilien», Ibidem.

[366] Beltrami (2013: 23).

disegna un quadro dettagliato (ma anche piacevolmente interessante e suggestivo) costruito sull'esperienza degli scavi archeologici del sito a Brucato e poggia sull'ispezione diretta degli oggetti materialmente rinvenuti. Il lavoro sulle parole è stato ulteriormente integrato con riferimenti agli scavi di Segesta, di Calathamet, di Iato, di Entella e di Palazzo Chiaramonte-Steri. All'opera va quindi riconosciuto, oltre ai meriti editoriali, il dettagliato quadro che riesce a fornire per la comprensione degli oggetti di uso quotidiano nella Sicilia del medioevo.

5.6 ANNOTAZIONI SULLA COSTITUZIONE DEL CORPUS

La codifica in GATTO dei nuovi testi non ha presentato particolari problemi[367]. Nessuno dei nuovi testi inseriti ha richiesto

---

[367] L'inserimento di MascalciaR2XVF ha generato un errore anomalo, localizzato, secondo il file di *output*, alla linea 18:

```
linea 18 posiz. 2  codice ANSI 46 caratt. "."
carattere non ammesso ad inizio riferimento organico
```

L'indicazione riportata da GATTO era palesemente fuoriviante in quanto alla linea 18 non si trovava nessun punto a inizio di riferimento organico. Eliminando l'intera linea, il programma continuava a restituire lo stesso errore ma anticipando il numero di riga. Eliminando una riga alla volta, partendo dalla prima, è stato possibile localizzare la causa dell'errore alla linea 24:

```
%8
# . v . @ di li infirmitati naturali et accedentali.
```

Sebbene il campo formula sia "destinato a contenere parti di testo che devono essere ignorate dal programma" il punto viene comunque tenuto in considerazione nel conteggio dei periodi. Si è quindi marcato il punto con il carattere "&", secondo quanto previsto dal manuale. Il malfunzionamento è stato segnalato all'Istituto Opera del Vocabolario (che peraltro era a conoscenza del bug).

la digitalizzazione delle edizioni di riferimento ma la versione per GATTO è stata realizzata partire da file digitali. Solo per i testi editi da Biondi (2001), i file forniti dall'editore hanno richiesto un confronto parola per parola con la stampa finale per integrare le modifiche subentrate in fase di correzione di bozze.

Per MascalciaR1XVF e MascalciaR2XVF, trattandosi di edizioni contenute in una tesi di dottorato[368], non è stata codificata la numerazione delle pagine e non si è tenuto conto delle righe. I testi sono stati ripartiti su due o più livelli di riferimenti organici[369] in accordo con l'organizzazione testuale dei due trattati, con un livello di complessità strutturale fino ad ora mai codificato nelle marcature di *Artesia*.

5.7 OSSERVAZIONI SUI CRITERI DI DATAZIONE

Dalla seconda metà del XX secolo, la lessicografia ha dimostrato una crescente sensibilità alla retrodatazione, già auspicata dall'importante intervento di Migliorini (1944-45)[370]. Lo sviluppo dei database, grazie alle grandi dimensioni e alla completezza dello spoglio, ha poi elevato le probabilità di rintracciare attestazioni sempre più antiche e oggi, a settant'anni di distanza dall'intervento di Migliorini e a dispetto dei successivi apporti del *DEI*, del *VEI*, del *DELI* e del *LEI*, le banche dati non mancano di autorizzare pratiche smaliziate, come per Parenti (2009: 223):

[368] Fichera (2015).

[369] Tre per MascalciaR2XVF.

[370] Ma si veda anche Migliorini (1956).

> È però doveroso aggiungere che la retrodatazione dei vocaboli, oltre a essere un «esercizio consapevolmente effimero» (come riconosce Cortelazzo 1989: 3), negli ultimi anni rischia di diventare un esercizio ozioso, perché chi è veramente interessato alla datazione di un termine sa che spesso è bene diffidare dei repertori e per maggior sicurezza ricava le informazioni autonomamente, soprattutto attraverso gli archivi elettronici, ogni giorno più ricchi e accessibili.

Più che all'apparente opposizione tra repertori e archivi elettronici, occorre badare alla demistificazione di quella concezione per la quale, «selon une vieille tradition de la lexicographie historique, la plus ancienne date est toujour considerée comme la meilleure date»[371]; l'obiettivo primario deve essere —per i dizionari ma anche e soprattutto per le banche dati — l'affidabilità dell'informazione. Lo stesso Migliorini (1944-45) lo aveva già intuito, con particolare riferimento ai falsi, e, con riferimento alla postdatazione, Zolli (1985) ha costruito un'interessante casistica dell'errore, dei «possibili trabocchetti per chi si avventura nell'irta foresta della lessicografia italiana[372]» di cui si dovrebbe adeguatamente tener conto. In informatica, la gestione delle date e la loro rappresentazione non presenta complicazioni tecniche — i più comuni database e linguaggi di programmazione hanno *tipi* di dati e *set* di istruzioni dedicate — ma logiche:

1. La definizione di precisi criteri per datare i testi e, di riflesso, le parole che attestano.
2. La predisposizione di un adeguato formato di rappresentazione nei metadati.

---

[371] Höfler (1986: 424).

[372] Zolli (1985: 152).

3. La definizione degli algoritmi per trattare le date. Questi non dovranno fare altro che di interpretare correttamente i metadati del punto 2 e trattarli coerentemente con i criteri stabiliti al punto 1.

I corpora storici presentano vasti gruppi di testi per i quali non si dispone di una data certa o precisa, opere per le quali la datazione è approssimata, incerta, o ricondotta a un periodo più o meno esteso e delimitato (con diverse modalità) da due o da un solo termine. Il trattamento computazionale deve affrontare il compito di "tradurre" i confini cronologici incerti in valori discreti e tra di loro equiparabili. Talvolta, datazioni che quotidianamente esprimiamo e comprendiamo senza sforzo mediante espressioni in linguaggio naturale, rivelano una certa complessità se trattate come concetti matematici.

Il passaggio al formalismo non è una conversione neutrale ma un'interpretazione che seleziona, tra le diverse scelte possibili, un modo particolare e in buona parte convenzionale per collocare il testo all'interno della collezione che lo contiene e per metterlo in relazione con tutti gli altri elementi del sistema diatestuale. Queste scelte condizioneranno i risultati delle interrogazioni, dalle concordanze alle analisi statistiche, dalle indagini filologiche a quelle lessicografiche. In altre parole, da criteri diversi in fase di *design* si otterranno risultati diversi in fase di consultazione.

I concetti temporali sfumati (es. "post 1320", "circa 1350", "inizi XV sec."), vanno perciò segnalati mediante attributi che ne precisano il grado di certezza e di precisione e/o ricondotti a un periodo delimitato. Le informazioni così prodotte andranno rappresentate in un formato che può essere gestito dal calcolatore. Gli algoritmi che si occuperanno di gestire queste informazioni dovranno essere in grado, a partire da una data formalizzata, di operare ordinamenti cronologici coerenti e stabilire l'appartenenza o la non appartenenza di un testo a un determinato segmento

temporale (rilevante anche per l'inclusione dei testi nei sottocorpora).

Artesia adotta gli strumenti dell'OVI, dove i testi sono catalogati utilizzando un doppio sistema: a una *data descrittiva* in forma discorsiva, viene affiancata una *data codificata* costituita da una stringa alfanumerica. La data codificata è incaricata di gestire l'effettivo ordinamento cronologico e può essere costruita dall'utente tramite una procedura guidata. Una volta inserita, la data codificata genera automaticamente altri due parametri, "anno iniziale" e "anno finale"[373]. Durante l'esecuzione di GATTO, i due valori agiscono in combinazione nel determinare la possibile inclusione di un testo nella definizione dei sottocorpora.

L'ordinamento cronologico dei testi e, di conseguenza, dei risultati delle ricerche in GATTO è basato solo sulla data finale. Gatto Web, invece, offre un'ulteriore funzionalità: l'ordinamento per data finale è un'impostazione predefinita ma è possibile selezionare l'anno iniziale nel criterio di ordinamento[374].

Negli algoritmi di GATTO si ritrova spesso una logica lessicografica prima che computazionale, visto che, nel circolo virtuoso tra corpus e vocabolario, la rappresentazione dell'*output* si riflette nella voce[375]. Dopo l'immissione dei metadati nella base di

---

[373] Nel caso di anno singolo i due campi vengono popolati con lo stesso valore.

[374] L'opzione "cronologico" è, per la versione *desktop*, l'unica scelta possibile per disporre i testi secondo la datazione. La scelta dei criteri di ordinamento è uno dei punti in cui GATTO e Gatto Web divergono nelle opzioni. In GATTO, una volta aperto il corpus sul quale lavorare, l'ordinamento può essere modificato selezionando "ordinamento" dalla barra dei menu. In Gatto Web, dal menu "Altre funzioni" è disponibile la voce "Modifica criterio ordinamento testi".

[375] «Ai fini della citazione in **0.3** [Prima attestazione], il primo testo è quello presentato come tale dalla banca dati in GATTO come risultato della ricerca

dati bibliografica, è il software a interpretare la data e gli intervalli di date e a formulare un ordinamento. L'algoritmo è trasparente per il redattore che si occupa solo di riportare nella voce quanto restituito delle interrogazioni. L'approssimazione "per eccesso" all'anno finale potrebbe quindi lasciare intravedere un atteggiamento lessicografico e computazionale prudente, dove a comandare non è la media aritmetica tra i termini estremi ma la garanzia che le forme e i lemmi del corpus non siano mai collocati al di là dell'orizzonte della certezza (dove il concetto di certezza va proporzionato all'affidabilità delle informazioni che ci consegnano le conoscenze attuali), una condotta scientificamente rigorosa ed estranea, se vogliamo, a facili compiacimenti da retrodatazione (che, peraltro, il *TLIO* non segnala).

La logica adottata deve essere resa esplicita e documentata adeguatamente. Ignorare o sottovalutare questi passaggi impedirebbe, sia ai curatori del corpus che agli utenti, un approccio teoricamente corretto ai dati e porterebbe a risultati incomprensibili (se non paradossali) o, peggio, a interpretazioni aberranti.

Ad esempio, per filtrare i testi del *Corpus Artesia* riferibili ai primissimi anni del '300, si potrebbe creare un sottocorpus selezionando ingenuamente l'intervallo 1300-1303 e interpretare la presenza tra i risultati di MilanaXVC come indicazione di antichità. In realtà, MilanaXVC viene visualizzato solo come conseguenza di una datazione generica (XIV sec.), che viene formalizzata come intervallo tra gli anni 1300 e 1399. Gli algoritmi di selezione dei testi sulla base di metadati riconoscono il testo come idoneo sulla base della sovrapposizione, anche minima, dei periodi considerati.

---

cumulativa di tutte le forme grafiche individuate per il lemma», Beltrami (2013a: 40).

Analogamente, senza una conoscenza di base dei meccanismi soggiacenti alla ripartizione dei testi attraverso i sottocorpora su discriminanti diacroniche, la sommatoria delle occorrenze rilevate in *cluster* temporali contigui porterebbe a risultati incomprensibili. Dati tre sottocorpora definiti su diversi intervalli temporali:

- Sottocorpus A: anno iniziale 1300 - anno finale 1349
- Sottocorpus B: anno iniziale 1350 - anno finale 1399
- Sottocorpus C: anno iniziale 1300 - anno finale 1399

Una scarsa conoscenza delle logiche di ripartizione temporale lascerebbe presupporre che la somma delle occorrenze complessive di A più le occorrenze complessive di B sia uguale alle occorrenze complessive di C, mentre il corpus Artesia restituisce:

- Sottocorpus A: 313.773 occorrenze (35 testi)
- Sottocorpus B: 450.794 occorrenze (147 testi)
  (A + B = 764.567)
- Sottocorpus C: 656.464 occorrenze (177 testi)
  (A + B > C)

Testi contrassegnati con intervalli a cavallo tra A e B vengono considerati come appartenenti a entrambi i sottocorpora, come avviene appunto per MilanaXVC.

Le applicazioni informatiche nelle scienze umanistiche hanno sempre evidenziato la necessità di rendere espliciti e quantitativamente precisi tutti i concetti con cui l'umanista si trovava a lavorare senza complicazioni di tipo quantitativo/formale. È perfettamente lecito collocare DialaguXIVS nella prima metà del XIV sec. Forse lo è un po' meno, ma tuttavia accettabile, riportare solo questo tipo di datazione nelle banche dati (come fanno peraltro *Artesia* e *CASVI*/*SALVIt*). Cosa succede però se convertiamo così com'è la data in linguaggio naturale in un intervallo codificato? Il programma continuerebbe a eseguire i suoi algoritmi con inesorabile precisione. Filtrando tutti i testi del 1301 e generando l'incipitario si potrebbe ottenere quindi:

> Al nome di Dio amen. Incomincia il libro che si chiama il "Dyalagho". Questa opera si è facta per mano di Frate Giovanne Campoli di Messina dell'ordine de frati minori, ad istantia e devotione di nostra segniora Madonna Alionora regina di Siciglia;

con evidente anacronismo nel riferimento alla regina Eleonora, che in quel periodo non era stata ancora promessa a Federico III (II) d'Aragona. Per l'accordo di matrimonio si dovrà attendere la pace di Caltabellotta, stipulata il 29 agosto 1302[376]. La datazione del *Corpus TLIO*, che collocava il testo nel periodo 1302-1337, è più coerente col dato storico (ma dal 7 agosto 2015 il testo è datato circa 2015[377]). Nella prossima edizione del *Corpus Artesia* sarà quindi auspicabile una revisione sistematica di tutti i dati bibliografici al fine di garantire una datazione più precisa. I risultati potranno essere profittevolmente integrati nella banca dati del portale.

Sia concesso poi, spostandoci su un terreno propriamente filologico, di soprassedere alle questioni relative alle edizioni e al valore lessicografico della variante, per il quale ci si limiterà a rimandare a Nencioni (1960: 187)[378], e di limitare le osservazioni ai

---

[376] Cfr. Kieswetter (1993).

[377] La scheda bibliografica di GattoWeb, inoltre, riporta un avviso importante: «Parole e brani in c.vo provengono da testimoni seriori (sec. XV) più o meno toscanizzanti (cfr. ed. p. xiii), e per questo motivo il testo c.vo va utilizzato con cautela come es. di mess. [vp 29/08/01]».

[378] «La buona edizione, e a *fortiori* l'edizione tecnicamente 'critica', tendono a certificare la lingua individuale degli autori, recuperandola, quando è il caso, dalle sviste o dalle manomissioni arbitrarie di copisti e stampatori. Ma quest'ultime, che per il filologo editore sono veri e propri guasti, per lo storico della lingua e per il lessicografo sono interpretazioni o, per tenersi in limiti più specifici, traduzioni nella lingua del copista, del tipografo o del correttore di bozze; la quale, benché sia anch'essa, a rigore, individuale, dovrà rassegnarsi, salvo il caso che quegli individui acquistino un rilievo singolare, a fungere da

nuovi problemi che sorgono se vogliamo dar conto dei testimoni, come intende fare programmaticamente *Artesia*:

> Mi preme sottolineare che nella costituzione del corpus ci si colloca, oltre che nella ovvia prospettiva dei testi, ove possibile, anche in quella dei testimoni; ciò significa che, per le opere che ci sono pervenute in più di un ms., *ARTESIA* intende editare l'intera tradizione o quanto meno i testimoni più significativi[379].

Di conseguenza, la data di un termine contenuto in un testo si frammenta tra la data del testimone e la data di composizione, e si dovrebbe a questo punto ammettere che:

> Normally is extremely difficult to be certain that a word appearing in a particular manuscript was definitely used (or conversely, was not used) by the original author. It's normally even more difficult to ascertain whether the spelling form in a particular manuscript reflects that the original author may have used[380].

Per questo motivo, una referenza dell'*OED* assume una forma del tipo:

> c1230 (▸?a1200) Ancrene Riwle (Corpus Cambr.) (1962)

Dove la prima data (c1230) è quella del manoscritto, la seconda, preceduta dal simbolo ▸, è la data di composizione, mentre la terza (1962) si riferisce all'anno di pubblicazione

---

testimonianza della lingua cosiddetta collettiva, cioè dell'uso linguistico del tempo e del luogo dove il manoscritto fu copiato o composto tipograficamente».

[379] Pagano (2009: 299).

[380] *Dating Middle English evidence in the OED*, in rete: <http://public.oed.com/aspects-of-english/english-in-time/dating-middle-english-evidence-in-the-oed/>, ultima consultazione: 29 luglio 2015.

dell'edizione. L'architettura informatica adottata da *Artesia* non consente però di rappresentare contemporaneamente due date e gestirle di conseguenza durante l'interrogazione del *Corpus*; i testi devono perciò essere associati univocamente associato a una sola data. Per l'*Istoria di Eneas* è stato pertanto indicizzato anche il ms. B, la cui edizione è stata approntata da Simona Spampinato all'interno dell'attività di ricerca dell'Università di Catania:

Ed. Folena:
EneasXIVF, Angilu di Capua, Istoria di Eneas, ms. A
Anno inizio: 1316. Anno fine: 1337.

Ed. Spampinato:

EneasXVS, Angilu di Capua, Istoria di Eneas, ms. B
Anno inizio: 1475. Anno fine: 1475.

La presenza di entrambe le edizioni restituisce quel confronto tra i due testi attraverso il quale «nella prospettiva diacronica si può leggere esattamente un secolo e mezzo di storia linguistica siciliana, proprio come nella sovrapposizione tra due radiografie si legge la storia di un caso clinico[381]».

Così, EneasXIVF, ripropone l'edizione critica di Folena, fondata sistematicamente sul ms. A, ma con qualche lezione di B accolta con riadattamenti alla norma linguistica dominante in A (Folena 1956: 262). La datazione segue, per l'ed. Folena, il criterio di base nel *Corpus*, quindi basato sulla data di redazione, sebbene sia stato appurato che tra l'originale e A — che ritorna, giustamente anche nel titolo abbreviato di GATTO — ci sia «una trafila che appare, allo stato del testo, piuttosto complessa[382]». EneasXVF,

[381] Folena (1956: LXI).

[382] Folena (1956: LVIII).

invece, riporta unicamente la data del manoscritto. L'intransigente approccio informatico/relazionale denuncerebbe un caso di incoerenza (dati non omogenei sullo stesso campo); con un giusto spirito pragmatico, il compromesso deve semplicemente essere reso esplicito e tenuto in dovuta considerazione, almeno fino a quando GATTO non permetterà la creazione di campi personalizzati nelle schede bibliografiche[383].

5.8 OSSERVAZIONI SUI METADATI DEI DOCUMENTI

La redazione dei metadati relativi ai nuovi inventari, lineare per quanto riguarda la datazione[384], ha però sollevato alcuni dubbi sulla compilazione del campo autore. Il confronto delle vecchie schede con le nuove ha sollevato alcuni dubbi sull'uso, forse troppo generico, del campo autore.

---

[383] La possibilità di introdurre campi nuovi non implica che tali campi siano suscettibili di influenzare la logica del programma.

[384] Per dovere di completezza andrebbe precisato che in diplomatica si rifiuta «il preconcetto dell'unicità assoluta della data, nonché la coincidenza necessaria del *datum* e dell'*actum*, distinguendo pure i diversi momenti che possono avvicendarsi nella fattura e nella spedizione di un documento. Un conto è il momento in cui è stata finita l'elaborazione di un documento, ed un altro conto è il tempo in cui un tale documento è stato consegnato al destinatario, oppure pubblicizzato» (De Lasala / Rabikauskas 2003: 36). Di più, siccome le date non si limitano agli anni dell'era cristiana, ma venivano aggiunte indizioni e anni di regno, l'ignoranza degli *scriptores* e guasti della tradizione aggiungono possibili perturbazioni (Pepe 1998: 38). Tuttavia si tratta di oscillazioni contenute e perfettamente trascurabili o di inesattezze da risolvere in altri contesti.

In alcuni casi il notaio è identificato con l'autore sulla base del registro di provenienza o del fondo dal quale proviene il testo. La pratica è poco rigorosa per più di un motivo.

Prima di tutto, non si tiene conto che la segnatura di un registro include eventuali fogli sciolti tra le carte che lo compongono, come nel caso del testamento della prostituta Rosa di Catania nei registri del notaio Pittacolis (Bresc-Bautier / Bresc 2014: 610) e nell'inventario del defunto Busacca Azaruti, a beneficio della vedova che intendeva recuperare la dote, allegato alle carte del notaio Randisi (Bresc-Bautier / Bresc 2014: 1488).

Non si può poi escludere che sui cartulari siano riportati testi di persone estranee allo studio notarile, come nell'inventario del monastero di San Salvatore[385], redatto in occasione della morte dell'abadessa da Perna della Rocca, alla presenza di nove sorelle con funzione di testimoni, in cui forse si potrebbe ipotizzare un divieto di ingresso, in un monastero femminile, per il notaio e i suoi assistenti.

Va ricordato poi che nel sistema del documento possono essere rintracciati almeno quattro tipi di soggetti:

1. L'*autore dell'azione giuridica* (*Urheber*), cioè colui che compie l'azione giuridica da cui nasce il documento.
2. L'*autore del documento*, cioè colui per ordine del quale o in nome del quale si scrive[386]. Può non coincidere con l'*autore dell'azione giuridica.*

---

[385] Bresc-Bautier / Bresc (2014: 740).

[386] Esistono anche punti di vista diversi. Per es. Bresslau, (1998: 11) non distingue tra *Aussteller* e *Urheber* e neppure tra autore e estensore e riduce, di fatto, i soggetti a due: «All'emissione di un documento partecipano nella maggior parte dei casi due persone o parti [autore e destinatario]. Chiamiamo autore (emittente, *Aussteller*) colui il quale richiede o dispone la redazione di un

3. Il destinatario o beneficiario (un acquirente o un erede).
4. L'estensore — o scrittore — cioè colui che esegue materialmente la stesura del documento[387].

Secondo una terminologia diplomatica per Bresc/2014 (352), ad esempio, si avrebbero 3 soggetti: il notaio Comito è l'estensore; l'autore dell'azione giuridica è Martino di Anselmo (dei beni del quale si redige l'inventario); l'autore del documento sarebbero i due fedecommissari che ne chiedono la redazione, cioè Bartolomeo di Maestro Antonio e maestro Giacomo di Ziza.

L'estensore può inoltre comprendere più individui reali: nelle cancellerie operavano più scrittori, a loro volta affiancati da vari collaboratori; in alcuni casi si può distinguere tra il *dictator*[388], o redattore del testo, e lo *scriptor*, o estensore materiale. La pluralità degli scrittori può essere qui ovviamente trascurata — così come il larghissimo ricorso ai formulari[389] — ma, anche solo per

---

documento, sia che egli abbia partecipato personalmente alla sua produzione, lo abbia egli stesso scritto o sottoscritto, oppure no. Il documento viene fatto risalire a lui, anche se egli ha soltanto dato l'incarico di redigerlo».

[387] De Lasala / Rabikauskas (2003: 44-45).

[388] A questa fase è riferibile la stesura di una eventuale minuta: «due sono soprattutto gli atti da considerare nel processo di produzione di un documento: la stesura del documento, compresa la confezione della minuta, nei casi in cui vi sia, e la preparazione della stesura a buono che il destinatario conserva come testimonianza. Il Medioevo definisce la prima attività con il termine *dictare* già in uso nel periodo romano tardoantico; coloro che svolgono tale attività sono i dettatori dei documenti. Per indicare il procedimento di confezione del *mundum* i manuali e le regole di cancelleria medioevali adottano spesso l'espressione *grossare*, *ingrossare* a causa dei caratteri più spessi e grandi usati; corrispondentemente noi parliamo di scrittori o grossatori dei documenti. È chiaro che molto spesso il dettato e l'ingrossatura di un documento derivano dalla stessa persona» (Bresslau 1998: 12).

[389] Si veda Pepe (1998: 45-48).

coerenza con il sistema di schedatura delle opere letterarie compresenti nell'archivio, occorrerebbe quantomeno valutare la distinzione tra autore e estensore, soprattutto in considerazione che «solo in casi rari gli autori di un documento lo stendevano o lo mettevano per iscritto personalmente[390]».

Eventuali regnanti (si consideri ad esempio Rinaldi/2005 (30) – 1364 Ordine di Federico IV) sono da considerare autori dell'azione giuridica e hanno una responsabilità diversa sul contenuto linguistico rispetto all'autore di un componimento in versi presente sullo stesso archivio testuale.

Il terzo problema è che il campo contiene, indiscriminatamente, singoli individui, istituzioni e denominazioni di soggetti collettivi.

Nell'inserimento dei metadati associati agli inventari di Bresc-Bautier / Bresc (2014) sono state rispettate le precedenti norme, ma, per le prossime edizioni sarebbe auspicabile per i casi enunciati una segnalazione tramite parentesi graffe. Per il momento si segnalano le referenze sulle quali sarebbe opportuna l'applicazione del nuovo criterio[391] nella versione 2016 del *Corpus*:

Bresc/2014 (201): R. Pittacolis
Bresc/2014 (247): G. Bonafede
Bresc/2014 (266): A. Zuccalà
Bresc/2014 (294): G. de Nava
Bresc/2014 (315): G. Traversa
Bresc/2014 (322): G. Bonafede
Bresc/2014 (346): A. Aprea
Bresc/2014 (352): G. Comito
Bresc/2014 (442): G. Comito
Bresc/2014 (447): A. Aprea
Bresc/2014 (450): G. Comito
Bresc/2014 (461): G. Traversa
Bresc/2014 (463): G. Comito
Bresc/2014 (465): G. Comito
Bresc/2014 (476): G. Saladino
Bresc/2014 (487): G. Traversa

[390] Bresslau (1998: 12).

[391] È stata anche creata una copia del database bibliografico con le parentesi già inserite, in vista di una possibile implementazione immediata nella futura versione di sviluppo del *Corpus Artesia 2016*.

Bresc/2014 (379): G. Comito
Bresc/2014 (402): G. Comito
Bresc/2014 (420): N. Maniscalco
Bresc/2014 (431): G. Traversa
Bresc/2014 (434): G. Traversa
Bresc/2014 (490): N. Aprea
Bresc/2014 (511): E. Pittacolis
Bresc/2014 (517): N. Grasso
Bresc/2014 (526): G. Vulpi

5.9 *CORPUS* SU CD-ROM: INTERFACCIA E DOCUMENTAZIONE

La valutazione di un sistema per la creazione e la gestione di corpora testuali — ma il principio può essere esteso a tutte le applicazioni informatiche per le scienze umane — deve tener conto di un ventaglio di caratteristiche eterogenee, da rintracciare in ogni singola fase della ricerca intrapresa, compresi i mezzi di pubblicazione dei risultati prodotti. A questo proposito, GATTO dispone di un'applicazione web che consente la pubblicazione e l'analisi on-line dei corpora, Gatto Web[392].

La controparte web — non autonoma, ma dipendente da GATTO per la creazione delle banche dati, per la compilazione dei metadati bibliografici e per la lemmatizzazione — oltre a valorizzare l'intero sistema e ad assolvere a un compito indispensabile per la maggior parte dei progetti di ricerca, ha il merito di allineare ai paradigmi attuali un sistema immaginato alla fine del secolo scorso, in un panorama tecnologico molto diverso da quello attuale, ancora estraneo alle applicazioni web. Grazie a Gatto Web basta un computer connesso a internet per consultare

[392] Iorio-Fili (2006).

liberamente la versione più aggiornata del *Corpus Artesia*, pubblicato regolarmente all'indirizzo <http://artesia.ovi.cnr.it>. Tuttavia, chi è abituato a utilizzare GATTO in locale nel proprio sistema trova spesso più comoda, veloce e stabile la versione *desktop* rispetto a quella on-line, soprattutto quando se ne fa un uso intensivo. GATTO inoltre conserva alcune funzioni in più nell'ambiente ricerche.

Parallelamente agli aggiornamenti regolari sul sito dell'Opera del Vocabolario, rilasciare periodicamente i file del *Corpus* per l'utilizzo con la versione desktop di GATTO, risponde alla volontà di permettere a chiunque di poter analizzare i testi della raccolta in locale, nell'ambiente software principale in cui il corpus viene creato e gestito.

Inoltre, com'è stato ampiamente discusso[393], una referenza sempre aggiornata, in continuo mutamento, è utile ma scientificamente sfuggente. Per questo motivo, dal momento in cui l'aggiornamento della versione su Gatto Web implica la rimozione delle versioni precedenti del *Corpus*, ai *Quaderni di Artesia* potrebbe spettare il compito di archiviare le versioni più significative del *Corpus*. Per questo principio il lavoro sul *Corpus Artesia 2015* è stato corredato con gli strumenti informatici di supporto alla pubblicazione su CD-ROM e dalla redazione della documentazione di supporto per l'interrogazione. L'interfaccia include un pacchetto di installazione guidata per il *Corpus*, il pacchetto di installazione di GATTO, e tutta la documentazione in formato elettronico. Una descrizione e gli *screenshot* di alcune schermate sono riportate in appendice.

Non sono stati ignorati i sei anni di distanza dalla prima pubblicazione su CD-ROM hanno visto la progressiva scomparsa dei lettori su molti portatili e computer *all-in-one* e sollecitano il

[393] §3.4.

superamento del supporto ottico come forma di pubblicazione esclusiva[394]. Pertanto ogni copia della prossima edizione includerà un codice per scaricare da internet la copia del *Corpus*.

La documentazione raccolta, una guida rapida all'utilizzo del *Corpus*, è invece riportata in appendice.

Per concludere, si potrebbe ribadire come la rappresentatività lessicale sia un obiettivo solo ideale, un «valore limite»[395], una chimera[396], e perseguire la rappresentatività di un corpus è un'attività di progressivo miglioramento. *Corpus Artesia 2015* che ha il merito, se non altro, di riprendere, dopo oltre quattro anni, la serie di aggiornamenti annuali del Corpus e ripristinare la continuità[397] interrotta, riesce in parte a colmare un'estesa lacuna la cui copertura era stata auspicata da più parti e in più occasioni[398], ma lancia alcune sfide da raccogliere per la prossima edizione, tra cui l'indicizzazione del *VSL* di Scobar e una completa revisione dei metadati.

---

[394] In controtendenza (ma l'esempio vale esclusivamente come ricordo affettuoso), erano le esplicite richieste di Alberto Varvaro per la redazione del *VSES*, il quale si rifiutava di scaricare le copie 'di servizio' aggiornate scaricabili per uso interno dal *server* di *Artesia*, e preferiva ricevere gli aggiornamenti masterizzati su CD-ROM e inviati per posta ordinaria.

[395] Lenci (2005: 40).

[396] Manning-Schütze (1999: 21).

[397] *Corpus Artesia 2008*, *Corpus Artesia 2009, Corpus Artesia 2010, Corpus Artesia 2011*, con una l'edizione su CD-ROM (Pagano 2008).

[398] Il problema del vuoto per i testi non letterari nel XV secolo è stato già sottolineato in Arcidiacono (2011: 286), Pagano / Arcidiacono (2013), Pagano (2012: 120).

6 VERSO IL VOCABOLARIO DEL SICILIANO MEDIEVALE

6.1 DAL *CORPUS ARTESIA* AL *VSM*

Un corpus informatizzato racchiude in potenza un insieme infinito di lessici più o meno complessi, alcuni dei quali generabili mediante procedimenti algoritmici. La definizione della popolazione e del campione, operata nelle prime fasi di *corpus design*, imprime alla collezione di dati lessicali una fisionomia propria, in virtù della quale taluni tra gli infiniti percorsi lessicografici possono sembrare implicitamente suggeriti, se non chiaramente orientati, dalla natura stessa della collezione di dati. Non sorprende quindi se l'idea di un *Vocabolario del Siciliano Medievale*, progetto auspicato già da Ruffino (1989: 337), di recente ripreso da Brincat (2011) e, parallelamente, da Pagano (2011) e Pagano (2012), fosse ben presente già prima del concreto avvio dei lavori per *Artesia – Archivio Testuale del Siciliano Antico*[399], come testimoniato dalle prime bozze dei testi per la homepage del *Portale*, puntualmente citate alla prima presentazione del progetto, in occasione del *V Convegno Internazionale Interdisciplinare su Testo, Metodo, Elaborazione Elettronica*[400].

Oggi, a parecchi anni di distanza da quella presentazione,

[399] <http://artesia.unict.it>; <http://artesia.ovi.cnr.it>. Per una presentazione del progetto si veda Pagano (2009). Alcune considerazioni sul campionamento operato per il *Corpus Artesia* in Arcidiacono (2011: 285-286).

[400] Pagano / Spampinato (2007).

l'idea del *Vocabolario* è in procinto di passare da ipotesi di lavoro a progetto, e non soltanto in virtù dei risultati conseguiti da *Artesia* o per la riflessione che non ha cessato di stimolare il gruppo di ricerca in questi anni (in particolar modo da quando sono state avviate le prime esperienze di lemmatizzazione[401]); si manifesta con crescente urgenza il bisogno di poter contare su una prospettiva definita che orienti le scelte e offra una prima impalcatura teorica e tecnica a sostegno delle problematiche lessicali indotte dal continuo sviluppo della base di dati.

Tra i primi risultati del processo pre-lessicografico [402] avviato, alcune prime coordinate sono rintracciabili in Pagano (2011) e Pagano (2012) e una prima versione di alcune delle acquisizioni presentate in questo lavoro si ritrova in Arcidiacono (2013).

A partire dal *Corpus Artesia* [403] viene stabilito un

---

[401] Si tratta di episodi, per uso interno, limitati ai due trattati di mascalcia tràditi dal ms. Riccardiano 2934 inclusi nella versione 2015 del *Corpus*. Per l'avvio di una lemmatizzazione si attende il rilascio della versione 4 di GATTO che implementerà, oltre alle tre modalità di lemmatizzazione classiche (*sequenziale su singolo testo*, *diretta su singolo testo* e *lemmatizzazione sul* corpus) un lemmatizzatore automatico (il cui uso è limitato alla lingua italiana moderna) e un lemmatizzatore semiautomatico, come documentato da Iorio-Fili (2007) e Iorio-Fili (2010). Per la descrizione di un sistema di lemmatizzazione esaustivo per lingue non standardizzate e un confronto con alcune soluzioni alternative si veda Gleßgen / Kopp (2005) (in merito a PHOENIX si veda anche Gleßgen 2006). Cfr. anche quanto riportato su LGeRM al §2.6.

[402] Atkins / Rundell (2008: 15-246).

[403] Fanno parte del *Corpus*, e di conseguenza della documentazione del *Vocabolario*, testi letterari, paraletterari e documentari appartenenti ad un arco cronologico compreso tra gli inizi del XIV secolo, periodo a cui risalgono i primi testi in volgare siciliano, e la prima metà del XVI, periodo in cui il siciliano è sostituito dal toscano come lingua dell'amministrazione. Per una descrizione aggiornata del *Corpus Artesia* si rimanda a Pagano / Arcidiacono (2013) e alla *Guida ai contenuti* (<http://artesia.ovi.cnr.it/HelpCorpora/Infoart.html>).

fondamentale punto di riferimento nel *Tesoro della Lingua Italiana delle Origini* (*TLIO*), modello programmatico per la concezione lessicografica e per gli strumenti informatici[404]. Al pari del *TLIO*, il *VSM* sarà infatti un vocabolario elettronico, in continuità con la vocazione digitale del progetto *Artesia*, nel quale mira a integrarsi organicamente. Il *VSM* si basa sulla banca dati elettronica del *Corpus Artesia*, e può contare sull'accelerazione dei processi di spoglio e sulla generazione automatica di indici, concordanze, formari e lemmari gestita dal software GATTO. Le successive scelte devono muovere da questa base di partenza e in sostanziale accordo col modello lessicografico del *TLIO*, ben sintetizzato nelle preziose *Norme per la redazione del* Tesoro della Lingua Italiana delle Origini[405].

L'aspirazione a collocare il *VSM* sulla linea tecnologicamente avanzata della lessicografia storica è da ricondurre alla necessità di disporre di metodi e strumenti per ottimizzare le risorse disponibili e assicurare la permanenza dei dati attraverso il tempo e l'intervento di possibili cambiamenti teorici e operativi, da cui la necessità di una riflessione preliminare: puntare sulla tecnologia come elemento strategico implica che la conoscenza (anche approssimativa) dell'entità dei vantaggi acquisibili sia preceduta da un serio studio di fattibilità.

La chiarezza di intenti del *VSM* si ridurrebbe a poca cosa, affidando entusiasmi e speranze a un generico concetto di dizionario elettronico, di per sé nebuloso: le agevolazioni auspicate potrebbero essere vanificate o, peggio, mutate di segno, laddove l'utilità degli esiti non riesca a bilanciare i costi richiesti

---

[404] Il *Corpus Artesia* è gestito dal software GATTO (Gestione degli Archivi Testuali del Tesoro delle Origini), programma realizzato da Domenico Iorio-Fili per il *TLIO*. Cfr. Iorio-Fili (2006), Iorio-Fili (2007) e Iorio-Fili (2012).

[405] Beltrami (2013a).

dall'informatizzazione, generando una perdita di efficienza generale.

6.2 AVVIO DEL FLUSSO DI LAVORO

Nella pianificazione di un'operazione lessicografica, il protrarsi dei tempi di redazione causa una variabilità, imputabile al fattore umano[406], che favorisce l'incongruenza e le imperfezioni; lo strumento informatico deve fungere da antidoto, forzando la consistenza e l'integrità referenziale anche attraverso intervalli temporali molto estesi.

Un *topos* ricorrente nella letteratura delle *digital humanities* è quello dell'esplicitazione rigorosa e la formalizzazione degli elementi e delle pratiche di ricerca come portato del trattamento algoritmico: «computing imposes a specificity which may at first seem alien to the more 'nuanced' in humanities scholarship[407]».
Tra gli infiniti esempi, Burdik et al. (2014: 14) riconducono il fenomeno a una prospettiva più ampia e radicale e ne ravvisano la propagazione degli effetti alle radici della conoscenza umanistica, nelle «premesse su cui poggiano tali concezioni» ampliandone la portata all'intera «gamma di spiegazioni in merito alla natura della conoscenza e del mondo, e alla capacità umana di determinare una valida forma di conoscenza»; chiaramente, la formalizzazione concorrerebbe a un cambio di paradigma [408] con un raggio infinitamente ampio.

Nella lessicografia computazionale non è data eccezione; il

[406] Trotter (2011: § 12).

[407] Hockey (2000: 2), cit. in Ciotti (2005: 9).

[408] Cfr. §1.3

primo lettore di un dizionario elettronico è il calcolatore: la macchina memorizza il vocabolario e a ogni consultazione lo rilegge e lo ricompila dinamicamente, prima ancora che il lettore umano possa visualizzarlo a schermo. Totalmente indifferente al significato e alla qualità delle informazioni, il computer è tra i destinatari più esigenti che il lessicografo possa incontrare per l'aspetto logico-formale, incapace com'è di soprassedere alla minima incongruenza e all'ambiguità referenziale. Nella lessicografia computazionale la forma diventa così un valore di primaria importanza e le imperfezioni emergono clamorosamente[409], non di rado con implicazioni problematiche.

> Le versioni elettroniche se, da un lato, esaltano quanto c'è di buono in un dizionario, liberandolo dalla camicia di forza dell'ordine alfabetico, facendo risaltare anche quando è nascosto tra le pieghe di una definizione lunga e complessa, dall'altro impietosamente rivelano all'occhio attento i difetti di un dizionario, le sue incongruenze, e sono quindi un ottimo strumento di igiene lessicografica, di allenamento per futuri lessicografi[410].

Nel processo senza fine della lessicografia scientifica un piano redazionale tecnologicamente dovrebbe inoltre recepire la flessibilità del digitale al fine di agevolare i lavori e pianificare gli obiettivi in maniera funzionale. Come si è visto[411] l'estremizzazione e la messa a sistema della continua modificabilità del dizionario elettronico conduce al cosiddetto paradigma evolutivo. Un flusso di lavoro che nasce con un carattere evolutivo ha la capacità di creare gli strumenti per distribuire le fasi di realizzazione in modo modulare e ottimizzare le risorse: focalizzando l'attenzione, di

[409] Cfr. Trotter (2011: §15) a proposito della digitalizzazione dell'*AND*.

[410] Marello (1996: 164).

[411] §3.4

volta in volta, su un singolo insieme circoscritto di problemi, la lessicografia evolutiva può procedere su segmenti contenuti del processo lessicografico e, per ciascuno di questi, produrre prodotti utilizzabili da subito, da integrare successivamente su un'architettura più ampia.

Affinché questo si realizzi è necessario che il flusso di lavoro sia in grado di affrontare e sfruttare le riconversioni tecniche senza traumi, e capace di rendere le informazioni (codificate in formati *standard*) non più semplici prodotti di funzioni di esportazione o garanzia solo teorica di trasportabilità, ma un baricentro vincolante, attorno al quale modellare tutti gli applicativi[412].

La persistenza dell'*hardware* e del software è realizzabile solo a breve termine, l'obsolescenza è un punto di arrivo per qualsiasi macchina e qualsiasi programma[413]. La stabilità richiesta da un approccio che miri a dirsi evolutivo andrà quindi cercata altrove. La proposta che qui si avanza è di spostare la struttura della voce al centro dell'architettura del progetto informatico per il *VSM*.

Per le prime prove di redazione, ma questo non esclude che del principio si possa fare una pratica di lessicografia evolutiva e politica editoriale, è stato stabilito di redigere voci di prova riferibili a insiemi di lessico compatti simili a quelli dei tradizionali quadri onomasiologici degli atlanti linguistici e, a questo scopo, occorrerà preparare il terreno per la redazione delle voci di prova. L'idea iniziale e il modello del *TLIO* vedono il *VSM* come dizionario semasiologico. Se le attività di redazione verranno

---

[412] Similmente a quanto auspicato da Leonardi (2013: 200) per i database.

[413] La proposta non deve lasciare intendere che la sensibilità ai dati può mettere in ombra l'attenzione alle procedure: la trasparenza del software e una critica sistematica degli strumenti informatici, così come auspicato da Gleßgen (2006: 16 e 20), rimangono ugualmente tra i desiderata più urgenti.

svolte, come proposto, per fasi discrete segnate da campi semantici di riferimento più o meno ampi, il vocabolario svilupperebbe anche una dimensione onomasiologica che, senza interferire con la macrostruttura di riferimento, può essere preservata nella struttura informatica tramite un'etichettatura non invasiva delle entrate, con un meccanismo di marcatori semantici simile a quello già visto sull'*OED*.

6.3 MICROSTRUTTURA

Il dizionario elettronico agisce sulla microstruttura con gli strumenti propri dell'informatica, ne rende espliciti forma e contenuto informativo, vi impone nuove regole, ne aumenta le potenzialità di organizzazione interna e accesso. La fisionomia della voce può uscirne stravolta, arricchita, e impaginata in modi nuovi, può adattarsi interattivamente alla consultazione o alle scelte dell'utente. Ma una voce rimane sempre l'unità fondamentale nell'organizzazione e nella consultazione del dizionario; anche nei casi più estremi che si possano immaginare, con informazioni aggregate a partire da risorse diverse e occorrenziari costruiti a *runtime*, è sempre possibile distinguere le voci tra di loro e tutti gli elementi che contengono. Forse, in un'opera che non ha più un inizio o una fine, lo stesso non potrà dirsi per la macrostruttura, riorganizzata a ogni aggiornamento del sito che la ospita e a ogni consultazione dei lettori che ne richiedono i contenuti parcellizzati secondo le richieste di interrogazione.

L'analisi formale sulla microstruttura, in un dominio di trattamento computazionale, è un esercizio propedeutico alle fasi successive di sviluppo. Sul piano dell'informazione, la microstruttura è la chiave d'accesso al modello di dati del sistema

informativo lessicografico[414]; analizzarla in termini computazionali significa portarsi al livello di astrazione nel quale tutte le voci del dizionario non sono che istanze di un oggetto composto da unità informative discrete che memorizzano tipologie di informazioni omogenee. Riemerge quindi quanto rilevato per i database, ma, a differenza del modello relazionale, gli elementi informativi sono organizzati secondo una particolare modalità gerarchica (anche ricorsiva). Da una parte riproducono la struttura della voce (lemma, etimo, significato, citazione, ecc.) ma dall'altra richiedono un'ulteriore scomposizione affinché ciascun punto della voce non includa sequenze indistinte di informazioni eterogenee impossibili da separare automaticamente.

> Il modello relazionale dei dati, attualmente il più diffuso, permette di definire tipi per mezzo del costruttore relazione, che consente di organizzare i dati in insiemi di record a stuttura fissa[415].

La struttura fissa è una gabbia troppo stretta per un vocabolario. Si pensi semplicemente agli elementi che ricorrono all'interno di una voce con frequenza variabile (forme grafiche, sensi, ecc.). Gli elementi si dovrebbero separare su più tabelle e combinare di volta in volta tramite *query*. La flessibilità dei formati di dati semistrutturati invece risolve alla radice alcuni gravi problemi logici nella rappresentazione formale del dizionario, dal momento che

> contrary to the situation in relational databases, the actual possible structures for such documents are not strictly limited to one single tree, but may vary according to design choices, for instance when one wants to iterate over senses in a

[414] Cfr. §3.1.

[415] Atzeni et al. (2002: 3).

dictionary, or make senses occur recursively within other senses[416].

La formalizzazione della microstruttura contribuirà inoltre alla realizzazione di un *modello formale* per dominare la complessità del trattamento computazionale, una prassi ben conosciuta da qualsiasi sviluppatore: «i programmi con una notevole estensione sono difficili da progettare, da modificare e da comprendere. Aumenteranno di molto le probabilità che abbiamo di realizzare un programma ben riuscito, se siamo in grado di creare un modello astratto relativamente semplice e quindi sviluppare il nostro sistema come attuazione pratica di quel modello[417]».

Si presenterà nelle prossime pagine un primo tentativo di definizione dalla microstruttura del *VSM*, identificando le informazioni pertinenti e individuando i relativi contesti di occorrenza di ciascun elemento. Si tenterà, parallelamente, una formalizzazione degli elementi di volta in volta individuati e la loro rappresentazione in XML/TEI.
L'esame dei marcatori permetterà di valutare e precisare un certo numero di informazioni lessicografiche, nell'intento di dotare il *VSM* di un modello che, per quanto volutamente provvisorio e necessariamente imperfetto, rappresenterà un ambiente di lavoro sul quale avviare i primi test di redazione e di una griglia informativa sulla quale basare lo sviluppo dei sistemi informatici per la redazione delle voci, l'archiviazione, la presentazione e l'interrogazione.

La provvisorietà programmatica del risultato proposto è un invito a procedere con una modellizzazione ciclica, articolata su

[416] Lemnitzer / Romary / Witt (2010).

[417] Grishman (1986: 5).

alcune fasi di sviluppo/test, in linea con un approccio evolutivo[418]: lo 'stile' di marcatura proposto prevede un progressivo irrigidimento dei vincoli di codifica (per quanto possibile) da una prima fase relativamente permissiva verso un graduale rafforzamento delle costrizioni formali, con l'obiettivo finale di raggiungere un *set* di regole estremamente precise e ben formalizzate[419] solo quando il *VSM* avrà raggiunto una maturità sufficiente.

Il delicato compito è apprezzabilmente agevolato dalla possibilità di avvalersi, con i dovuti riadattamenti, del raffinato modello di voce del *TLIO*; conformando il *VSM* allo schema elaborato dall'OVI si perviene al doppio risultato dell'adeguamento a una forma per la lessicografia italoromanza e all'acquisizione di un impianto lessicografico collaudato e di alto valore scientifico, impossibile da sviluppare autonomamente in tempi ragionevoli.

Se sarà conveniente mantenere minimo lo scarto con il *TLIO*, con divergenze limitate ai soli punti del *VSM* irriducibili al modello, una certa autonomia sarà conferita, almeno inizialmente, alla struttura tecnica, per i seguenti motivi:

- Contrariamente al rilascio degli strumenti informatici per la gestione dei corpora e alla pubblicazione del *Normario* per la redazione del *TLIO*, l'OVI non ha reso di pubblico dominio il software di gestione *server-side* del *Vocabolario* on-line.

- La redazione delle voci del *TLIO* viene effettuata su documenti di testo redatti con Microsoft Word per mezzo di soluzioni non *standard* (come l'utilizzo di particolari combinazioni di caratteri) sviluppate talvolta per funzionare

[418] Cfr. §3.4.

[419] *TEI Guidelines* (2013: xxxii).

in combinazione con i motori di indicizzazione per la messa in linea del *Vocabolario*.

- L'*editor* per la redazione del *TLIO* è ancora provvisorio. La versione definitiva è in fase di elaborazione[420].

6.4 REQUISITI DELLA MARCATURA

L'autonomo sviluppo di soluzioni per l'avvio dei lavori del *VSM* dovrà prevedere:

- La rigida separazione tra dati e presentazione degli stessi (da consegnare successivamente a uno o più fogli di stile o di trasformazione). Per la presentazione finale del prodotto sarà bene seguire le direttive del *Normario*, ma, a differenza del *TLIO*, non saranno usati elementi di formattazione durante la redazione del vocabolario.

- L'esplicitazione, tramite appositi marcatori, di tutte quelle parti di voce che sul *TLIO* sono segnalate da elementi tipografici o caratteri speciali.

- L'aderenza a *standard* internazionali.

Come si è visto, sono possibili infinite soluzioni in grado di soddisfare i primi due punti, ed esistono molte tecniche per adempiere anche al secondo, ma poche sono in accordo con l'ultimo: lo *standard* ISO 1951[421], il più recente Lexical Markup

[420] Cfr. Beltrami (2013a: 13, 21, 35, 66, 68).

[421] <http://www.iso.org/iso/catalogue_detail.htm?csnumber=36609>. Ultima consultazione: 26 ottobre 2015.

Framework (ISO ISO-24613:2008)[422] e il più noto XML/TEI[423]. Quest'ultimo può essere eletto a tecnologia di riferimento per le caratteristiche della marcatura potente e flessibile, per il prestigio accademico acquisito, per il numero e la qualità degli strumenti e delle risorse a disposizione (compresi gli studi disponibili e la manualistica tecnica) e per la volontà manifestata dall'OVI di orientarsi verso l'integrazione di questo *standard* nel proprio panorama tecnologico[424].

In questa fase di pianificazione, la formulazione di un primo impianto XML/TEI ha valore principalmente astratto per la costruzione dei dati e delle relazioni che intrattengono tra di loro. È del tutto irrilevante come saranno realmente redatti in futuro gli articoli (è auspicabile la realizzazione di una maschera di inserimento con scelte multiple, funzioni di autocompletamento e correzione automatica) e altrettanto irrilevante è il sistema col quale saranno archiviati e interrogati i dati: l'importante è che qualunque successivo strumento sia conforme a questa impalcatura formale e autodescrittiva.

6.5 LA VOCE

La voce del *VSM*, sul modello del *TLIO*, dovrà sempre riportare:

- Tutti i significati accertati, accompagnati dalla prima attestazione di ognuno di essi.

[422] Cfr. §4.3.

[423] Cfr. §4.2.

[424] Cfr. Iorio-Fili (2007).

- L'elenco delle forme presenti nel *Corpus Artesia*. Sulla piattaforma che ospita il TLIO, che viene eseguito sugli stessi server che gestiscono il *Corpus*, ciascuna forma è direttamente collegata a GATTO WEB per la ricerca automatica di tutte le occorrenze. *Artesia* non potrà contare su questo livello di integrazione, a meno che non si riesca a trovare una soluzione compatibile con le sessioni di navigazioni sulle quali Gatto Web è profondamente basato.

- L'etimologia, di solito costituita da un rinvio ai lessici.

- L'indicazione del testo più antico in cui nel *Corpus* viene attestato il lemma.

Rispetto al *TLIO* viene meno l'obbligo di inserire un campo per l'indicazione delle prime attestazioni del lemma per ogni varietà dell'italiano antico ma vanno mantenute invece le informazioni facoltative riguardanti l'uso in antroponimi e toponimi, i sinonimi e antonimi, i rapporti di derivazione e le note linguistiche.

6.6 L'ENTRATA

Il *VSM*, sul modello del *TLIO*, prevede entrate distinte sulla base delle distinzioni di categoria grammaticale, di etimo e di genere[425]. A causa dell'ampia varietà nella struttura dei dizionari esistenti, la Text Encoding Initiative ha previsto tre tipi di marcatori per segnare l'inizio e la fine dell'entrata lessicografica:

[425] Per i criteri e le eccezioni sull'entrata lessicale si rimanda a Beltrami (2013a: 19).

- *entry*: utile per la maggioranza dei dizionari convenzionali.

- *entryFree*: ha meno restrizioni dell'elemento <*entry*>, per permettere di codificare possibili dizionari atipici. Ove possibile è da preferire *entry* a *entryFree*.

- *superEntry*: riunisce una sequenza di entrate o un *set* di omografi.

Per gli omografi è prevista la possibilità di inclusione all'interno di un unico elemento <*entry*>:

```
<entry>
      <hom n="1">
            <sense n="1">
                  <!--…-->
            </sense>
            <sense n="2">
                  <!--…-->
            </sense>
      </hom>
      <hom n="2">
            <sense n="1">
                  <!--…-->
            </sense>
            <sense n="2">
                  <!--…-->
            </sense>
      </hom>
</entry>
```

Nel *TLIO* gli omografi costituiscono entrate differenti che «si distinguono con numeri tra parentesi prima della categoria grammaticale»[426]. Si riprodurrà quindi questa struttura anche nella marcatura, evitando quindi il *tag* <*hom*>, e si introdurranno tanti elementi <*entry*> quanti sono gli omografi, corredati dall'attributo *type* con valore *"hom"*.

[426] Beltrami (2013a: 22).

```
<entry n="531" type = "hom"><!--…--></entry>
<entry n="532" type = "hom"><!--…--></entry>
```

Le *Guidelines* suggeriscono di inserire la forma dell'entrata[427] direttamente tra le proprietà del *tag* <*entry*>[428], come per esempio avviene nella versione elettronica del Du Cange[429]:

```
<entry xml:id="PUTAGIUM" rend="ducange">
```

Questa soluzione tuttavia potrebbe condizionare la progettazione di un sistema organico di identificativi, in special modo nel caso di un dizionario nato in forma digitale. Più versatile la soluzione dell'*AND*[430]:

```
<entry type="main" key="janglure" id="AND-201-7A6EC4E5-
B1205CE-838CC89F-2AA9BAA2"status="61" lead="gdw">
```

Nell'elenco delle forme (di cui si parlerà in dettaglio nella prossima sezione) quella selezionata come vedetta avrà il suo elemento <*form*> che dovrà essere considerato, per così dire, il luogo di registrazione principale. Il lemma però viene riportato in

[427] Per la scelta della forma cfr. infra.

[428] «No restrictions are placed on the method used to construct xml:ids; one convenient method is to use the orthographic form of the headword, appending a disambiguating number where necessary. Identification codes are sometimes included on machine-readable tapes of dictionaries for in-house use», (*TEI Guidelines* 2013: 952).

[429] Du Cange s.v. *Putagium*.

[430] §2.5.

un attributo nel *tag* <*entry*>: per evitare di utilizzare un attributo non *standard* come *key*, si opterà per *n*[431]:

```
<entry n="acquistari" id="000">
      <-- ... -->
</entry>
```

Assegnando a ogni voce un numero progressivo, gli identificatori delle sezioni che compongono la voce saranno generati facendo seguire a questo numero una sequenza di cifre separate da punti, così da riprodurre la struttura gerarchica e la numerazione delle sezioni nella voce del *TLIO*[432]. La stringa identificativa comincerà con "VSM" perché il valore richiesto è di tipo NCName[433].

6.7 INTESTAZIONE DELLA VOCE (PUNTI 0.1-0.8)

6.7.1 Punto 0.1 - Forme grafiche

Il *tag* <*form*>[434] raccoglie le informazioni sulle forme

---

[431] «@n(numero) assegna un numero (o altra etichetta) a un elemento che non è necessariamente unico all'interno del documento», <http://www.tei-c.org/release/doc/tei-p5-doc/it/html/ref-att.global.html>, ultima consultazione 27 ottobre 2015.

[432] Per la numerazione potrà essere di qualche utilità l'utilizzo del *tag* <*head*>, come in Graiger (2006: 7), al quale si rimanda anche per altri esempi di marcatura.

[433] Cfr. <http://www.w3.org/TR/xmlschema-2/#NCName>, ultima consultazione: 5 dicembre 2015.

[434] *TEI Guidelines* (2013: 257): «Dictionary entries most often begin with information about the form of the word to which the entry applies. Typically, the orthographic form of the word, sometimes marked for syllabification or

riguardanti la voce corrente [435] . Per l'attributo *type* relativo all'elemento <*form*>, la Text Encoding Initiative suggerisce i seguenti valori:

- *simple* (single free lexical item).
- *lemma* (the headword itself).
- *variant* (a variant form).
- *compound* (word formed from simple lexical items).
- *derivative* (word derived from headword).
- *inflected* (word in other than usual dictionary form).
- *phrase* (multiple-word lexical item).

La forma a lemma avrà la seguente marcatura:

```
<form xml:id= "335.0.1.1" type= "lemma">
      <orth>acquistari</orth>
</form>
```

In un sistema dove il lemmatizzatore del corpus è integrato al dizionario e dove potrebbero esserci differenze tra lemma nel lemmatizzatore e lemma dell'entrata, si potrebbe usare un attributo

---

hyphenation, is the first item in an entry. Other information about the word, including variant or alternate forms, inflected forms, pronunciation, etc., is also often given».

[435]A differenza del *TLIO* non sono qui previsti i marcatori *a* per le forme fuori corpus e *x* per lo schedario xerografico.

type con valore *headword*, valore non suggerito ma attestato nelle *Guidelines*[436].

Col valore *variant* qui si intende una variante del lemma con la stessa flessione. Tutte le altre forme attestate saranno *inflected* (che si potrà quindi omettere o generare in automatico).

Andrà aggiunta una segnalazione per le eventuali forme non attestate del lemma che, per il momento, può essere rappresentata per mezzo dello stesso attributo *type* ma stavolta sull *tag* <orth>.

6.7.2 Sull'attestazione moderna

Un criterio generale, in una prospettiva evolutiva, dovrebbe essere di limitare per il momento la marcatura solo agli elementi strettamente funzionali.

L'utilizzo degli altri valori dovrà essere attentamente valutato in futuro, ma può qui trovare già applicazione la proposta di Pagano (2012: 131[437]) di integrare la struttura del *TLIO* con un campo *Attestazione Moderna*, mediante l'aggiunta di un valore personalizzato per la proprietà *type*. Grazie alla marcatura XML l'auspicato allestimento di una tavola di corrispondenza tra forme antiche e moderne potrà essere automatizzato. L'attestazione, dove possibile, potrà contenere un riferimento al corrispondente lemma

---

[436] Cfr. §2.6 sulla distinzione tra <ved> e <lem> nel Dictionnaire du Moyen Français.

[437] «Rispetto alla struttura della voce del *TLIO*, che prevede *Lista forme*, *Nota etimologica*, *Prima attestazione*, *Distribuzione geolinguistica*, *Note linguistiche*, *Note*, *Lista definizioni*, *Redattore*, c'è da riflettere se non sia opportuno aggiungere un campo *Attestazione moderna*, facendo ricorso alle voci del *VS*. L'ipotesi nasce dalla constatazione che nel *Corpus* vi sono lessemi con un'unica o con pochissime attestazioni, che, ben presenti nel siciliano moderno, devono essere stati certamente vitali nel corso dei secoli (ma per il *Corpus* valgano le variabili di ciò che ci è pervenuto e/o di ciò che ci è noto)», Pagano (2012: 131).

del *VS*. Per segnalare il rinvio, una prima proposta è di utilizzare, qui come in tutti gli altri luoghi in cui figurano referenze bibliografiche, un riferimento autosufficiente (<*bibl*>); in futuro, previo allestimento di una bibliografia integrata nel Vocabolario, si potrà effettuare una conversione automatica in puntatore bibliografico. Si avrà dunque:

```
<form xml:id="335.0.1.1.1" type="mod">
      <orth>acquistari</orth>

      <bibl>
            <title>VS</title>
            <biblScope>acquistari</biblScope>
      </bibl>

</form>
```

Corredando con il *tag* per le informazioni grammaticali <*gramGrp*>, con riferimento alla classificazione e alle sigle del *TLIO*[438]:

```
<entry n="531" key="acquistari">

      <gramGrp>
            <pos>v.</pos>
      </gramGrp>

      <form xml:id= "335.0.1.1" type= "lemma">
            <orth>acquistari</orth>
      </form>

      <form xml:id="335.0.1.1.1" type="mod">
            <orth>acquistari</orth>

            <bibl>
                  <title>VS</title>
```

[438] Per le quali si fa riferimento alla *Grammatichetta* di Esperti (1979) e alle successive modifiche dell'OVI.

```
                <biblScope>acquistari</biblScope>
            </bibl>

        </form>

        <form xml:id="335.0.1.1.2" type="variant">
            <orth>aquistari</orth>
        </form>

        <form xml:id="335.0.1.3" type="inflected">
            <orth>acquistirai</orth>
        </form>

        <form xml:id="335.0.1.4" type="inflected">
            <orth>acquistatu</orth>
        </form>

        <!--…-->

</entry>
```

6.7.3 Scelta della forma vedetta

Il lemma è un elemento liminare, tra testo e dizionario[439], tra macrostruttura e microstuttura, e sulla sua collocazione esatta non c'è accordo tra gli specialisti[440]. Può inoltre essere visto come una pura entità lessicografica o, soprattutto se lo si avvicina al lessema, come una manifestazione della teoria linguistica soggiacente. Secondo l'uso comune, l'entrata è data dal singolare per i sostantivi, dal maschile singolare per gli aggettivi e dall'infinito per i verbi.

---

[439] «La lemmatisation place le texte concordé à mi-chemin entre le texte lui-même (suite de formes graphiques et fait de discours ou performance) et le dictionnaire (suite d'entrées et fai de systeme ou Compétence)», Hanon (1990: 1564).

[440] Büchi (1996: 43).

Il *TLIO* adotta, così come l'*OED*[441], il principio della maggiore conformità possibile alla lingua moderna: «base generale per l'entrata lessicale è la forma moderna dell'italiano *standard*, con riferimento ai lessici principali (principalmente, nell'ordine, *GraDIt*, *DEncI*, *VLI*, *GDLI*[442])». La scelta è perfettamente giustificata per un vocabolario storico della lingua italiana con una prospettiva di lungo periodo di cui il *TLIO* rappresenta solamente la prima parte. L'impostazione di *Artesia* è però estremamente diversa.

Pagano (2012: 129-130) propone criteri diversi per la lemmatizzazione del *VSM*. La forma del lemma è da reperire all'interno di quelle attestate nel *Corpus Artesia* (o da ricostruire) al fine di mettere in evidenza anche gli elementi di alterità rispetto al siciliano moderno.

> Adottando questa prospettiva, sarebbe improprio, per es., per il sost. 'pace' lemmatizzare *paci* (32 occ.) e non *pachi* (335 occ.), dato che, al di là del dato quantitativo, <ch> per [tʃ] è un tratto pertinente della grafia del siciliano medievale; per l'agg. 'bianco' andrà lemmatizzato *blancu* (410 occ.) e non *biancu* (31 occ.), valendo, in questo caso, il dato quantitativo; per 'fanciullo', 'ragazzo', la forma del lemma dovrà essere *citellu*, dato che, tra le possibili alternative a <c>, [ts] non è mai rappresentato dai grafemi <ç>, <cz>, <z>.

Martin (1998) scompone la lemmatizzazione ideale in due fattori difficilmente conciliabili: la fedeltà all'uso dominante e la facilità di accesso:

---

[441] <http://public.oed.com/the-oed-today/rewriting-the-oed/editing-of-entries/> (ultima consultazione: 28 ottobre 2015).

[442] Beltrami (2013a: 20).

> L'expérience du *Anglo-Norman Dictionary*, qui retient pour chaque entrée la graphie la plus fréquente, conduit à une nomenclature représentative mais graphiquement disparate (les hasards des fréquences pour les mots peu usités entraînant d'inévitables inégalités de traitement) et, du fait même, peu maniable pour le consultant (qui n'a pas connaissance des fréquences au moment de la consultation).
>
> Dès lors que le lexicographe s'efforce de systématiser la matière pour en faciliter l'accès, sa Nomenclature devient un artefact sans garantie philologique: ainsi pour T-L[443], de consultation aisée, mais dont les entrées, systématiquement reconstruites, sont loin de correspondre toujours aux graphies effectivement attestées[444].

Da queste considerazioni, in opposizione alla soluzione adottata dall'*AND*, Martin elegge la facilità di consultazione a principio guida del *DMF*, da far prevalere sulla "fedeltà illusoria" delle attestazioni grafiche (rimandando al corpo della voce per le grafie effettive) e limitando la funzione degli elementi della nomenclatura a «des étiquettes commodes pour l'accès à l'information[445]».

Un primo risultato dello schema TEI qui formulato è di rendere di facile risoluzione il conflitto lessicografico rilevato da Martin, e invalidare la mutua esclusione tra «*usage dominant*» e «*aisé à l'information*», incorporando in un unico nodo XML le informazioni utili: puntando il motore di ricerca sull'elemento <*form*> di tipo "*lemma*", ed estendendo la ricerca a tutti i nodi figlio (di tipo "*orth*" e "*mod*"), si consentirà di utilizzare un

---

[443] Sull'eccessiva normalizzazione del Tobler-Lommatzsch, si veda anche Trotter (2013: 664).

[444] Martin (1998: 971).

[445] Ibidem.

lemmario corrispondente a forme realmente attestate e conformi all'uso grafico del siciliano medievale, annullando l'incertezza derivante dall'oscillazione delle grafie, preservando la piena compatibilità con il modello del *TLIO* e migliorando l'esperienza di ricerca più di quanto possa fare un lemmario concepito come insieme di "etichette comode[446]".

Non va trascurata inoltre la possibilità di etichettare e, di conseguenza, gestire adeguatamente le forme ricostruite mediante un attributo personalizzato applicato al *tag* <*orth*>.

6.7.4 Punto 0.2 - Etimologia

Il *VSM*, così come il *TLIO*, riporta l'etimo con «la funzione di identificare il lemma rispetto a omografi e omofoni reali o virtuali [447] ». Nel nostro caso le direttive del *Normario* vanno integrate con l'adozione del *VSES* come punto di riferimento privilegiato. Per i derivati da lemmi volgari previsti dal *VSM* si rinvierà ad altre voci per mezzo del puntatore <*ref*>.

> In general, however, normal modern bibliographic practice, and these Guidelines, distinguish between a bibliographic reference, which is a self-sufficient description of a bibliographic item, and a bibliographic pointer, which is a short-form citation (e.g. Baxter, 1983) which serves usually as a place-holder or pointer to a full long-form reference found elsewhere in the text. The usual encoding of short-form references such as Baxter, 1983 is not as <bibl> elements but as cross-references to such elements[448].

Non è ancora prevista la possibilità di registrazione di

[446] Si ricordi anche la funzione del marcatore <VED> nel *DMF* nello *snippet* al §2.6.

[447] Beltrami (2013a: 33).

[448] *Tei Guidelines* (2013: 116).

prefissi e suffissi (che molto probabilmente andrà rimandata a un'altra fase), per la quale è comunque possibile utilizzare lo spazio previsto per eventuali annotazioni e rinvii.

<*ref*> sarà qui utilizzato nei casi di participi e sostantivi derivati da aggettivi, per i quali si seguiranno le indicazioni contenute nel *Normario* del *TLIO*.

Lo schema segnalato da Beltrami (2013a: 33)

> Lingua etimo (Lessico s.v. Voce) || Eventuali annotazioni e rinvii

integrato con quanto stabilito in precedenza a proposito dei riferimenti bibliografici, può essere rappresentato come segue:

```
<etym xml:id="L335.0.2">

      <lang>Lat.</lang>
      <mentioned>Adventare</mentioned>

      <bibl>
            <title>VSES</title>
            <biblScope>Abbintári</biblScope>
      </bibl>

      <note></note>
</etym>
```

6.7.5 Punto 0.3 - Prima attestazione

Nel *TLIO* la prima attestazione è indicata attraverso l'abbreviazione bibliografica, seguita dal rinvio alla prima definizione sotto la quale il testo si trova citato.

Analogamente a quanto riscontrato al punto 0.2 [449], la bibliografia dei testi inseriti nel *Corpus Artesia* non è integrata nel

[449] §6.7.4.

Vocabolario. Un legame relazionale su questo punto ridurebbe la ridondanza dei dati bibliografici, replicati su numerose risorse del progetto: la voce, il *Corpus*, il *Portale* (alle voci "Bibliografia Testi" e "Bibliografia Documenti" nella sezione "Banca Dati") e, con coincidenza solo parziale, le voci della "Banca dati" del *Portale* contrassegnate dall'icona di GATTO. Si converrà che lo scopo che qui ci si prefigge non si estende a tale livello di generalità e che la gestione organica di tutte le risorse del progetto non può e non deve condizionare la progettazione microstrutturale. È altresì auspicabile che il dizionario in XML contenga tutte le informazioni minime per la corretta consultazione e che dovrà essere indipendente dalle altre risorse. Il file XML potrà contenere alcuni riferimenti alle entità informative del sito o del *Corpus* per garantire l'integrità referenziale, ma la progettazione del sistema risiede su un altro livello di analisi.

Sulla base di quanto affermato, si ricorrerà per il momento al marcatore autosufficiente <*bibl*>, da riconvertire in puntatore una volta realizzato un sistema dinamico per l'integrazione e la condivisione dei dati. La coerenza dell'impianto può essere garantita utilizzando la sigla GATTO come identificativo univoco[450]. Il titolo abbreviato verrà visualizzato come referenza sintetica, in accordo con le norme editoriali della bibliografia del progetto. Sebbene l'abbreviazione del titolo contenga anche un chiaro riferimento all'autore, verrà inserito un campo <*author*> per semplificare successive operazioni di trattamento automatico.

```
<bibl type="primAtt" xml:id="335.0.3" >
```

[450] Anche il *TLIO* prevede l'inserimento della sigla come campo nascosto per favorire il trattamento automatico dei dati. Si veda a tal proposito Beltrami (2013a: 109).

```
      <idno type="siglaGATTO">ENE</idno>[451]
      <author>Angilu di Capua</author>
      <title>EneasXIVF - Angilu di Capua, Istoria di
      Eneas, ms. A</title>
      <biblScope>Lxxx.1.5</biblScope>

</bibl>
```

Punto 0.4 - Tipo di attestazione della voce

La sezione 0.4, marcata attraverso il *tag <dictScrap>*, registra le seguenti informazioni:

- Attestazione unica nel corpus:

```
<dictScrap xml:id="L335.0.4">
      Att. unica nel Corpus.

</ dictScrap>
```

- Attestazione in un solo testo:

```
<dictScrap xml:id="L335.0.4">

      Att. solo in

      <bibl type="attTxtUnico"xml:id="L335.0.3">
            <idno type="siglaGATTO">ENE</idno>
            <author> Angilu di Capua</author>
            <title>EneasXIVF - Angilu di Capua,
            Istoria di Eneas, ms. A</title>
      </bibl>

</dictScrap>
```

- Attestazione solo nelle opere di un unico autore:

```
<dictScrap xml:id="L335.0.4">
```

[451] L'identificativo riporta l'abbreviazione del titolo così come inserito nella base di dati, così come suggerito da Beltrami (2013a: 109).

```
        Att. solo in
        <bibl>
              <author>Angilu di Capua</author>
        </bibl>

</dictScrap>
```

Si tralasciano i casi particolari e i commenti (punto 04 N[452]) la cui marcatura non presenta particolari difficoltà. Per i casi di documentazione ripartita in più voci, così come descritto in Beltrami (2013a: 46), si farà uso del *tag* <*ref*>.

Punto 0.5 - Note linguistiche

Si tralasciano, per il momento, le note linguistiche, per le quali si rimanda a Beltrami (2013a: 51); si propone invece di segnalare, col marcatore <*dictScrap*>, l'elenco delle poliremetiche, che sul *TLIO* viene inserito in coda a questa sezione e distinto dalle eventuali note linguistiche col solo ausilio di un rientro tipografico. Ogni entrata riporterà in coda il riferimento (<*ref*>) alla relativa definizione.

Punto 0.6 – Annotazioni varie

Il punto 0.6 raccoglie un insieme eterogeneo di blocchi informativi, contrassegnati da lettere, la cui marcatura dipenderà da future scelte redazionali:

A: Antroponimi

T: Toponimi

O: Scheda onomasiologia

D: Rapporti di derivazione

[452] Beltrami (2013a: 44).

V: Rinvii ad altre voci

N: Annotazioni

Punto 0.7 – Riepilogo della struttura della voce

Il punto 0.7 del *TLIO* contiene un riepilogo delle definizioni di primo e secondo livello[453]. Dal *Normario* si deduce che a questa sezione è associata anche la funzione dell'indicizzazione del dizionario[454].

Sul metodo di indicizzazione del *VSM* è ancora prematuro esprimersi. Per ridurre la ridondanza dei dati, abbassare la possibilità di introdurre refusi e snellire il file XML il punto 0.7 potrebbe essere omesso, consegnando alla marcatura delle definizioni il compito di compilarlo automaticamente in futuro.

Punto 0.8 – Firma del redattore

Con molta probabilità la firma del redattore sarà inserita all'interno della voce per mezzo del marcatore <*respons*>. Tuttavia le proprietà da attribuire all'elemento e ulteriori accorgimenti per tracciare la cronologia delle modifiche saranno determinati, più che dall'impianto lessicografico o dalla struttura tecnica, dalla concreta organizzazione delle procedure di redazione e revisione.

7 I SIGNIFICATI

> How is a lexicographer to know what *any* word means? How are the public features of word menaning and to be identified? I suggest that at least three components are necessary for the construction of an accurate account of the

[453] Cfr. infra.

[454] «Per consentire ricerche nel vocabolario, il punto 0.7 deve essere compilato anche per voci di estensione minima, in cui appaia inutile» (Beltrami 2013a: 60).

> meanings and conventional functions of a word. The first is a body of evidence — citations, indices, concordances to a corpus, and so on. The 200econdi s the personal knowledge or intuitions about word meanings which native speakers have, although these are notoriously difficult ti access directly. The third is the body of statements (true or false, accurate or inaccurate, as the case may be) to be found in existing dictionaries, grammars, and other language studies[455].

Nella lessicografia di prima mano l'individuazione del significato deve sempre partire dall'interpretazione dei contesti di occorrenza individuati per ciascun lemma:

> L'individuazione e l'articolazione dei significati devono basarsi sull'esame delle attestazioni nel corpus […] e non sui lessici, ai quali si ricorre solo per controllo[456].

Sulla scorta di Beltrami (2013°: 62), il marcatore <*sense*> riproduce la struttura dei significati di un vocabolario anche con impieghi ricorsivi. La successione è stabilita dal redattore secondo due criteri: partendo, se possibile, dal significato più vicino all'etimo e procedendo per successivi scostamenti, secondo la scala che va dal significato "proprio" al "figurato" o dal significato «più prossimo al valore referenziale» a significato "esteso" o "traslato". Privilegiando l'impostazione "a cascata" piuttosto che quella "ad albero"[457], come sul *TLIO*, la sezione 0.1 presenterà una marcatura del tipo:

```
<sense xml:id="L335.1.1">
```

---

[455] Hanks (1990: 34).

[456] Ivi, p. 62.

[457] Il *TLIO* prevede la possibilità di voci 'ad albero' per locuzioni o espressioni fraseologiche ma, parallelamente, sconsiglia questo metodo, raccomandando di citare le polirematiche sotto le diverse definizioni.

```
Significato A
<!-- ... -->

<sense xml:id="L335.1.1.1">
      Significato B
      <!-- ... -->
</sense>

<sense xml:id="L335.1.1.2">
      Significato C
      <!-- ... -->
</sense>

<sense xml:id="L335.1.2">
      Significato D
      <!-- ... -->
</sense>

</sense>
```

Le citazioni andranno segnalate con i tag <quote> e <cit>, con l'utilizzo dell'attributo "n" per la numerazione progressiva all'interno del significato.

Prima della definizione (*tag* <ù>) vanno collocate le marche grammaticali, semantiche e d'uso. Le sigle impiegate saranno mutuate dalla Tabella delle abbreviazioni del *TLIO*[458].

Le marche grammaticali saranno segnalate da <*gramGrp*>, e collocate all'interno del *tag* <*pos*>, con l'eventuale ausilio di <*subc*> nei casi in cui occorre specificare l'uso del verbo. Marche semantiche e marche d'uso impiegheranno il *tag* <*usg*>, opportunamente differenziato dall'attributo *type* (con valore "*style*" per le marche semantiche e "*dom*" per le marche d'uso):

[458] La tabella è accessibile all'indirizzo <http://tlio.ovi.cnr.it/TlioAbbr>. È possibile autenticarsi con nome utente "guest" e lasciando il campo password vuoto. Ultima consultazione: 7 dicembre 2015.

```
<usg type="dom">anat.</usg>

<usg type="style">fig.</usg>
```

I commenti aggiunti alla fine della definizione possono essere racchiusi da un *tag* <*note*>.

Il sistema per la registrazione delle poliremetiche andrà trattato separatamente, vagliando, tra l'altro, l'opportunità di creare una tabella (come per il *TLIO*) opportunamente collegata alle voci, se non generata dinamicamente. Per il momento è lecito chiedersi se non sia utile segnalare la presenza di una poliremetica all'interno di un significato per mezzo del marcatore <*formtype* = "*phrase*">, eventualmente corredato da ulteriori attributi che consentano di specificare meglio il tipo di entità immessa nella sezione.

7.1 PER UN SISTEMA DI REDAZIONE ASSISTITA DAL CALCOLATORE

A conclusione del lavoro è stato sviluppato un prototipo di interfaccia per l'inserimento guidato delle voci e la compilazione automatica del file XML/TEI.

Per consentire una redazione collaborativa, con eventuali strumenti di gestione del flusso di lavoro e divisione dei compiti all'interno della redazione, ma soprattutto per mantenere alta l'integrazione con il sistema di pubblicazione on-line, la maschera di inserimento guidato è stata concepita come un modulo di un sistema più ampio per la gestione del vocabolario ed è stata implementata direttamente su piattaforma LAMP (Linux [459] ,

[459] È il sistema operativo del *web server* che gestisce l'applicazione.

Apache[460], MySQL[461], PHP[462]). Chiaramente le tecnologie sono perfettamente trasparenti e all'utente finale basterà semplicemente accedere al sistema tramite un browser, collegato a internet, a una rete *lan*[463] o alla piattaforma LAMP eseguita in locale sullo stesso computer.

Anche se l'applicazione ha attualmente complessità e dimensioni contenute, è stata impiantata un'architettura adeguata a sostenere la futura realizzazione di un sistema più articolato, tramite l'uso di librerie distribuite su più livelli di astrazione e di una programmazione orientata agli oggetti[464].

Una libreria di supporto (identificata dal nome *Entèca_db*) — che contiene essenzialmente un oggetto per le connessioni al database — crea un'interfaccia tra l'applicazione e il database MySql. La maggior parte delle classi del sistema derivano invece dalle classi astratte[465] contenute nella libreria *Entèca*; gli oggetti di questa classe instanziano la classe contenuta in *Entèca_db* nel proprio costruttore. Una terza classe (*Ninive*) gestisce i

---

[460] È il *web server*, software che si incarica di gestire le richieste di trasferimento delle pagine e dei dati e dell'interpretazione del linguaggio PHP.

[461] <https://www.mysql.it/>. MySQL è un sistema di gestione per database relazionali libero, sviluppato per essere il più possibile conforme agli standard ANSI SQL e ODBC SQL.

[462] PHP (acronimo ricorsivo di "PHP: Hypertext Preprocessor") è un linguaggio di programmazione interpretato, usato principalmente per realizzare pagine web dinamiche.

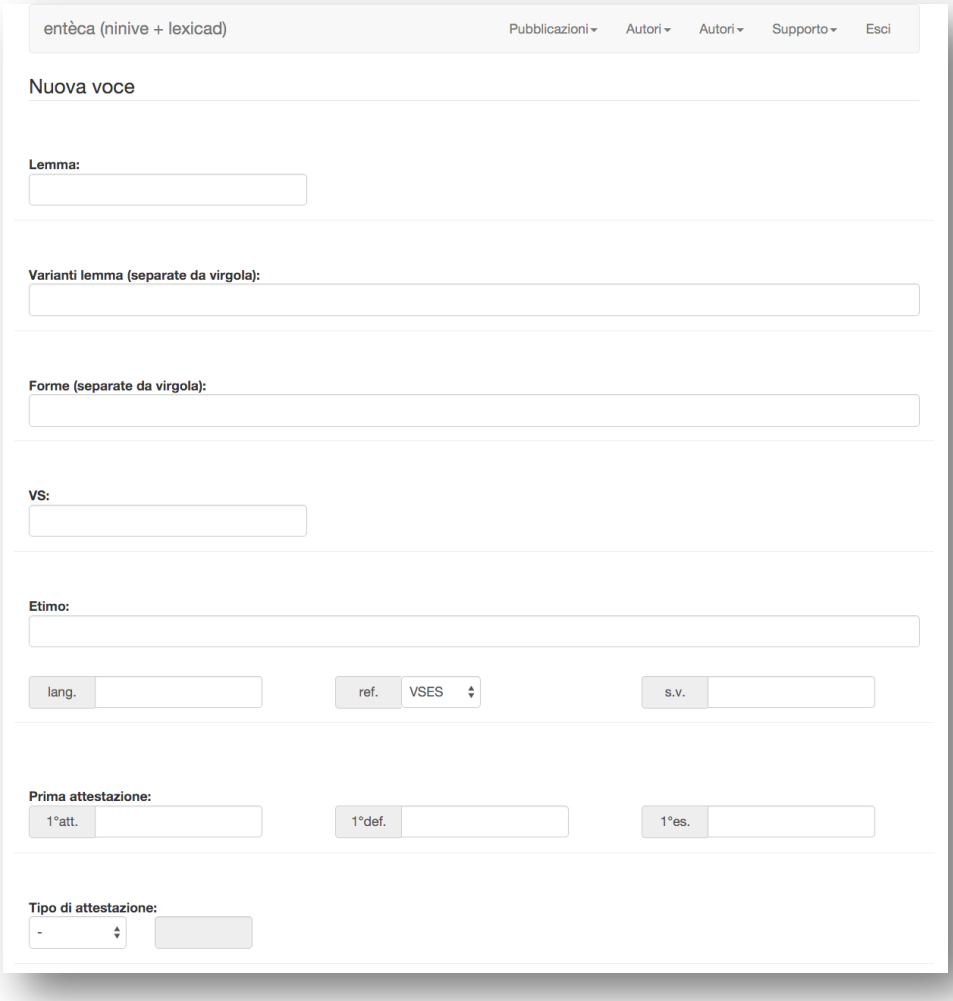
[463] Local Area Network (LAN), rete di computer locale.

[464] La programmazione orientata agli oggetti (OOP, *Object Oriented Programming*) è un paradigma di programmazione (o paradigma) che si basa principalmente sul raggruppamento all'interno di un'unica entità (la classe) delle strutture dati (proprietà) e delle procedure (metodi) che su queste operano.

[465] Nella programmazione orientata agli oggetti, le classi astratte definiscono una classe senza implementarla concretamente.

comportamenti della base di dati bibliografica; tutte le classi qui contenute sono derivate dalle classi di *Entèca*. I componenti direttamente connessi alla redazione delle voci fanno invece riferimento a un terzo modulo (*LexiCad*) che contiene i file e le classi per la gestione delle funzioni lessicografiche..

La base del sistema di redazione è quindi realizzata attraverso un modulo di *input,* o *form*, che tipicamente assume la forma rappresentata nella figura sottostante.



Il *form* è stato realizzato con un *design* responsivo[466], cioè in grado di adattarsi perfettamente allo schermo del dispositivo sul quale viene visualizzato (dagli *smarthpone*, ai *tablet*, ma anche su computer con *monitor* molto grandi o estremamente definiti), mantenendo un alto livello di usabilità e leggibilità. Sugli schermi piccoli, ad esempio, i campi si dispongono singolarmente in sequenza verticale, come nella maggior parte delle *App*, dedicando la maggior parte dello spazio disponibile al singolo campo sul quale si sta lavorando.

CSS[467] e Javascript[468] assistono la compilazione. A un primo livello, le due tecnologie operano rimodulando parzialmente la maschera nel corso dell'uso, bloccando o abilitando i campi a seconda delle scelte (per es., nella sezione *Tipo di attestazione*, selezionando il valore "Att. solo in" viene sbloccato un campo per la selezione del testo; una volta sbloccato, il campo diventa obbligatorio nella validazione) e creandone di nuovi (come per l'aggiunta di un significato).

Tramite Ajax[469], invece, viene interrogata una tabella del database contenente l'archivio bibliografico (sostanzialmente una riconversione in MySql del file di bibliografia, così come realizzato da GATTO, della più recente versione del *Corpus Artesia*). Un

---

[466] Per accelerare lo sviluppo, ma anche per mantenere al minimo la scrittura di codice relativo alla grafica, è stato utilizzato Bootstrap un *framework* HTML, CSS e Javascript per il *front-end web development*.

[467] I CSS (Cascading Style Sheets) sono i fogli di stile con i quali si definisce la presentazione dei file HTML, XHTML e XML.

[468] JavaScript è un linguaggio per la programmazione di *script* lato *client*. Javascript è orientato agli oggetti e agli eventi, ed è usato per implementare comportamenti dinamici e interattivi all'interno delle pagine web.

[469] Ajax (Asynchronous JavaScript and XML), è una tecnica di sviluppo software che permette uno scambio di dati in *background* fra *browser* e *server* e l'aggiornamento di parti della pagina senza doverla aggiornare interamente.

comportamento simile dovrà essere riprodotto per la futura funzione di aggiunta di un rinvio (la maschera interroga in tempo reale il *Vocabolario* e crea un elenco di voci sulle quali effettuare la selezione). Nella maggior parte delle interazioni col database, *Ajax* lavora in tempo reale, creando suggerimenti istantanei e funzioni di autocompletamento capaci di accelerare la stesura delle voci.

A intervalli di tempo regolari, inoltre, Ajax raccoglie silenziosamente i dati fino a quel momento inseriti e ne effettua un *backup* in un database temporaneo[470]. Il *backup* in automatico è uno stratagemma per permettere un buon livello di affidabilità alle applicazioni basate su piattaforma web.

Al termine dell'inserimento i campi vengono singolarmente verificati dal sistema (validazione) e, se l'operazione va a buon fine, vengono passati a PHP.

A questo punto le informazioni inviate dal *form* vengono raccolte e convertite in XML/TEI secondo lo schema descritto negli ultimi paragrafi. Il modulo di conversione potrà essere usato in futuro per creare una funzione di anteprima in tempo reale del file XML.

A questo punto il sistema crea il file XML e lo archivia in una cartella sul *server*, effettuando anche una copia di *backup* dell'intero contenuto su una tabella del database; i dati vengono parallelamente trattati singolarmente per un parallelo inserimento "disaggregato" su un'ulteriore tabella del database.

---

[470] Attualmente la funzione è sviluppata solo approssimativamente. Il sistema converte i dati in XML/TEI ed effettua un *backup* grezzo in una tabella dedicata del database. Il *backup* viene effettuato su tutti i campi, anche nel caso in cui non siano state fatte modifiche dal momento dell'ultimo backup. Lo scopo della funzione, fino a questo momento, è semplicemente quello di poter recuperare i dati in caso di disconnessioni della linea. Per evitare la chiusura accidentale della pagina, è stato creato uno *script* in Javascript che chiede conferma se si tenta di abbandonare la pagina in presenza di modifiche non salvate.

Al termine dell'inserimento viene visualizzato il contenuto XML e, successivamente, una versione impaginata dell'articolo.

Allo stato attuale il sistema mira ad essere solamente un prototipo dimostrativo e mancano, come ovvio, molte delle funzioni basilari per poter parlare di un sistema completo per la compilazione e la gestione di un vocabolario on-line. La sua vera funzione è però quella di suggerire una strada concreta per realizzare concretamente quanto descritto finora.

8 BIBLIOGRAFIA

Sigle *Artesia*:

Bresc/2014 (201) = Bresc-Bautier / Bresc (2014: 652-653).

Bresc/2014 (247) = Bresc-Bautier / Bresc (2014: 739-740).

Bresc/2014 (266) = Bresc-Bautier / Bresc (2014: 786-787).

Bresc/2014 (293) = Bresc-Bautier / Bresc (2014: 825-826).

Bresc/2014 (294) = Bresc-Bautier / Bresc (2014: 826-827).

Bresc/2014 (298) = Bresc-Bautier / Bresc (2014: 830-834).

Bresc/2014 (315) = Bresc-Bautier / Bresc (2014: 864-875).

Bresc/2014 (322) = Bresc-Bautier / Bresc (2014: 896-898).

Bresc/2014 (346) = Bresc-Bautier / Bresc (2014: 991-995).

Bresc/2014 (352) = Bresc-Bautier / Bresc (2014: 1004-1007).

Bresc/2014 (379) = Bresc-Bautier / Bresc (2014: 1084).

Bresc/2014 (393) = Bresc-Bautier / Bresc (2014: 1130-1132).

Bresc/2014 (402) = Bresc-Bautier / Bresc (2014: 1159-1161).

Bresc/2014 (420) = Bresc-Bautier / Bresc (2014: 1202).

Bresc/2014 (431) = Bresc-Bautier / Bresc (2014: 1227-1228).

Bresc/2014 (434) = Bresc-Bautier / Bresc (2014: 1235-1235).

Bresc/2014 (442) = Bresc-Bautier / Bresc (2014: 1269-1274).

Bresc/2014 (447) = Bresc-Bautier / Bresc (2014: 1286-1297).

Bresc/2014 (450) = Bresc-Bautier / Bresc (2014: 1303-1304).

Bresc/2014 (461) = Bresc-Bautier / Bresc (2014: 1328-1331).

Bresc/2014 (463) = Bresc-Bautier / Bresc (2014: 1332-1337).

Bresc/2014 (465) = Bresc-Bautier / Bresc (2014: 1344-1345).

Bresc/2014 (476) = Bresc-Bautier / Bresc (2014: 1376-1382).

Bresc/2014 (487) = Bresc-Bautier / Bresc (2014: 1458-1460).

Bresc/2014 (490) = Bresc-Bautier / Bresc (2014: 1479-1484).

Bresc/2014 (511) = Bresc-Bautier / Bresc (2014: 1556-1561).

Bresc/2014 (517) = Bresc-Bautier / Bresc (2014: 1586-1591).

Bresc/2014 (526) = Bresc-Bautier / Bresc (2014: 1216-1217).

Conf3XVB = Branciforti (1953: 154-178).

DeclarusXIVM = Marinoni (1955).

DialaguXIVS = Santangelo (1933).

EneasXIVF = Folena (1956).

EneasXVS = Spampinato (2002).

MilanaXVC = Cusimano (1951: 87-91).

Rinaldi/2005 (1) = Rinaldi (2005: 8-13).

Rinaldi/2005 (19) = Rinaldi (2005: 22-23).

Rinaldi/2005 (30) = Rinaldi (2005: 87).

Rinaldi/2005 (61) = Rinaldi (2005: 146-151).

Rinaldi/2005 (75) = Rinaldi (2005: 165-166).

Sardina/2006 (8) = Sardina (2006: 442).

SonLibArbXIVC = Cusimano (1951: 36).

ValMaxXIVU = Ugolini (1967).

_____

AA. VV., 1985. *La Crusca nella tradizione letteraria e linguistica italiana, Atti del Congresso internazionale per il IV centenario dell'Accademia della Crusca (Firenze 29 settembre - 2 ottobre 1983)*, Firenze, Accademia della Crusca.

—, 1992. *Bibliografia dei testi in volgare fino al 1375 preparati per lo spoglio lessicale*, Firenze, Opera del Vocabolario Italiano.

—, 2011. *Reinventing reseach? Information pratices in the humanities*, RIN (Research Information Network), in rete: <http://www.rin.ac.uk/system/files/attachments/Humanities_Case_Studies_for_screen_2_0.pdf>. Ultima consultazione: 09 settembre 2015.

—, 2014. *Una lingua e il suo vocabolario*, Firenze, Accademia della Crusca.

Altenberg, Bengt / Sylviane Granger, 2002. *Lexis in Contrast: Corpus-based Approaches*, Amsterdam-Philadelphia, John Benjamins Publishing.

Antonelli, Roberto, 2011. *Premessa*, in AA.VV. *LirIO. Corpus della Lirica Italiana su CD-ROM*, I, Firenze, Edizioni del Galluzzo (*Archivio romanzo*, 20 – *Lirica europea*, 4), pp. VII-VIII.

Arcidiacono, Salvatore, 2011. «Risorse per lo studio dei volgarizzamenti in siciliano: il Corpus Artesia», in Sergio Lubello (a cura di) *Volgarizzare, Tradurre, interpretare nei secc. XIII-XVI*, Strasbourg, Eliphi – Éditions de linguistique et de philologie (Bibliothèque de Linguistique Romane, 8), pp. 283-293.

—, 2013. «Percorsi di lessicografia computazionale per un Vocabolario del Siciliano Medievale (VSM)», in *Bollettino del Centro di studi filologici e linguistici siciliani*, 24, pp. 87-108.

Atkins, Sue, 1992. «Putting Lexicography on the Professional Map: Training Needs and Qualifications of Career Lexicographers», in AA. VV. *EURALEX '92 Procedings*, Tampere, Tampereen Yliopisto.

Atkins, Sue / Jeremy Clear / Nicholas Ostler, 1991. «Corpus Design Criteria», in *Literary and Linguistic Computing*, 7/1, pp.1-16.

Atkins, Sue / Michael Rundell, 2008. *The Oxford Guide to Practical Lexicography*, Oxford, Oxford University Press.

Atzeni, Paolo et al., 2002. *Basi di dati. Modelli e linguaggi di interrogazione*, Milano, McGraw-Hill.

Avalle, D'Arco Silvio, 1979. *Al servizio del vocabolario della lingua italiana*, Firenze, Accademia della Crusca.

—, 1985a. «I Canzonieri: definizione di genere e problemi di edizione», in *La critica del testo. Problemi di metodo ed esperienze di lavoro, Atti del Convegno di Lecce (22-26 ottobre 1984)*, a cura di Enrico Malato, Roma, Salerno Ed., pp. 363-82.

—, 1985b. «Lessicografia dei testi antichi», in AA.VV. 1985.

Barbera, Manuel, 2009. *Schema e storia del* Corpus Taurinense, Alessandria, Edizioni dell'Orso. In rete: <https://archive.org/details/ManuelBarberaSchemaEStoriaDelCorpusTaurinense>. Ultima consultazione: 4 dicembre 2015.

—, 2013. *Linguistica dei corpora e linguistica dei corpora italiana. Un'introduzione*, Milano, Qu.A.S.A.R.
Ed. elettronica, da cui si cita: <http://www.bmanuel.org/man/cl-HOME.htm>. Ultima consultazione: 14 ottobre 2015.

Barbera, Manuel / Elisa Corino / Cristina Onesti, 2007. *Corpora e linguistica in rete*, Perugia, Guerra Edizioni. In rete: <http://archive.org/details/BarberaCorinoOnestiCorporaELinguisticaInRete>. Ultima consultazione: 14 ottobre 2015.

Barbi, Michele, 1935. «Crusca, lingua e vocabolari», in *Pan*, III/9, pp. 13 – 24. Poi in Barbi / Pasquali / Nencioni 1957, pp. 7-35, da cui si cita.

Barbi, Michele / Giorgio Pasquali / Giovanni Nencioni, 1957. *Per un grande vocabolario storico della lingua italiana*, Firenze, Sansoni. [Rist. anastatica: Firenze, Le Lettere, 2012].

Beccaria, Gian Luigi, 1994. *Dizionario di linguistica e di filologia, metrica, retorica*, diretto da G.L. Beccaria, Einaudi, Torino.

Beltrami, Pietro, 2005. «Il "Battaglia" visto dal cantiere del "Tesoro della Lingua Italiana delle Origini"», in Gian Luigi Beccaria / Elisabetta Soletti (a cura di), *La lessicografia a Torino dal Tommaseo al Battaglia. Atti del Convegno (Torino-Vercelli, 7-9 novembre 2002)*, Alessandria, Edizioni dell'Orso, pp. 309-321.

—, 2008. «Nuova lessicografia dell'italiano antico: il Tesoro della Lingua Italiana delle Origini», in *Bollettino dell'Atlante Lessicale degli Antichi Volgari Italiani*, 1, Pisa-Roma, Fabrizio Serra.

—, 2009a. «The Lexicography of Early Italian: its Evolution and Recent Advances», in Bruti / Cella / Foschi Albert 2009, pp. 29-53. In rete: <https://www.academia.edu/6233012/Pietro_G._Beltrami_The_Lexicography_of_Early_Italian_its_Evolution_and_Recent_Advances>. Ultima consultazione: 17 ottobre 2015.

—, 2009b. «Past and Present of Italian Historical Dictionaries», in *«Quo vadis lexicography?», 6th working meeting of German language Academy dictionaries*, *Berlin 2.-5. September 2009*, pp. 1-13. In rete: < http://dwb.bbaw.de/tagung09/pdf/Beltrami.pdf>. Ultima consultazione: 2 novembre 2015.

—, 2010a. «Lessicografia e filologia in un dizionario storico dell'italiano antico», in *Storia della lingua e filologia: Atti del convegno ASLI (Pisa-Firenze, 18–20 dicembre 2008)*, Firenze, Cesati, pp. 235–248.

—, 2010b. «Textual criticism and historical dictionaries», in *Variants*, 10, pp. 41-59.

—, 2013a. *Norme per la redazione del Tesoro della Lingua Italiana delle Origini, versione aggiornata 2013*, in rete <http://tlio.ovi.cnr.it/TLIO /NormeTLIO.pdf>. Distribuita il 25/07/2013; ultima consultazione: 29 gennaio 2014; prima in *Bollettino dell'Opera del Vocabolario Italiano*, III.

—, 2013b. «Theory of Dictionary Management», in Gouws et. al. 2013, pp. 524-530.

Beltrami, Pietro / Simone Fornara, 2004. «Italian Historical Dictionaries: from the Accademia della Crusca to the Web», in *International Journal of Lexicography*, Vol. 17, No. 4.

Bergenholtz, Henning, 2011. «Access to and Presentation of Needs-Adapted Data in Monofunctional Internet Dictionaries», in Fuertes-Olivera / Bergenholtz 2011, pp. 30-53.

Biber, Douglas, 1986. «Spoken and Written Textual Dimensions in English: Resolving the Contradictory Findings» in *Language*, 62, pp. 384-414.

—, 1988. *Variation across Speech and Writing*, Cambridge, Cambridge University press.

—, 1990. «Methodological Issues Regarding Corpus-based Analyses of Linguistic Variation» in *Literary and Linguistic Computing*, Vol. 5 (4), pp. 257-69.

—, 1993. «Representativeness in Corpus Design», in *Literary and Linguistic Computing*, vol. 8 (4); ora in W. Teubert and R. Krishnamurthy (Eds.), *Corpus linguistics: Critical concepts in linguistics* (Vol. II), London, Routledge, pp. 134-165. In rete: <http://otipl.philol.msu.ru/media/ biber930.pdf>. Ultima consultazione: 7 ottobre 2015.

Biffi, Marco, 2009. «Accademia della Crusca's Online Dictionaries» in Bruti / Cella / Foschi Albert 2009, pp. 240-285.

—, 2013. «Un nuovo sito web per l'Accademia della Crusca», in Maraschio / De Martino / Stanchina 2013, pp. 127-137.

—, 2014. «La lessicografia della Crusca in rete», in AA.VV. 2014, pp. 113-127.

Biondi, Clara, 2000. «Vita quotidiana e cultura materiale a Scicli», in *Siculorum Gymnasium*, LIII.

Blum, Claude, 2002. *Frédéric Godefroy. Dictionnaire de l'ancienne langue française et de tous ses dialectes du IX^e^ au XV^e^ siècle. Édition électronique*, Paris, Champion.

Blumenthal, Peter / Achim Stein, 2002. *Tobler-Lommatzsch: Altfranzösisches Wörterbuch. Elektronische Ausgabe*, Stuttgard, Franz Steiner Verlag (4 CD-ROM e manuale di istruzioni).

Bonavita, Massimo, 2008. «Effetto farfalla», in *Enciclopedia della Scienza e della Tecnica,* Roma, Treccani.

Branciforti, Francesco, 1953. *Regole, costituzioni, confessionali e rituali*, a cura di F. Branciforti, Palermo, Centro di studi filologici e linguistici siciliani (Collezione di testi siciliani dei secoli XIV e XV, 3).

Bresc, Henri, 1995. «Une maison de mots: inventaires palermitains en langue sicilienne (1430-1456)», in *Bollettino del Centro di studi filologici e linguistici siciliani*, 18, pp. 109-187.

Bresc-Bautier, Genevieve / Henri Bresc, 2014. *Une maison de mots. Inventaires de maisons, de boutiques d'ateliers et de châteaux de Sicile (XIIIe-XVe siècles)*, Palermo, Associazione Mediterranea.

Brincat, Giuseppe, 2011. «Per un vocabolario del siciliano antico: l'apporto dei documenti di Malta», in Gruppo di ricerca dell'Atlante Linguistico della Sicilia (a cura del), *Per i linguisti del nuovo*

millennio. Scritti in onore di Giovanni Ruffino, Palermo, Sellerio, pp. 304-311.

Braga, Daniele / Alessandro Campi / Stefano Ceri, 2005. «XML: rappresentare e interrogare dati semi-strutturati», in *Mondo Digitale*, 2, pp. 45-58. In rete: <http://archivio-mondodigitale.aicanet.net/Rivista/ 05_numero_tre/ Braga,_Ceri_p._45-58.pdf>. Ultima consultazione: 14 ottobre 2014.

Bresslau, Harry, 1998. *Manuale di diplomatica per la Germania e l'Italia*, Roma, Ufficio Centrale per i Beni Archivistici (Sussidi, 10). In rete: <http://www.archivi.beniculturali.it/dga/uploads/documents/Sussidi/Sussidi_10_1_a.pdf>. Ultima consultazione: 25 ottobre 2015.

Bruni, Francesco, 1980. «La cultura e la prosa volgare nel '300 e nel '400», in *Storia della Sicilia*, Società editrice Storia di Napoli del Mezzogiorno continentale e della Sicilia, Palermo, vol. IV, pp. 179-279.

Bruti, Silvia / Roberta Cella / Marina Foschi Albert, 2009. *Perspectives on Lexicography in Italy and Europe*, Cambridge, Cambridge Scholars.

Buchi, Eva, 2013. «La lessicografia storica condotta dall'ATILF: ancoraggio lessicologico, complementarietà interna e internazionalità crescente», in Maraschio / De Martino / Stanchina 2013, pp. 3-10. *Preprint* in rete: <https://halshs.archives-ouvertes.fr/halshs-00910113/document>. Ultima consultazione: 19 ottobre 2015.

Buchi, Eva / Pascale Renders, 2013. «Gallo-Romance I: Historical and Etymological Lexicography», in Gouws et al. 2013, pp. 653-662. *Preprint* in rete: <https://halshs.archives-ouvertes.fr/halshs-00421893>. Ultima consultazione: 10 novembre 2015.

Budin, Gerhard / Stefan Majewski / Karlheinz Mörth, 2012. «Creating Lexical Resources in TEI P5» in *Journal of the Text Encoding Initiative*, 3. In rete: <http://jtei.revues.org/522 ; DOI: 10.4000/jtei.522>

Burdick et al. 2012. *Digital Humanities*, Cambridge (MA) - London, MIT Press. In rete: <https://mitpress.mit.edu/sites/default/files/titles/content/97802 62018470Open_Access_Edition.pdf>. Ultima consultazione: 4 ottobre 2015; Trad. it. da cui si cita, *Umanistica Digitale*, Milano, Mondadori, 2014.

Burnard, Lou / Michael C. Sperberg-McQueen, 2005. *Il manuale TEI Lite. Introduzione alla codifica elettronica dei testi letterari*, a cura di Fabio Ciotti, Milano, Edizioni Sylvestre Bonnard.

Busa, Roberto, 1951. *Sancti Thomae Aquinatis hymnorum ritualium varia specimina concordantiarum. Primo saggio di indici di parole automaticamente composti e stampati da macchine IBM a schede perforate*, Milano, Bocca.

—, 1990-1991. «Préface», in Cignoni / Peters, vol. II, pp. IX-XIII.

—, 1992. *Sancti Thomae Aquinatis opera omnia cum ipertextibus*, Milano, Editel (in CD-ROM).

—, 2005. *Index Thomisticus, edizione web* a cura di E. Bernot e E. Alarcón, 2005, online: <http://www.corpusthomisticum.org/it/>. Ultima consultazione: 3 ottobre 2015.

Bush, Vannevar, 1945. «As We May Think», in *The Atlantic*, 146 (July), pp. 101-108. In rete: <http://www.theatlantic.com/magazine/archive/1945/07/as-we-may-think/303881/>. Ultima consultazione: 3 ottobre 2015.

Buzzetti, Dino / Leonardo Quaquarelli, 1995. «Informatica umanistica», in *Schede Umanistiche*, n.s. 1995, 1.

—, 2014. *Osservazioni critiche dell'AIUCD sull'ASN*, in rete: <http://www.roars.it/online/osservazioni-critiche-dellaiucd-sullasn/>. Ultima consultazione: 22 ottobre 2015.

Casapullo, Rosa, 1995. «Bibliografia dei testi siciliani dei secoli XIV e XV», in *Bollettino del Centro di studi filologici e linguistici siciliani*, 18, pp. 13-34.

Castells, Manuel, 2002. *Galassia Internet*, Milano, Feltrinelli.

CD-ROM Multilingue = *CD-ROM Multilingue - CD-Rom multilingual dictionary database. Chinese, Dutch, English, French, German, Italian, Japanese, Spanish*, Bologna, Zanichelli, 1987.

Chauveau, Jean-Paul / Buchi, Eva (2011). «État et perspectives de la lexicographie historique du français», in *Lexicographica. International Annual for Lexicography*, 2011, 27, pp.101-122.

Chiari, Isabella, 2006. «Performance Evaluation of Italian Electronic Dictionaries: user's needs and requirements», in Corino / Marello / Onesti 2006, pp. 141-146. In rete: <http://www.euralex.org/elx_proceedings/Euralex2006/018_2006_V1_Isabella%20CHIARI_Performance%20Evaluation%20of%20Italian%20Electronic%20Dictionaries_Users%20Need.pdf>. Ultima consultazione: 2 novembre 2015.

—, 2007a. *Introduzione alla linguistica computazionale*, Roma-Bari, Laterza.

—, 2007b. «Dizionari elettronici italiani in glottodidattica», in Monica Barni / Donatella Troncarelli / Carla Bagna (a cura di), *Lessico e apprendimenti. Il ruolo del lessico nella linguistica educativa, Atti del XIV Convegno Nazionale GISCEL*, Milano, Franco Angeli, pp. 227-233.
In rete: <http://www.alphabit.net/PDF/Chiari_GISCEL2006_impaginato.pdf>. Ultima consultazione: 17 ottobre 2015.

—, 2012. «Il dato empirico in lessicografia: dizionari tradizionali e collaborativi a confronto», in *Bollettino di Italianistica. Per Tullio De Mauro*, II, pp. 94-125.
In rete: < http://www.alphabit.net/PDF/Pubblicazioni/2012_chiari_

dato_empirico_lessicografia.pdf>. Ultima consultazione: 17 ottobre 2015.

Chomsky, Noam, 1962. «A Transformational Approach to Syntax», in Archibald A. Hill (ed.), *Proceedings of the Third Texas Conference on Problems of Linguistic Analysis in English, May 9–12, 1958*, Austin, University of Texas Press.

Cignoni, Laura / Carol Peters (eds.), 1990-1991. *Computational Lexicology and Lexicography*, Pisa, Giardini (*Linguistica Computazionale*, *Special Issue dedicated to Bernard Quemada*, VI-VII), 2 voll.

Ciotti, Fabio, 1995. «Testi elettronici e banche dati testuali. Problemi teorici e tecnologie», in *Schede Umanistiche*, n.s., 1995, 2, pp. 147-148.
In rete: <http://disi.unitn.it/~poesio/Teach/IU/Testi_elettronici_e_banche_dati_testuali.pdf>. Ultima consultazione: 12 ottobre 2015.

Colistra, Vincenzo, 2008. «Ubiquitous computing», in *Enciclopedia della Scienza e della Tecnica*, Roma, Treccani. In rete: < http://www.treccani.it/enciclopedia/ubiquitous-computing_(Enciclopedia_ della_Scienza_e_della_Tecnica)/>. Ultima consultazione: 1 dicembre 2015.

Corino, Elisa / Carla Marello / Cristina Onesti (a cura di), *Atti del XII Congresso Internazionale di Lessicografia, Torino, 6-9 settembre 2006 - Proceedings of the XII EURALEX International Congress*, Torino, Italia, 6th-9th September 2006, Alessandria, Edizioni dell'Orso (2 voll.).

Curti, Luca, 1972. «Antichi testi siciliani in volgare», in *Studi mediolatini e volgari*, 20, pp. 49-139.

Cusimano, Giuseppe, 1951. *Poesie siciliane dei secoli XIV e XV*, a cura di G. Cusimano, I, Palermo, Centro di studi filologici e linguistici siciliani (collezione di testi siciliani dei secoli XIV e XV, 1).

De Blasi, Nicola / Francesco Montuori, «Per un dizionario storico del napoletano», in Emanuela Cresti (a cura di), *Prospettive nello studio del lessico italiano. Atti SILFI 2006*, Firenze, FUP, Vol. 1, pp.85-92.

DEI = Carlo Battisti / Giovanni Alessio, *Dizionario etimologico italiano*, Firenze, Barbèra, 1950-1957.

DELI = Manlio Cortelazzo / Paolo Zolli, *Dizionario etimologico della lingua italiana*, Bologna, Zanichelli, 1979 e sgg.

De Robbio, Antonella, 2007. *Archivi aperti e comunicazione scientifica*, Napoli, ClioPress, 2007. In rete: <http://www.fedoa.unina.it/1093/1/derobbio.pdf >. Ultima consultazione: 25 novembre 2015.

De Lasala, Fernando / Paulus Rabikauskas, 2003. *Il documento medievale e moderno: panorama storico della diplomatica generale e pontificia*, Roma, Editrice Pontificia Università Gregoriana – IPSAR.

De Robertis, Domenico, 1985. «L'ufficio filologico dell'Opera del Vocabolario, il suo impianto, il suo lavoro», in AA.VV. 1985, pp. 443 – 451.

De Shryver, Gilles-Maurice, 2003. «Lexicographers' Dream in the Electronic Age», *in International Journal of Lexicography*, Vol. 16, No. 2, pp. 143-199.

Devlin, Keith, 1997. *The End of Logic and the Search of a New Cosmology of the Mind*, New York, John Wiley & Sons [Trad it. da cui si cita, *Addio Cartesio*, Torino, Bollati Boringhieri, 1999].

Di Girolamo, Costanzo, 2007. «Esperienze filologiche nella rete», in *Ecdotica*, 4, pp. 160-167.In rete: <http://www.ecdotica.org/pdf/2007 /girolamo.pdf>. Ultima consultazione: 15 ottobre 2015.

Di Tonto, Giuseppe, 2003. «Scrivere per I nuovi media: dal testo cartaceo alla scrittura digitale», in Numerico / Vespignani 2003.

DO = Devoto, Giacomo / Gian Carlo Oli, *Il Dizionario della lingua italiana*, a cura di Luca Serianni e Maurizio Trifone, Firenze, Le Monnier, 1971-2014.

DMF = Sylvie Bazin-Tacchella, / Robert Martin / Gilles Souvay, *Dictionnaire du Moyen Français: version DMF 2012*, Nancy, ATILF, 2012, in rete: <http://www.atilf.fr/dmf>. Ultima consultazione: 17 ottobre 2015.

DMP = Tullio De Mauro, *Dizionario della Lingua Italiana*, Torino, Paravia, 2000.

Du Cange = Charles Du Fresne, sieur du Cange et al., *Glossarium mediæ et infimæ latinitatis*, Niort, L. Favre, 1883-1887.
Ed. Elettronica all'indirizzo <http://ducange.enc.sorbonne.fr/>, ultima consultazione: 09 ottobre 2015.

Duro, Aldo, 1985. «L'impianto del nuovo Vocabolario: profilo storico», in AA.VV. 1985, pp. 431-442.

DZ = AA.VV., *Dizionario Garzanti di Italiano*, Milano, Garzanti, 2006.

EAGLES Guidelines = *Expert Advisory Group on Language Engineering Standards Guidelines*, in rete: <http://www.ilc.cnr.it/ EAGLES/browse.html>. Ultima consultazione: 14 ottobre 2015.

Elliott, Laura / Sarah Williams, 2006. «Pasadena. A New Editing System for the *Oxford English Dictionary*», in Corino / Marello / Onesti 2006, pp. 257-264. In rete: <http://www.euralex.org /elx_proceedings/Euralex2006/033_2006_V1_Laura%20ELLIOTT,% 20Sarah%20WILLIAMS_Pasadena_A%20New%20Editing%20Syste m%20for%20the%20Oxford%20English%20Dictionary.pdf>. Ultima consultazione: 22 ottobre 2015.

Esperti, Piero, 1979. «Grammatichetta della lingua italiana ad uso del calcolatore», in Avalle 1979, pp. 123-187.

Fanfani, Massimo, 2014. «Vene moderne nel Vocabolario», in AA.VV. 2014, pp. 73-110.

Fanfani, Massimo / Marco Biffi, 2006. «La lessicografia della Crusca in rete», in Corino / Marello / Onesti 2006, pp. 409-416. In rete: < http://www.euralex.org/elx_proceedings/Euralex2006/051_2006_V1_Massimo%20FANFANI,%20Marco%20BIFFI_La%20Lessicografia%20della%20Crusca%20in%20Rete.pdf>. Ultima consultazione: 02 novembre 2015.

Fillmore, Charles J., 1992. «"Corpus Linguistics" or "Computer-aided Armchair Linguistics"», in Jan Svartvik (ed.), *Directions in Corpus Linguistics: Proceedings of Nobel Symposium 82, Stokholm, 4-8 August 1991*, Berlin, Walter de Gruyter, pp. 35- 60.

Fiormonte, Domenico, 2003. *Scrittura e filologia nell'era digitale*, Bollati Boringhieri, Torino.

Folena, Gianfranco, 1956. *La Istoria di Eneas vulgarizzata per Angilu di Capua*, a cura di G. Folena, Palermo, Centro di studi filologici e linguistici siciliani (Collezione di testi siciliani dei secoli XIV e XV, 7).

Francopulo, Gil, 2013. *LMF — Lexical Markup Framework*, ed. by G. Francopulo, London, ISTE/Wiley.

Francopulo, Gil / Chu-Ren Huang, 2914. «Lexical Markup Framework: an ISO *standard* for Electronic Lexicons and its Implications for Asian Languages», in *ASIALEX* 2014/1, pp. 37-51.

Francopoulo, Gil et al. 2006. *Lexical Markup Framework (LMF). International Conference on Language Resources and Evaluation - LREC 2006, 2006, Gênes/Italie*. In rete: <https://hal.inria.fr/inria-00121468>. Ultima consultazione: 14 ottobre 2015.

Fuertes-Olivera, Pedro, 2009. «The Function Theory of Lexicography and Electronic Dictionaries: Wikitionary as a Prototype of Collective Free Multiple-Language Internet Dictionary», in Henning Bergenholtz / Sandro Nielsen / Sven Tarp (eds.), *Lexicography at a Crossroads. Dictionaries and Encyclopedias Today, Lexicographical Tools Tomorrow*, Bern, Peter Lang, pp. 99-134.

Fuertes-Olivera, Pedro / Henning Bergenholtz (eds.), 2011. *E-Lexigography, The Internet Initiatives and Lexicography*, London-New York, Continuum.

Fuertes-Olivera, Pedro / Marta Niño-Amo, 2011. «Internet Dictionaries for Communicative and Cognitive Functions: *El Diccionario Inglés-Español de Contabilidad*» in Fuertes-Olivera / Bergenholtz 2011, pp. 187-207.

Galimberti, Umberto, 2000. *Psiche e Techne: l'uomo nell'età della tecnica*, Milano, Feltrinelli.

Gandin, Stefania, 2005. «Linguistica dei corpora e traduzione: definizioni, criteri di compilazione e implicazioni di ricerca dei corpora

paralleli» in *Annali della Facoltà di Lingue e Letterature Straniere dell'Università di Sassari*, Vol. 5/2005 (2009).

Garrett et al., 2005. «The contribution of corpus linguistics to lexicography and the future of Tibetan dictionaries», *Revue d'Etudes Tibétaines*, no. 32, pp. 51–86. In rete: <http://www.researchgate.net/publication/275275442_The_contribution_of_corpus_linguistics_to_lexicography_and_the_future_of_Tibetan_dictionaries>. Ultima consultazione: 13 ottobre 2015.

GAVI = *Glossario degli Antichi Volgari Italiani*, a cura di Giorgio Colussi, Holsinki-Foligno, Helsinki University Press – Editoriale Umbra, 1983-2006.

GDLI = *Grande Dizionario della Lingua Italiana*, a cura di Salvatore Battaglia e Giorgio Bàrberi Squarotti, Torino, UTET, 1961-2002, 21 voll.

Gigliozzi, Giuseppe, 1987a. *Studi di codifica e trattamento automatico dei testi*, a cura di G. Gigliozzi, Roma, Bulzoni.

—, 1987b. «Codice, testo e interpretazione», in Gigliozzi 1987a, pp.65-84.

—, 1997. *Il testo e il computer. Manuale di informatica per gli studi letterari*, Milano, Bruno Mondadori.

Giorgini-Broglio = Emilio Broglio / Giovan Battista Giorgini, *Novo Vocabolario della lingua italiana secondo l'uso di Firenze*, Firenze, Cellini, 1870-1897.

Gleick, James, 1987. *Chaos*, New York, Viking Penguin; trad. it. da cui si cita *Caos, nascita di una nuova scienza*, Milano, Rizzoli, 2001.

Gleßgen, Martin-Dietrich, 2006. «Esigenze della tecnologia informatica nella filologia e lessicografia storica», in Schweickard 2006, pp.15-24.

Gleßgen, Martin-Dietrich / Mathias Kopp, 2005. «Linguistic annotation of texts in non-standardized languages: the program procedures of the tool PHOENIX», in Pusch et al. 2005, pp. 147-154.

Gouws et al. 2013. *Dictionaries. An International Encyclopedia of Lexicography. Supplementary Volume: Recent Developments with Focus on Electronic and Computational Lexicography*, Berlin, De Gruyter Mouton.

Gouws, Rufus, 2011. «Learning, Unlearning and Innovation in the Planning of Electronic Dictionaries», in Fuertes-Olivera / Bergenholtz 2011, pp. 17-29.

GRADIT = *Grande Dizionario Italiano dell'Uso*, diretto da Tullio De Mauro, Torino, UTET, 2000, 6 voll.

Graiger, Christian, 2006. *TEI P5: Encoding Dictionaries & More*, in rete <https://www.textgrid.de/fileadmin/publikationen/graiger-2006.pdf>. Ultima consultazione: 30 gennaio 2014.

Granger, Sylviane, 2012. «Introduction: Electronic lexicography: From Challenge to Opportunity», in Granger / Paquot 2012, pp. 1-11. In rete: <http://www.researchgate.net/publication/259640872_Electronic_lexicography_From_challenge_to_opportunity>. Ultima consultazione: 10 novembre 2015.

Granger, Sylviane / Magali Paquot, 2012. *Electronic Lexicography*, Oxford, Oxford University Press.

Gregory, Stewart / William Rothwell / David A. Trotter, 2005. *Anglo-Norman Dictionary. Second edition*, Leeds, Maney Publishing.

Grishman, Ralph, 1986. *Computational linguistics*, Cambridge, Cambridge University Press; trad. it. da cui si cita: *Linguistica computazionale*, Milano, Tecniche Nuove, 1988.

Hanks, Patrick, 1990. «Evidence and intuition in lexicography», in J. Tomaszczyk and B. Lewandowska-Tomaszczyk *Meaning and Lexicography* (Linguistic and Literary Studies in Eastern Europe 28), pp. 31-42.

—, 2003. «Lexicography», in Ruslan Mitkov (ed.), *The Oxford Handbook of Computational Linguistics*, Oxford, Oxford University Press.

—, 2012. «The Corpus Revolution in Lexicography», in *International Journal of Lexicography*, Vol. 25, No.4, pp. 398-436.

Hanon, Souzanne, 1990. «La concordance», in Gouws et al. 2013, pp. 1562, 1567.

Harold, Elliotte Rusty / William Scott Means, 2001. *XML. Guida di riferimento*, Milano, Apogeo/O'Reilly.

Hiltrud Gerner, 2005. «La *Base des Lexiques du Moyen Francçais* (*BLMF*): Première étape vers le *DMF* électronique», in Pusch / Rabatek / Raible 2005a, pp. 155-162.

Hockey, Susan, 2000. *Electronic Texts in the Humanities: Principles and Practice*, Oxford-New York, Oxford University Press.

Höfler, Manfred, 1986. «Typologie des erreurs de datation dans la lexicographie française», Revue de Linguistique Romane, 50, pp. 423-441. <http://dx.doi.org/10.5169/seals-399799>. Ultima consultazione: 30/07/2015.

Huguet = Huguet, Edmond, 1925-1967. *Dictionnaire de la langue française du seizième siècle*, Parigi, Champion-Didier, 7 voll.

Iorio-Fili, Domenico, 2006. «Estratto dalla guida di GattoWeb», in Bollettino dell'Opera del Vocabolario Italiano, XI, Alessandria, Edizioni dell'Orso, pp. 273-397.

—, 2007. «Breve storia, stato attuale e prospettive del software GATTO», in Bollettino dell'Opera del Vocabolario Italiano, XII, Alessandria, Edizioni dell'Orso, pp. 365-386.

—, 2010. «Un nuovo strumento di lemmatizzazione automatica per corpora testuali di ridotte dimensioni. Applicazione all'italiano antico», in Bollettino dell'Opera del Vocabolario Italiano, XI, Alessandria, Edizioni dell'Orso, pp. 367-390.

—, 2012. Manuale Gatto. In rete <http://www.ovi.cnr.it/index.php?page=gmanuale>. Ultima consultazione: 05 ottobre 2013

Kiesewetter, Andreas, 1993. «Eleonora d'Angiò, regina di Sicilia», in *Dizionario Biografico degli Italiani* – Vol. 42 (1993)
In rete: <http://www.treccani.it/enciclopedia/eleonora-d-angio-regina-di-sicilia_(Dizionario_Biografico)/>. Ultima consultazione: 01 settembre 2015.

Kilgarriff, Adam, 2012. Recensione a Fuertes-Olivera / Bergenholtz (2011), in Kernerman Dictionary News, July 2012. In rete <http://www.sketchengine.co.uk/documentation/attachment/wiki/AK/Papers/kdn20_2012_pp26-29.pdf?format=raw>. Ultima consultazione: 5 ottobre 2013.

Klein, Wolfgang, 2013. «Declino e crescita della lessicografia tedesca», in Maraschio / De Martino / Stanchina 2013.

Kuhn, Thomas, 1978 (1° ed. 1962). *The Structure of Scientific Revolutions, Chicago*, University of Chicago Press; trad. it. *La struttura delle rivoluzioni scientifiche*, Torino, Einaudi.
<http://projektintegracija.pravo.hr/_download/repository/Kuhn_Structure_of_Scientific_Revolutions.pdf>. Ultima consultazione: 3 ottobre 2015.

Lee Berg, Donna / Gaston H. Gonnet / Frank WM. Tompa 1990-1991. «The New *Oxford English Dictionary* Project at the University of Waterloo», in Cignoni / Peters 1990-1991, pp. 29-63.

LEI = Max Pfister, *Lessico Etimologico Italiano*, Reichert, Wiesbaden, 1979- .
In rete: <http://www.uni-saarland.de/lehrstuhl/schweickard/lei/pubblicazioni.html>. Ultima consultazione: 15 ottobre 2015.

Lemnitzer, Lothar / Laurent Romary / Andreas Witt, 2010. «Representing Human and Machine Dictionaries in Markup Languages (SGML, XML)», in Gouws et al. 2013, pp. 1195-1209.

Lenci, Alessandro / Montemagni, Simonetta / Pirrelli, Vito, 2005. *Testo e computer. Elementi di linguistica computazionale*, Roma, Carocci.

Leonardi, Lino, 2013. *Introduzione* a «A che servono i DATABASES? Esperienze di informatica per la filologia romanza», in *Le forme e la storia*, n.s. 1, pp. 199-202.

Li Gotti, Ettore, 1951. *Volgare nostro siculo. Crestomazia di testi siciliani del sec. XIV. Parte I*, a cura di E. Li G., Firenze, La Nuova Italia

LIE = *Letteratura Italiana Einaudi*, consulenza scientifica di Alberto Asor Rosa, Torino, Einaudi, 1999 / 2002 (10 CD-ROM).

LIZ = *Letteratura Italiana Zanichelli 4.0*, a cura di Pasquale Stoppelli ed Eugenio Picchi, Bologna, Zanichelli, 2001 (in CD-ROM).

Lo Piparo, Franco, 1987. «Sicilia linguistica», in M. Aymard – G. Giarrizzo, Storia d'Italia, La Sicilia, Torino, Einaudi, pp. 735-807.

Losee, John, 2009. *Filosofia della scienza*, Milano, Il Saggiatore. [Trad. di A Historical Introduction to the Philosophy of Science, Third Edition, Oxford, Oxford University Press, 1993 (1° ed. 1972)].

Mandelbrot, Benoit. 1954. «Structure formelle des textes et communcation» in *Word*, X, pp. l-27.

Manning, Christopher D. / Hinrich Schütze, 1999. *Foundations of Statistical Natural Language Processing*, Cambridge (MA) - London, MIT Press.

Manovich, Lev, 2002. *Il linguaggio dei nuovi media*, Milano, Edizioni Olivares.

Maraschio, Nicoletta / Domenico De Martino / Giulia Stanchina (a cura di), 2013. *L'italiano dei vocabolari. Atti della sesta edizione della Piazza delle Lingue (6-7 novembre 2012)*, Firenze, Accademia della Crusca, (La Piazza delle lingue, 4).

Maraschio, Nicoletta / Teresa Poggi Salani, 2014. «La prima edizione del *Vocabolario degli Accademici della Crusca*», in AA.VV. 2014, pp. 25-66.

Marello, Carla, 1996. *Le parole dell'italiano. Lessico e dizionari*, Bologna, Zanichelli.

Marinoni, Augusto, 1955. *Dal* Declarus *di A. Senisio i vocaboli siciliani*, a cura di A. Marinoni, Palermo, Centro di Studi Filologici e Linguistici Siciliani (Collezione di testi siciliani dei secoli XIV e XV, 6).

Martin, Robert, 1998. «Le *Dictionnaire du moyen français* (*DMF*)», in *Comptes-rendus de l'Académie des Inscriptions et Belles-Lettres*, 142^{e} année, N.4, pp. 961-982. In rete : <http://www.persee.fr/doc/crai_0065-0536_1998_num_142_4_15923>. Ultima consultazione : 30 novembre 2015.

—, 2007. *Dictionnaire du Moyen Français DMF (1330–1500). Seconde version: DMF2. Présentation.* In rete: <http://atilf.atilf.fr/gsouvay/dmf2/ PresentationDMF2.pdf>. Ultima consultazione: 14 ottobre 2011.

Martin, Robert / Hiltrud Gerner / Gilles Souvay, 2010. «Présentation de la seconde version du *DMF* (*Dictionnaire du Moyen Français*)», *in* Maria Iliescu et al. (éd.), *Actes du XXV^{e} Congrès International de Linguistique et de Philologie Romane (Innsbruck, 3-8 septembre 2007)*, Berlin, De Guyter, Tome VI, pp. 213-220.

—, 2008. «Perspectives de la lexicographie informatisée» in Jacques Durand / Bruno Habert / Bernard Laks, *Congrès Mondial de Linguistique Française – CMLF'08*, Paris, Institut de Linguistique Française, pp. 1251-1256.

Matsumura, Takeshi, 2003a. «Sur la versione électronique de Godefroy», in Frédéric Duval, *Frédéric Godefroy. Actes du X*e *colloque international sur le Moyen Français,* Paris, Champion, pp. 405-408. [Il contributo è visualizzato integralmente nell'anteprima del volume in Google Libri. Link diretto abbreviato: <http://*bit*.ly/1WRsLtk>. Ultima consultazione: 11 novembre 2015].

—, 2003b. «Review of *Godefroy* (CD-ROM) and *T-L* (CD-ROM)», in Revue de Linguistique Romane, 67, pp. 265-272.

McEnery, Tony / Andrew Wilson, 2001. *Corpus Linguistics: An Introduction*, Edinburgh, Edinburgh University Press.

McLuhan, Marshall, 1964. *The Gutenberg Galaxy: the Making of Typographic Man*, Toronto, University of Toronto Press.

—, 1964. *Understanding Media*, New York, McGraw-Hill [Trad. it. da cui si cita: *Gli strumenti del comunicare*, Milano, Net].

Meyer, Charles F., 2012. *English Corpus Linguistics: an Introduction*, Cambridge, Cambridge University Press.

Migliorini, Bruno, 1944-45. «L'atto di nascita dei vocaboli», in *Lingua Nostra* VI, 1944-45, pp. 6-10.

—, 1956. «Alcune retrodatazioni lessicali», in *Studi letterari. Miscellanea in onore di Emilio Santini*, Palermo, Manfredi, pp.189-195.

Mugglestone, Lynda, 2000. «Pioneers in the Untrodden Forest: The *New* English Dictionary», in L. Mugglestone (ed.), *Lexicography and the OED: Pioneers in the Untrodden Forest*, Oxford, Oxford University Press.

Murray, James, 1879. *Appeal to the English-speaking and English-reading public*. In rete: <http://public.oed.com/history-of-the-oed/archived-documents/april-1879-appeal/>. Ultima consultazione: 18 ottobre 2015.

Murray, K. M. Elisabeth, 1979. *Caught in the Web of Words: James A. H. Murray and the Oxford English Dictionary*, Oxford, Oxford University Press.

Muta, 1773. *Pragmaticae Regni Siciliae*, Palermo, Tomo III, p. 183, *Titulus XXV De Notariato numero, examine e officio* (pp. 178-187).

Nencioni, Giovanni, 1955. «Relazione all'Accademia della Crusca sul Vocabolario della Lingua Italiana (1955)», in *Studi di Filologia Italiana*, XIII, pp. 395-420. Poi in Barbi / Pasquali / Nencioni (1957), da cui si cita.

—, 1961. «Filologia e lessicografia a proposito della “variante”», in *Studi e problemi di critica testuale. Atti del Convegno di studi di filologia italiana*, Bologna, aprile 1960, Bologna, Commissione per i testi di lingua, pp. 183-192; poi in Nencioni 1983.
In rete: <http://nencioni.sns.it/fileadmin/template/allegati/pubblicazioni/1961/Studi_1961.pdf>

—, 1983. *Di scritto e parlato. Discorsi linguistici*. Bologna, Zanichelli.

—, 1985. «Verso una nuova lessicografia», in *Studi di lessicografia italiana*, VII, 1985, pp. 5-19.

—, 1987. «Lessico tecnico e difesa della lingua», in *Studi di lessicografia italiana*, IX, 1987, pp. 5-20; poi in Cignoni / Peters 1990-1991, vol. II, pp. 157-175, ca cui si cita.
Le due edizioni in rete: <http://nencioni.sns.it/index.php?id=670>. Ultima consultazione: 20 ottobre 2015.

Numerico, Teresa / Arturo Vespignani, 2003. *Informatica per le scienze umanistiche*, Bologna, Il Mulino.

OED = *Oxford English Dictionary*, Oxford, Oxford University Press. (Supplement 1933; 2nd ed. 1972-1986; 3rd ed. 1989). In rete: <http://www.oed.com>.

OED Compact = *The Compact Edition of The Oxford English Dictionary, Complete Text Reproduced Micrographically (in slipcase with reading glass)*, Oxford, Oxford University Press, 1979

Orlandi, Tito, 1986. «Problemi di codifica e trattamento informatico in campo filologico», in Savoca 1985.
In rete: <http://www.cmcl.it/~orlandi/pubbli/info054.html>. Ultima consultazione 31 luglio 2015.

—, 1987. «Informatica umanistica», in Gigliozzi 1987, pp. 1-38.

—, 2002. «Is humanities computing a discipline?», in *Jahrbuch für Computerphilologie*, IV (2002), pp. 51-58.
In rete: <http://computerphilologie.uni-muenchen.de/jg02/orlandi.html>. Ultima consultazione: 22 ottobre 2015.

—, 2004. «Jaufré Rudel, ovvero Le disgrazie di un navigatore», in *La Cultura*, 2004/33, pp. 495-504. In rete: <http://www.cmcl.it/~orlandi/pubbli/info113.pdf>. Ultima consultazione 23 febbraio 2014.

—, 2010. *Informatica testuale. Teoria e prassi*, Roma-Bari, Laterza.

Paccagnella, Ivano, 2012. *Vocabolario del Pavano (XIV-XVII secolo)*, Padova, Esedra Editrice.

Pagano, Mario, 2008. *Corpus Artesia*, Catania, Ed.it, (Quaderni di Artesia 2-3).

—, 2009. «Il progetto *ARTESIA* (Archivio Testuale del Siciliano Antico)», in *Le forme e la storia*, n.s. II, 2, pp. 295-303.

—, 2011. «Per un Vocabolario del siciliano medievale», in *Per i linguisti del nuovo millennio. Scritti in onore di Giovanni Ruffino*, a cura del Gruppo di ricerca dell'Atlante Linguistico della Sicilia, Palermo, Sellerio, 2011, pp. 312-317.

—, 2012. «Appunti sparsi per un Vocabolario del siciliano medievale (VSM)», in *Bollettino del Centro di studi filologici e linguistici siciliani*, 23, pp. 113-137.

—, 2013. «Il Medioevo. Testi in prosa», in *Lingue e Culture in Sicilia*, a cura di Giovanni Ruffino, Palermo, Centro di studi filologici e linguistici siciliani, 2013, pp. 792-817.

Pagano, Mario / Salvatore Arcidiacono, 2013. «Corpus Artesia (Archivio Testuale del Siciliano Antico)» in Emili Casanova Herrero / Cesáreo Calvo Rigual, (edd.), *Actes del 26é Congrés de Lingüística i Filologia Romàniques (València, 6-11 de setembre de 2010)*, Vol VIII, Berlin, Walter de Gruyter, pp. 253-262.

Pagano, Mario / Margherita Spampinato, 2007. «Filologia ed informatica in ambito romanzo: l'Archivio testuale del siciliano antico (ARTESIA)», in Domenico Antonio Cusato / Domenico Iaria / Rosa Maria Palermo (a cura di), *Atti del V convegno interdisciplinare su Testo, Metodo, Elaborazione elettronica* (Messina – Catania – Brolo, 16-18 novembre 2006), Messina, Andrea Lippolis Editore, pp. 323-338.

Parenti, Alessandro, 2009. «Nuove datazioni di termini della linguistica», in *Quaderni del Dipartimento di Linguistica dell'Università di Firenze*, 19 (2009), pp. 223-258. In rete: <https://www.yumpu.com/it/document/view/16149162/2009223-258-nuove-datazioni-di-termini-linguistica/35>, ultima consultazione 29/ luglio 2015.

Pasqual, Antonio, 2013. «Piccoli lavori e grandi risultati: c'è vita fuori del *Nuevo Diccionario Histórico* della Real Academia Española?», in Maraschio / De Martino / Stanchina 2013.

Pasquali, Giorgio, 1941. «Per un Tesoro della lingua italiana», in *Atti della Reale Accademia d'Italia, Rendiconti della Classe di scienze morali e storiche*, 7ª serie, II, pp. 490-521 [poi in Barbi / Pasquali / Nencioni 1957, da cui si cita].

Pepe, Gabriele, 1998. *Introduzione allo studio del medioevo latino*, Bari, Edizioni Dedalo.

Picchi, Eugenio, 1990-1991. «DBT: a Textual Database System», in Cignoni / Peters 1990-1991, vol. II, pp. 177-205.

Pierrel, Jean-Marie / Eva Buchi, 2009. «Research and Resource Enhancement in French Lexicography: the ATILF Laboratory's Computerized Resources», in Bruti / Cella / Foschi 2009, pp.79-117. *Preprint* in rete: <https://halshs.archives-ouvertes.fr/halshs-00258126/document>. Ultima consultazione: 17 ottobre 2015.

Pratesi, 1957. «Una questione di metodo: l'edizione delle fonti documentarie», in *Rassegna degli archivi di stato*, XVII/1.

Pruvost, Jean, 2000. «Colloquium report: Des dictionnaires papier aux dictionnaires électroniques. VII Journée des dictionnaires (22 mars 2000)», in *International Journal of Lexicography*, Vol. 13, No.3, pp. 187-193.

Pusch, Claus D. / Johannes Rabatek / Wolfgang Raible (eds.), 2005a. *Romanistische Korpuslinguistik II. Korpora und diachrone Sprachwissenschaft. Romance Corpus Linguistics II. Corpora and Diachronic Linguistics*, Tübingen, Gunter Narr Verlag.

—, 2005b. «Romance corpus linguistics and language change», in Pusch / Rabatek / Raible 2005a, pp. 1-10.

Quemada, Bernard, 1991. «Acquis et perspectives de l'informatique», in *Travaux de linguistique*, 23, pp. 17-21.

Raynaud, Savina, 2009. «Dall'indicizzazione all'ermeneutica testuale. Filosofia del linguaggio e linguistica computazionale», in *Informatica Umanistica*, 2009/2. In rete : <http://www.ledonline.it/informatica-umanistica/Allegati/IU-02-09-Raynaud.pdf>. Ultima consultazione : 12 ottobre 2015.

Raffaele, Ferdinando, 2009. *Lu raxunamentu di l'abbati Moises e di lu beatu Germanu supra la virtuti di la discretioni*, a cura di F. Raffaele, Palermo, Centro di studi filologici e linguistici siciliani (Supplementi al Bollettino, 17).

Roncaglia, Gino, 2002. «Informatica umanistica: le ragioni di una disciplina», in *Intersezioni*, 3/2002, pp. 353-376.

Rothwell, Andrew, 2005. «Computerization of the *AND*», in Gregory / Rothwell / Trotter 2005. In rete (da cui si cita): <http://cadair.aber.ac.uk/dspace/bitstream/handle/2160/2539/Introduction%20to%20the%20Anglo-Norman%20Online%20Hub.pdf?sequence=1>. Ultima consultazione: 11 novembre 2015.

Rothwell, William, 2005. «Anglo-French and the AND», in Gregory / Rothwell / Trotter 2005. In rete (da cui si cita): <http://cadair.aber.ac.uk/dspace/bitstream/handle/2160/2539/Introduction%20to%20the%20Anglo-Norman%20 Online%20Hub.pdf?sequence=1>. Ultima consultazione: 11 novembre 2015.

Ruffino, Giovanni, 1989. «La dialettologia siciliana tra consuntivi e programmi», in Günter Holtus / Michele Metzeltin / Max Pfister (a cura di), *La dialettologia italiana oggi: studi offerti a Manlio Cortelazzo*, Tübingen, Gunter Narr Verlag, pp. 325-343.

Rundell, Michael / Adam Kilgarriff, 2011. «Automating the Creation of Dictionaries: Where will it all End? », in Fanny Meunier / Sylvie De Cock / Gaëtanelle Gilquin / Magali Paquot (eds), *A taste for corpora. A tribute to professor Sylviane Granger*, Amsterdam, John Benjamins. In rete <http://www.sketchengine.co.uk/documentation/raw-attachment/wiki/AK/Papers/2011-RundellKilg-SylvianeFest-automating .doc>. Ultima consultazione: 09 dicembre 2013.

Sabatini, Francesco, 2006. «La storia dell'italiano nella prospettiva della *corpus linguistics*», in Corino / Marello / Onesti 2006, pp. 31-37; ora in Francesco Sabatini, *L'italiano nel mondo moderno. Saggi scelti dal 1968 al* 2009, a cura di Vittorio Coletti, Rosario Coluccia, Paolo D'Achille, Nicola De Blasi, Domenico Proietti e Riccardo Cimaglia, Napoli, Liguori, 2011. In rete: <http://www.euralex.org/elx_proceedings/ Euralex2006/004_2006_V1_Francesco%20SABATINI_La%20Storia %20dellItaliano%20nella%20Prospettiva%20della%20Corpus%20Linguistics. pdf>. Ultima consultazione: 2 novembre 2015.

—, 2007. «Storia della lingua italiana e grandi corpora. Un capitolo di storia della linguistica», in Barbera / Corino / Onesti 2007, pp. xiii-xvi.

—, 2014. «Un ponte fra l'età di Dante e l'Unità nazionale», in AA.VV. 2014.
In rete: <http://www.edizionidicrusca.it/download2/PDF/473_361.pdf>. Ultima consultazione: 2 novembre 2015.

Salarelli, Alberto / Anna Maria Tammaro, 2006. *La biblioteca digitale*, Milano, Editrice Bibliografica.

Salvi, Giampaolo / Lorenzo Renzi (a cura di), 2010. *Grammatica dell'italiano antico*, Bologna, Il Mulino (2 voll.).

Santangelo, Salvatore, 1933. *Libru de lu Dialugu de Sanctu Gregoriu traslatatu pir frati Iohanni Campulu de Missina*, a cura di S. Santangelo, Palermo, Scuola Tipografica «Boccone del Povero».

Sardina, Patrizia, 2006. «Il notaio Vitale de Filesio, vicesecreto di Agrigento nell'età dei Martini (1392-1410)», in *Mediterranea*, 3, pp. 423-442.

Sardo, Rosaria, 2008. *«Registrare in lingua volgare». Scritture pratiche e burocratiche in Sicilia tra '600 e '700*, Palermo, Centro di studi filologici e linguistici siciliani.

Savoca, Giuseppe, 1986. *Lessicografia, filologia e critica. Atti del Convegno internazionale di studi (Catania-Siracusa, 26-28 aprile 1985)*, a cura di G. Savoca, Firenze, L. S. Olschki, 151-175.

—, 1994. *L'automa. Lettera aperta a Oreste Macrì*, Firenze, 1994.

Schäfer, Jürgen. 1980. *Documentation in the O.E.D. Shakespeare and Nashe as test cases*, Oxford, Clarendon Press.

Schweickard, Wolfgang (ed.), 2006. *Nuovi media e lessicografia storica: Atti del colloquio in occasione del settantesimo compleanno di Max Pfister*, Berlin, Walter de Gruyter.

—, 2013. «Italian», in Gouws et al. 2013, pp. 672-687.

Scotti, Andrea, 2003. «La base di dati e la sua struttura: uno strumento per l'umanista», in Numerico / Vespignani 2003, pp. 91-109.

Sechi, Letizia, 2010. *Editoria digitale*, Milano, Apogeo.
In rete: <http://www.apogeonline.com/libri/9788850310975/scheda>. Ultima consultazione 13 ottobre 2015.

Segre, Cesare, 1981. «Testo», in *Enciclopedia Einaudi*, Torino, Einaudi, vol. XIV, pp. 276-277.

—, 1985. Avviamento all'analisi del testo letterario, Torino, Einaudi.

Serianni, Luca, 1999. *Dizionari di ieri e di oggi*, Milano, Garzanti.

Simpson, John, 2009. «What does the Future Hold for Historical Lexicography?», in Bruti / Cella / Foschi Albert 2009.

— 2013. «Il futuro dell'*OED*: dati lessicali sempre più accessibili», in Maraschio / De Martino / Stanchina 2013, pp. 35-51.

Sinclair, John McHardy, 1987a. «The Nature of the Evidence», in J. M. Sinclair (ed.), *Looking Up: An Account of the COBUILD Project in Lexical Computing and the Development of the Collins COBUILD English Language Dictionary*, London, Collins, pp. 150-159.

—, 1987b. «The Dictionary of the Future, Collins English Dictionary Annual Lecture, University of Strathclyde, 6 May 1987», in *Library Review*, Vol. 36 Iss: 4, pp.268 – 278.

Spampinato, Margherita, 2013. «La violenza verbale in un corpus documentario del tardo Medioevo italiano: aspetti pragmatici», in Emili Casanova Herrero / Cesáreo Calvo Rigual, (edd.), *Actes del 26é Congrés de Lingüística i Filologia Romàniques (València, 6-11 de setembre de 2010)*, Berlin, Walter de Gruyter, pp. 683-694.

Spampinato, Simona, 2002. *La versione quattrocentesca dell'Istoria di Eneas di Angelo di Capua: edizione interpretativa, studio linguistico-letterario e concordanza*, Tesi di dottorato, Catania, Dipartimento di Filologia Moderna.

Spina, Stefania, 2001. *Fare i conti con le parole. Introduzione alla linguistica dei corpora*, Perugia, Guerra edizioni.

Squillacioti, Paolo, 2002. «Il Tesoro della Lingua Italiana delle Origini», in *Verbum*, vol. IV/2002/2, pp. 503-516. In rete: <https://www.academia.edu/5797367/Il_Tesoro_della_Lingua_Italiana_delle_Origini>, ultima consultazione: 27 novembre 2015.

—, 2010. «Uno sguardo al Tesoro della Lingua Italiana delle Origini: procedure e prospettive del vocabolario storico dell'italiano antico», in *Dizionari e ricerca filologica. Atti della Giornata di studi in memoria di Valentina Pollidori, Firenze, 26 ottobre 2010*, Alessandria, Edizioni dell'Orso (Supplemento III al *Bollettino dell'Opera del Vocabolario Italiano*).

Summers, Dela, 1993. «Longman / Lancaster English Language Corpus – Criteria and Design», in *International Journal of Lexicography*, Vol. 6, No. 3, pp. 181-208.

Tarp, Sven, 2012. «Lexicographical and Other e-Tools for Consultation Purposes: Towards the Individualization of Needs Satisfaction», in Fuertes-Olivera / Bergenholtz 2012, pp. 54-70.

TEI Guidelines, 2013. *TEI consortium, TEI P5: Guidelines for Electronic Text Encoding and Interchange, Version 2.5.0. Last updated on 26th July 2013*, Virginia, Text Encoding Initiative Consortium Charlottesville.
In rete: <http://www.tei-c.org/Guidelines/P5/>. Ultima consultazione: 20 settembre 2014.

TLF = Paul Imbs / Bernard Quemada, *Trésor de la Langue Française. Dictionnaire de la langue du XIXe et du XXe siècle (1789-1994)*, Parigi, Éditions du CNRS-Gallimard, 1971-1994, 16 voll.
In rete: <http://atilf.atilf.fr/tlf.htm>. Ultima consultazione: 30 ottobre 2015.

TLF-Étym = Nadine Steinfeld, *Projet TLF-Étym (révision sélective des notices étymologiques du* Trésor de la Langue Française informatisé, Nancy, ATILF, 2005. In rete: <http://atilf.atilf.fr/tlf.htm>. Ultima consultazione: 30 ottobre 2015.

Tobler-Lommatzsch = Adolf Tobler / Erhard Lommatzsch, *Altfranzösisches Wörterbuch*, Berlin - Wiesbaden - Stuttgart, Weidmann - Steiner, 1925-2002 (11 voll.).

Tognini-Bonelli, Elena, 1996. «The Malvern Seminar: Towards Translation Equivalence from a Corpus Linguistics Perspective», in *International Journal of Lexicography*, Vol. 9, No. 3, pp. 197–217.

—, 2001. *Corpus Linguistics at Work*, Amsterdam - Philadelphia, John Benjamins Publishing Company, 2001 (Studies in Corpus Linguistics, 6).

Tommaseo-Bellini = *Dizionario della lingua italiana*, a cura di Nicolò Tommaseo e Bernardo Bellini, Torino, Unione Tipografico-Editrice Torinese, 1861-1879 (4 voll., ciascuno dei quali diviso in due tomi).

Trotter, David, 2011. «Bytes, Words, Texts: The Anglo-Norman Dictionary and Its Text-Base», in *Digital Medievalist*, 7. In rete <http://www.digitalmedievalist.org/journal/7/trotter/>. Ultima consultazione: 15 gennaio 2013.

—, 2013. «Gallo-Romance II – Synchronic Lexicography», in Gouws et al. 2013, pp. 663-672.

Turing, Alan, 1950. «Computer Machinery and Intelligence», in *Mind*, New Series, Vol. 59, No. 236, pp. 433-460.

Ugolini, Francesco A. 1957. *Valeriu Maximu translatatu in vulgar messinisi per Accursu di Cremona*, a cura di F. A. Ugolini, I-II, Palermo, Centro di studi filologici e linguistici siciliani (Collezione di testi siciliani dei secoli XIV e XV,10-11).

Unicode *standard* V.8.0 = Julie D. Allen et al. (eds.), *The Unicode standard* (*Version 8.0)*, Mountain View (CA), Unicode Consortium,

1991-2015.
In rete: <http://www.unicode.org/versions/Unicode8.0.0/> (la versione per il print-on-demand è in corso di pubblicazione). Ultima consultazione: 16 ottobre 2015.

Van Noppen, Jean-Pierre, 1983. Recensione a Jürgen Schäfer, *Documentation in the O.E.D. Shakespeare and Nashe as Test Cases*, Oxford, Clarendon Press, 1980, in *Revue belge de philologie et d'histoire*, 61/3, pp. 707-708.
In rete: <http://www.persee.fr/doc/rbph_0035-0818_1983_num_61_3_5917_t1_0707_0000_3 >. Ultima consultazione: 28/10/2015.

Vàrvaro, Alberto, 1995. «Südkalabrien und Sizilien», in Günter Holtus / Michael Metzeltin / Christian Schmitt (herausgegeben von) *Lexikon der Romanistischen Linguistik (LRL)*, *Volume II*, *2*, Tübingen, Max Niemeyer Verlag, pp. 228-237

—, 1998. «Storia della lingua e filologia (a proposito di lessicografia)», in *Storia della lingua italiana e storia letteraria. Atti del I Convegno ASLI (Firenze, 29-30 maggio 1997)*, a cura di Nicoletta Maraschio e Teresa Poggi Salani, Firenze, Franco Cesati Editore.

VEI = Angelico Prati, *Vocabolario etimologico italiano*, Milano, Garzanti, 1951.

Verlinde, Serge, 2012. «Modelling Interactive Reading. Traslating and Writing Assistants», in Fuertes-Olivera / Bergenholtz 2012, pp. 275-286.

Verlinde, Serge / Patrick Leroyer / Jean Binon, 2009. «Search and You Will Find. From Stand-Alone Lexicographic Tools to User Driven Task and Problem Oriented Multifunctional Leximats», in *International Journal of Lexicography*, Vol. 23, No. 1, pp. 1–17.

Verlinde, Serge / Thierry Selva, 2001. «Corpus-Based vs Intuition-Based Lexicography: Defining a Word List for a French Learners' Dictionary», in Paul Rayson et al. (eds.), *Proceedings of the Corpus Linguistics 2001 conference*, Lancaster University, pp. 594-598.
In rete: <http://www.kuleuven.be/grelep/publicat/verlinde.pdf>. Ultima consultazione: 9 ottobre 2015.

VES = Alberto Vàrvaro, *Vocabolario Etimologico Siciliano*, Palermo, Centro di studi filologici e linguistici siciliani, 1986 (Lessici siciliani, 3).

VLI, *Vocabolario della lingua italiana*, diretto da Aldo Duro, Roma, Istituto dell'Enciclopedia Italiana, 1980, 5 voll.

VS, *Vocabolario siciliano*, a cura di Giorgio Piccitto (vol. I), diretto da Giovanni Tropea (voll. II-IV), a cura di Salvatore Carmelo Trovato (vol. V), Palermo, Centro di studi filologici e linguistici siciliani, 1977-2002.

VSL = *Il Vocabolario Siciliano Latino di L. C. Scobar*, a cura di Alfonso Leone, Palermo, Centro di studi filologici e linguistici siciliani, 1990.

VSES = Alberto Varvaro, *Vocabolario Storico-Etimologico del Siciliano*, Palermo – Strasbourg, Centro di studi filologici e linguistici siciliani – Société de Linguistique Romane, 2014.

ZNG = Zingarelli, Nicola, *Lo Zingarelli 2015*, a cura di Mario Cannella, Beata Lazzarini, Milano, Zanichelli, 2014.

Zampolli, Antonio, 2003. «Le principali attività dell'Istituto di Linguistica Computazionale. Il punto di vista del Direttore», in Antonio Zampolli, Nicoletta Calzolari / Laura Cignoni (a cura di.), *Computational Linguistics in Pisa - Linguistica Computazionale a Pisa*, Pisa-Roma, IEPI (*Linguistica Computazionale*, Special Issue, XVI-XVII), tomo I, pp. xvii-lxx.

Zipf, George Kingsley, 1949. *Human Behaviour and the Principle of Least Effort: An Introduction to Human Ecology*, Oxford, Addison-Wesley Press; ristampa anastatica: Eastford, Martino Fine Books, 2012.

Zolli, Paolo, 1986. «Filologia e lessicografia: il problema della postdatazione», in Savoca 1986, pp. 151-175.

APPENDICE

INTRODUZIONE ALL'INTERROGAZIONE DEL *CORPUS ARTESIA 2015*

9 INSTALLAZIONE

9.1 REQUISITI DI SISTEMA

Il *Corpus Artesia* può essere utilizzato sulla gran parte dei *personal computer*, compresi modelli poco recenti. GATTO il motore di interrogazione del *Corpus*, funziona con tutte le versioni di Windows a 32 bit e 64 bit (da Windows 95 in poi).

REQUISITI MINIMI

Sistema Operativo: Microsoft Windows (Windows 95, Windows 98/ME, Windows XP, Windows Vista, Windows 7 e Windows 8)
Processore: Pentium III o superiore.
RAM: 64 MB.
Spazio su disco: 300 MB

9.2 L'INTERFACCIA DI INSTALLAZIONE E ACCESSO AI CONTENUTI

Il CD-ROM di installazione di *Corpus Artesia* 2015 è provvisto di un'interfaccia grafica (fig. 1) per l'installazione di tutti gli strumenti necessari alla corretta configurazione del sistema (compresa una versione su PDF di questa guida).

9.2.1 Avviare l'interfaccia da CD-ROM

L'interfaccia d'installazione si avvia automaticamente qualche istante dopo l'inserimento del disco di installazione.

Su alcuni sistemi l'avvio automatico da CD-ROM potrebbe essere disabilitato. In questo caso basta aprire la cartella principale del CD ed eseguire il file *artesia_avvia.exe*

Fig. 1 – L'interfaccia di installazione

9.2.1 Installazione on-line

Alcuni tra i più recenti modelli di computer portatili non sono dotati di lettore per CD e DVD-ROM. In questo caso è possibile scaricare la propria copia del *Corpus Artesia* dall'indirizzo riportato sulla custodia del CD-ROM o sulla quarta di copertina del volumetto allegato.

Per accedere ai contenuti il sito chiederà di inserire il numero di serie riportato sulla vostra copia di *Corpus Artesia 2015*. Una volta effettuato l'accesso, la pagina mostrerà la stessa schermata dell'interfaccia di installazione del CD-ROM e consentirà di effettuare le stesse operazioni con le stesse modalità riportate in questa guida[471].

[471] Con l'unica differenza che i pacchetti di installazione andranno prima scaricati in una cartella temporanea del proprio sistema. Una volta terminata la procedura, i pacchetti di installazione potranno essere cancellati.

9.2.2 L'installazione minima

Per predisporre il sistema all'interrogazione del *Corpus Artesia* si dovranno portare a termine due operazioni preliminari:

- Installazione del motore di interrogazione GATTO (Gestione degli Archivi Testuali del Tesoro delle Origini), sviluppato da Domenico Iorio-Fili per l'Opera del Vocabolario Italiano[472].
- Installazione dei file del *Corpus Artesia 2015*.

Entrambe le operazioni sono state semplificate per mezzo di due procedure di installazione guidata che verranno descritte nei prossimi due paragrafi.

9.3 INSTALLAZIONE DI GATTO

Per avviare l'installazione di GATTO, cliccare sul bottone "Installa GATTO" nell'interfaccia del CD-ROM[473] e accettare le condizioni di licenza selezionando "si" .

Alla schermata successiva il programma chiederà di chiudere altre eventuali applicazioni aperte prima di procedere con l'installazione. Dopo aver verificato che l'installazione di GATTO sia l'unica applicazione in esecuzione, cliccare su "ok" e, alla

[472] http://www.ovi.cnr.it

[473] Nelle edizioni su CD-ROM del *Corpus,* GATTO viene fornito nella versione più recente disponibile al momento della pubblicazione. In ogni caso è possibile, scaricare una copia aggiornata del programma dal sito dell'Opera del Vocabolario Italiano (http://ovi.cnr.it). L'eventuale compatibilità del *Corpus* con versioni di GATTO differenti da quella di riferimento non può essere garantita, specialmente per le versioni successive alla data di pubblicazione del *Corpus*.

finestra successiva, premere il pulsante con l'icona del computer (fig. 2)[474].

Alla finestra di dialogo successiva, cliccare su "Continue" per lanciare l'installazione.

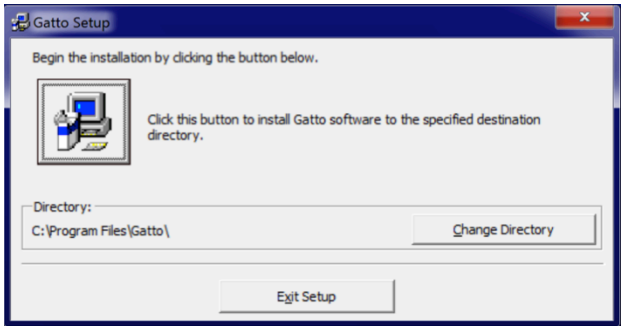


Figura 2. Avvio dell'installazione di GATTO

Durante il processo di configurazione di GATTO è molto probabile che venga visualizzata, anche più di una volta, una finestra di dialogo indicante un conflitto di versione ("*version conflict*") come quella in figura 4. Si tratta di un normale avviso che segnala la presenza nel sistema di una diversa versione delle librerie che servono per il corretto funzionamento del programma. Come raccomandato nel messaggio riportato dalla finestra, selezionare "Yes" per proseguire.

[474] Da questa schermata è possibile anche personalizzare il percorso di installazione, ma la scelta consigliata è di utilizzare le impostazioni predefinite.

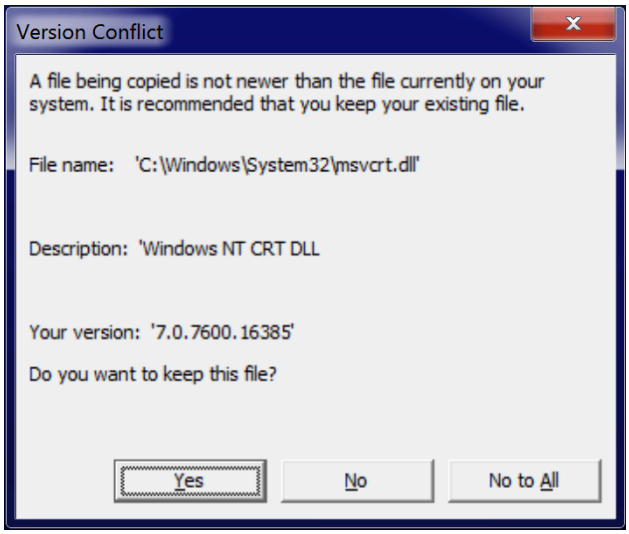


Figura 4 – Segnalazione di un conflitto di versione

9.4 INSTALLAZIONE DEL *CORPUS ARTESIA 2015*

Per facilitare l'installazione del *Corpus Artesia* è stato realizzato di uno strumento di installazione semplificata. Per copiare sul sistema tutti i file necessari per utilizzare il *Corpus*, basta cliccare sul bottone "Installa Corpus" nella schermata principale dell'interfaccia grafica del CD-ROM e seguire la procedura guidata.

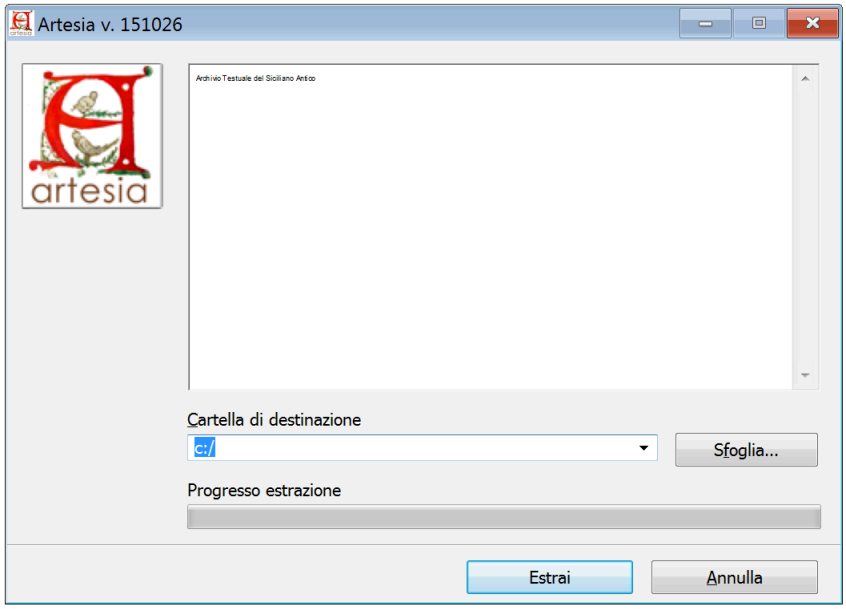


Figura 5 – Installazione del Corpus

Per avviare la copia dei file cliccare su "Installa"[475] (fig. 5). Al termine del processo la finestra si chiuderà automaticamente. Su alcuni sistemi potrebbe accadere che, prima dell'avvio dell'installazione del *Corpus*, appaia il messaggio "Impossibile verificare l'attendibilità dell'autore. Eseguire il software?". In questo caso il messaggio appare per motivi puramente tecnici e non denota un reale rischio per la sicurezza. Selezionare "Esegui" per avviare la procedura di installazione.

[475] Si consiglia di mantenere invariata la "Cartella di destinazione". Cambiare la cartella di destinazione predefinita (c:/) è possibile, a patto che si indichi la directory principale di una delle unità connesse al sistema (una partizione del disco rigido locale, un disco esterno, una penna USB, ecc.). Se installato in una posizione differente, il *Corpus* non verrà rilevato da GATTO.

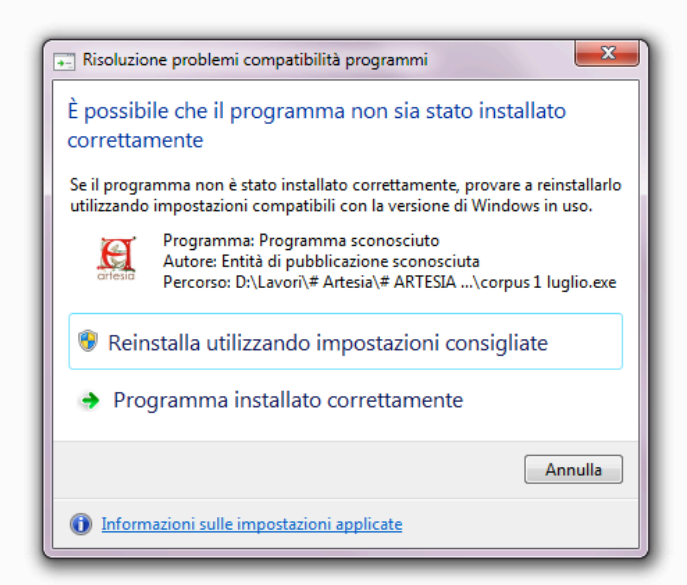


Figura 6 – Avviso di compatibilità

In alcuni casi è possibile che al termine del processo appaia la finestra in fig. 6, «è possibile che il programma non sia installato correttamente», anche a fronte di una corretta installazione. Selezionare l'opzione "Programma installato correttamente" per terminare la procedura.

10 INTERROGAZIONE

10.1 CONCETTI DI BASE

Il *Corpus Artesia 2015* è interrogabile per forme, ovvero per singole parole che si distinguono in virtù del loro aspetto grafico. Questa modalità di ricerca individua due concetti fondamentali che sarà utile chiarire preventivamente.

Occorrenza (***token***): è l'unità minima dei corpora elettronici. Ogni testo è costituito da una sequenza di parole, separate da spazi e segni di punteggiatura, ognuna delle quali costituisce un *token* o occorrenza.
Si consideri l'incipit dell'*Istoria di Eneas*, di Angilu di Capua, ms. A [476]:

In la memoria di li homini divinu sempri essiri li

excelenti facti et virtusi operi di li antiqui Romani

La stringa in esempio è composta da 19 occorrenze o *token* ("*in*", "*la*", "memoria", "*di*", ecc.), una per ciascun segmento testuale separato da spazi.

Forma (***type***): la forma è una parola che si distingue dalle altre esclusivamente per la sequenza di caratteri dai quali è composta. La forma si situa a un livello di astrazione più alto rispetto alle singole realizzazioni concrete (le occorrenze) attraverso le quali appare nei testi. Ritornando all'incipit del'*Eneas*, le quattro occorrenze di *il* sono realizzazioni di un'unica forma, così come le due occorrenze

[476] EneasXIVF.

di *di*. Il lacerto riportato conta quindi 19 occorrenze (token) ma solo 16 forme (*type*).

Gli omografi contano come unica forma, quindi una ricerca per forme di *divinu*, restituirebbe sia il *divinu* (ind. pres. 3° pers. plur. del v. *diviri*, 'dovere') rilevato nell'esempio precedente, sia la forma *divinu* (agg. m. sing., 'divino').

Nella sua interezza l'intero *Corpus Artesia 2015* è costituito da 1.148.564 forme e 69.661 occorrenze.

10.2 APERTURA E CHIUSURA DEL PROGRAMMA E DEL CORPUS

GATTO crea, in fase di installazione, un'icona sul *desktop* e una nel menù *Start* di Windows. Da questo collegamento è possibile avviare il programma.

Fig. 7 – Finestra principale di GATTO

Il programma va chiuso esclusivamente attraverso la voce "Fine" nella barra dei menu nelle schermate principali (fig. 7). Il pulsante chiudi nella barra del titolo della finestra (in alto a destra e in rosso in fig. 7) è disabilitato; le interruzioni dell'esecuzione del

programma effettuate senza usare il comando apposito potrebbero compromettere l'integrità delle banche dati aperte.

Una volta avviato GATTO selezionare l'ambiente "Ricerche" (prima voce della barra dei menu in fig. 6), quindi cliccare su "Corpus" (prima voce della barra dei menu in fig. 7).

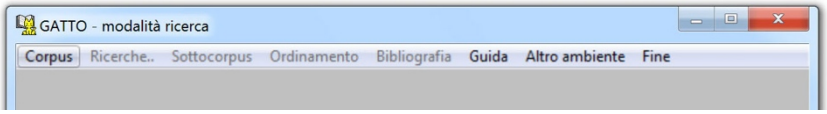


Figura 7 – Barra dei menu in modalità ricerca

Nella schermata di apertura del *Corpus*, selezionare la voce "Artesia" (in rosso in fig. 8) e cliccare sulla voce "Apri" (la prima a sinistra della barra dei menu in fig. 8).

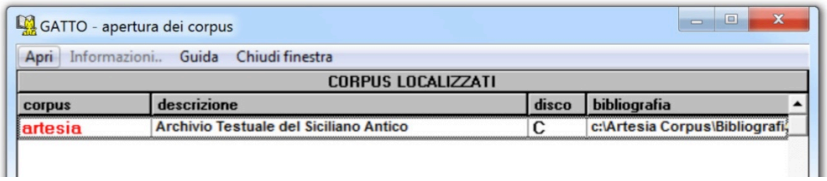


Figura 8 – Apertura del corpus

Attenzione: dopo aver cliccato su "Apri", nella barra dei menu si accenderà la voce "Informazioni" (la seconda da sinistra, in grigio, in fig.8). Questo è l'unico segnale dell'avvenuta apertura del *Corpus* e potrebbe passare inosservato, come spesso avviene quando si usa il programma per le prime volte, lasciando l'utente bloccato sulla schermata con la sensazione che non sia successo nulla. Il *Corpus*, invece, è già aperto. Selezionando "Informazioni" > "Statistiche", si potranno visualizzare i dati relativi al numero di occorrenze, numero di forme e numero di testi contenuti. Una volta aperto il *Corpus Artesia 2015*, per uscire questa pagina di selezione, cliccare sulla voce "Chiudi Finestra" (l'ultima da sinistra nella barra dei menu in fig. 8).

10.3 RICERCA SEMPLICE

Una volta chiusa la finestra di selezione del *Corpus*, selezionare la voce "Ricerche" dalla barra dei menu (seconda da destra in fig. 9), quindi la voce "per forme" dal menu a tendina (selezionata in azzurro in fig. 9).

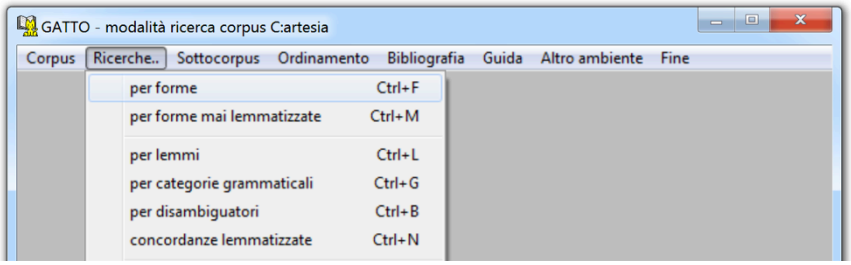


Figura 9 – Selezione della modalità di ricerca

Sono disponibili due modalità di immissione delle chiavi di ricerca: *Input stringhe* e *Input brano*.

10.3.1 Input stringhe

La finestra di ricerca *Input stringhe* (fig. 10) si aprirà automaticamente una volta selezionata la ricerca per forme ed è composta da 10 campi, chiamati *selettori*, ciascuno dei quali può contenere una chiave di ricerca per lanciare più interrogazioni in contemporanea.

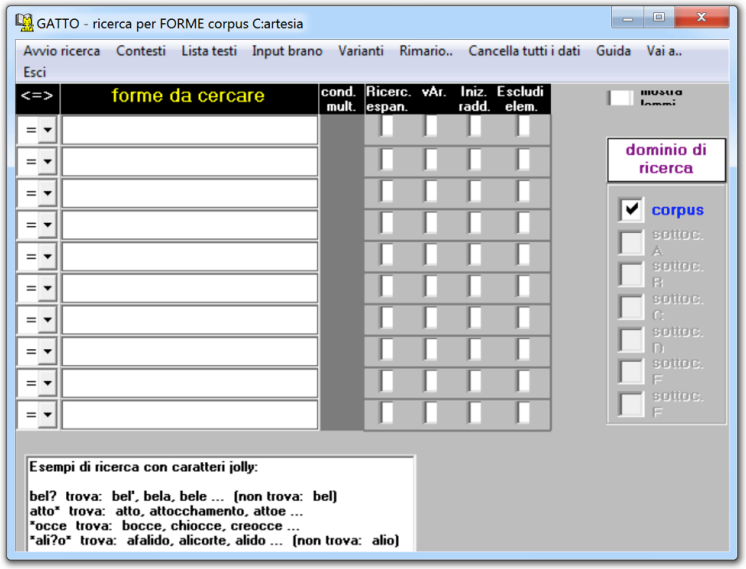


Figura 10 – Modalità di ricerca *Input stringhe*

Si può utilizzare solo un selettore qualsiasi o più selettori in contemporanea. Si immagini di voler ricercare contemporaneamente due termini dal significato simile, *strapuntinu* e *mataraciu*; inserendo le due forme in due selettori, come in fig. 11, il programma restituirà i risultati relativi a entrambe le forme.

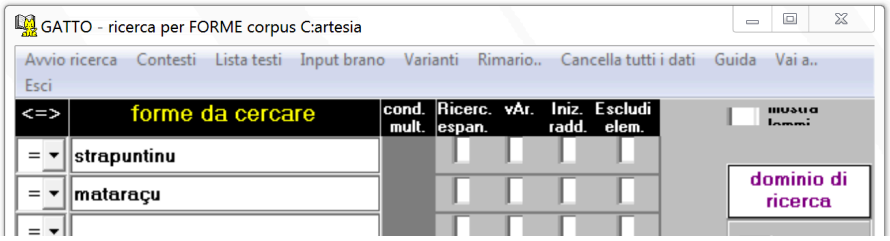


Figura 11 – Immissione di chiavi multiple

10.4 INPUT BRANO

Selezionando *Input brano* dalla barra dei menu, apparirà un campo testo in cui ogni parola inserita diventerà oggetto di una ricerca separata. Si immagini di voler cercare tutte le forme presenti in un frammento di un inventario (Bresc/2014 393) – 1444 Inventario) al fine di visualizzare rapidamente la loro distribuzione nel *Corpus*; inserendo nel campo il testo dell'inventario, il programma ricercherà tutte le parole contenute nel brano.

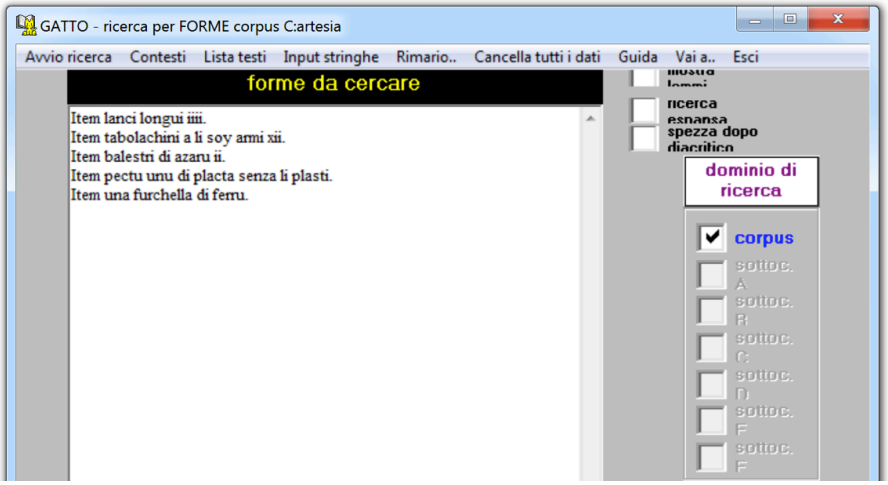


Figura 12 – Finestra *Input brano*

10.5 AVVIARE LA RICERCA

Una volta impostata la ricerca si possono imboccare tre percorsi:

1. Percorso completo: prevede due fasi di visualizzazione e raffinamento delle forme localizzate (pannello di selezione dei risultati e accumulatore) e si conclude con i contesti estratti.
2. Il passaggio diretto ai contesti estratti dalla ricerca
3. La visualizzazione della lista di testi localizzati nella ricerca.

Le tre scelte possono essere fatte attraverso le prime tre voci della barra dei menu: *Avvio ricerca*, *Contesti*, *Lista testi*.

10.6 VISUALIZZAZIONE E RAFFINAMENTO DEI RISULTATI

Lanciata la ricerca con "Avvio ricerca", il programma mostrerà una tabella con il numero di occorrenze trovate per ognuna delle forme localizzate nel *Pannello di selezione*. In questo pannello possono essere selezionate i risultati estratti dalla ricerca e rimuovere eventuali forme indesiderate.

Per es., cercando in due selettori diversi le forme *focu* e *acqua*, si otterrà:

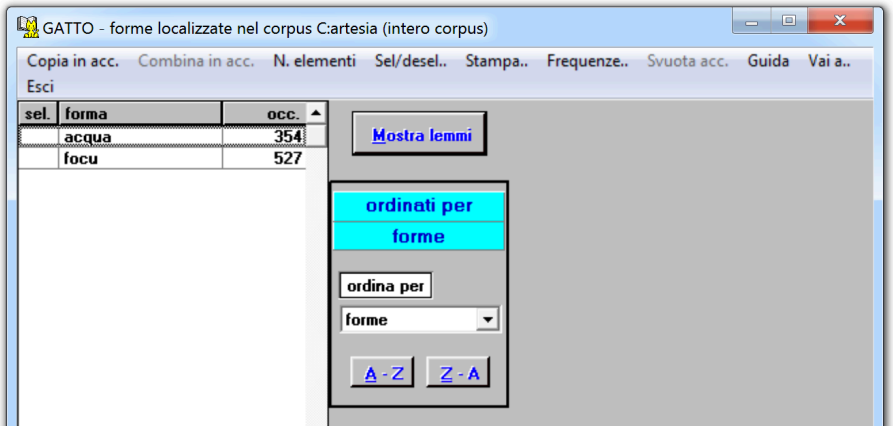


Figura 13 – Forme localizzate

Il programma segnala che sono state trovate 881 occorrenze totali, 354 per *acqua* e 527 per *focu*[477]. In questa finestra possiamo ordinare le occorrenze dal pannello di ordinamento a destra della casella con i risultati e selezionare le forme sulle quali vogliamo continuare a lavorare.

10.7 GENERAZIONE DEL FORMARIO E DELL'INDICE DI FREQUENZA

Cercando * con uno dei sistemi previsti, il pannello di visualizzazione dei risultati mostrerà il *formario* del *Corpus*, cioè l'insieme di tutte le forme che il corpus contiene. Cliccando su *N. elementi*, dalla barra dei menu, possiamo subito visualizzare il totale delle forme, 69.517. Ordinando per *forme* (*A-Z*) dal box di ordinamento alla destra della tabella, il formario si disporrà in ordine alfabetico.

Dallo stesso riquadro è possibile ordinare le quasi 70.000 forme per numero di occorrenze (ordine Z-A), ottenendo un **indice**

[477] La somma delle occorrenze, il numero di forme trovate e il numero di forme selezionate si ottiene dalla voce *N. elementi* della barra dei menu.

di frequenza assoluta decrescente, cioè una lista di parole ordinata in base al totale delle volte in cui compare nel *Corpus*.

Nell'indice di frequenza decrescente del Corpus Artesia 2015, di cui si riporta la prima parte, la forma di *primo rango*, ovvero quella con il numero più elevato di occorrenze, è *di*.

forma	occ.
di	57445
et	53586
lu	41773
la	36055
li	34507
a	23568
in	19659
ki	18894
per	18577
non	13405

10.8 SELEZIONE DELLE FORME E PASSAGGIO ALL'ACCUMULATORE

La finestra delle forme localizzate permette inoltre di selezionare le forme desiderate escludendo eventuali elementi non previsti. Cliccando nel riquadro bianco della colonna *sel*. a sinistra della voce corrispondente, la voce verrà selezionata e nel riquadro apparirà una *x* rossa. Per selezionare più di una voce, mentre si clicca, basterà tenere schiacciato il tasto CTRL nella tastiera. In alternativa è possibile spuntare la casella *blocco selezionati* (sotto il riquadro di ordinamento) per bloccare gli elementi già contrassegnati e selezionarne di nuovi con il solo click del mouse.

Terminata la cernita, le voci selezionate potranno essere passate all'accumulatore selezionando la prima voce della barra dei menu "Copia in acc".

Se l'accumulatore non è vuoto, verrà abilitata la voce "Svuota acc." nella barra dei menu. Con questo comando è possibile eliminare tutte gli elementi presenti nell'accumulatore prima di immettere le nuove forme.

10.9 ACCUMULATORE

L'accumulatore permette di selezionare e combinare i risultati di tutte le ricerche recenti. Il funzionamento è simile a quello della finestra delle forme localizzate e anche qui si disporrà del pannello di ordinamento, della casella per il blocco dei record selezionati, le funzioni di stampa e il comando per svuotare l'accumulatore.

10.10 VISUALIZZAZIONE DELLE CONCORDANZE

Le concordanze sono elenchi di parole, ordinate secondo criteri diversi in cui, per ogni parola, è mostrato un riferimento che ne consenta l'individuazione nel punto in cui appare nell'edizione di riferimento e da un contesto linguistico per permetterne l'interpretazione del significato[478].

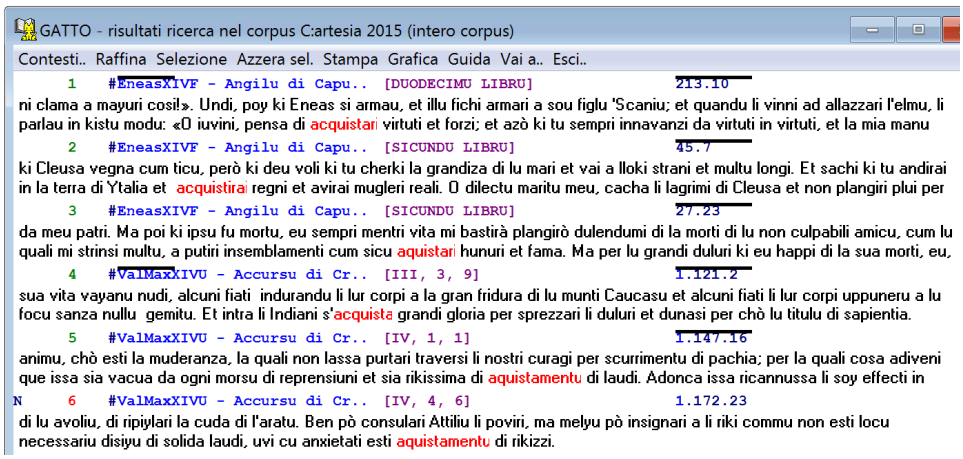


Figura 14 – Visualizzazione dei contesti

[478] Qualora vengano superate i 32.000 contesti, appariranno due barre di scorrimento. La barra a destra, permette di muoversi tra gruppi di 32.000 contesti La prima a sinistra permette lo scorrimento normale all'interno di ciascun gruppo.

Nella prima riga di ciascun contesto vengono mostrati, da sinistra a destra:

- Il numero progressivo del contesto (in verde in fig. 14).
- Titolo abbreviato (in blu in fig. 14). Cliccando su questa stringa, verrà visualizzata la scheda con i dati bibliografici e statistici relativi al testo.
- Riferimento organico (in viola in fig. 14). Cliccando su questa stringa, verrà visualizzato il riferimento organico completo relativo al contesto.
- Numero di verso (in verso, solo per i brani in versi).
- Le intestazioni di *Corpus Artesia* si chiudono con il riferimento all'eventuale numero di volume, al numero di pagina e all'eventuale numero di colonna.

Cliccando sul numero di contesto, questo viene selezionato per l'esportazione (la selezione verrà segnalata da un carattere in blu, come nel contesto numero 6 in fig. 14).

In alcuni contesti potrebbe apparire il simbolo "A", in marrone, a indicare la presenza di eventuali note in GATTO. Cliccando su questo simbolo sarà possibile visualizzare le note associate al testo.

Dalla barra dei menù, selezionando "contesti" > "tipo", sarà possibile selezionare la modalità *Kwic* (*Key Word in Context*); in questa modalità i contesti saranno selezionati incolonnando la forma oggetto di ricerca.

Cliccando su un punto qualunque del testo riportato dal contesto si aprirà la finestra di visualizzazione per il *Contesto Singolo*. Usando i comandi "allarga" e "restringi" è possibile agire sulla lunghezza del contesto visualizzato. L'allargamento massimo previsto è di 15 periodi per contesto limitato, in ogni caso, dai confini del riferimento organico corrente.

11 APPROFONDIMENTI

11.1 RICERCA AVANZATA (CARATTERI JOLLY)

? Il punto interrogativo sta per un carattere qualunque, non nullo. Es. *sichilian?* restituirà *sichiliani* (39) e *sichilianu* (1).

* L'asterisco trova qualsiasi stringa, anche vuota. Es. *attra** trova *attrahiri* (1), *attrahyri* (1) e *attrattu* (1).

[…] Le parentesi quadre permettono di specificare un insieme di lettere o un intervallo da includere nella ricerca. Es. b[li]ancu trova blancu (160) e biancu (13).

< … > Le parentesi angolari consentono di specificare più stringhe alternative di lunghezza diversa, anche nulla, separate da virgole. Possono includere i caratteri jolly ? e *. Es. *abra<cz, z>ari* trova *abrazari* (10) e *abraczari* (4). La stringa può essere anche nulla, per es. *abra<c,>zari* ottiene lo stesso risultato di prima.

11.2 COMBINARE I SELETTORI DI RICERCA

Cliccando nella colonna grigia *cond. mult.* a destra dei selettori, il programma troverà solo i risultati che soddisfano entrambe le condizioni. Si veda un esempio classico dove la combinazione dei due campi riesce a produrre risultati impossibili da ottenere su un campo singolo, anche ricorrendo ai caratteri jolly[479]:

[479] La difficoltà è data dalle due *a*. Il ricorso al carattere jolly su una sola stringa (ama*ari) individuerebbe tutte le forme in fig. 12 meno *amari* poiché, anche se l'asterisco può indicare una stringa vuota, si la presenza delle due a troverebbe, al massimo, *amaari*.

Figura 11 – Immissione di chiavi multiple

L'asterisco indica una qualunque stringa di caratteri, e individua un insieme di forme. La ricerca troverà tutte le parole che cominciano per *ama* e finiscono per *ari*, comprese le 136 occorrenze di *amari*.

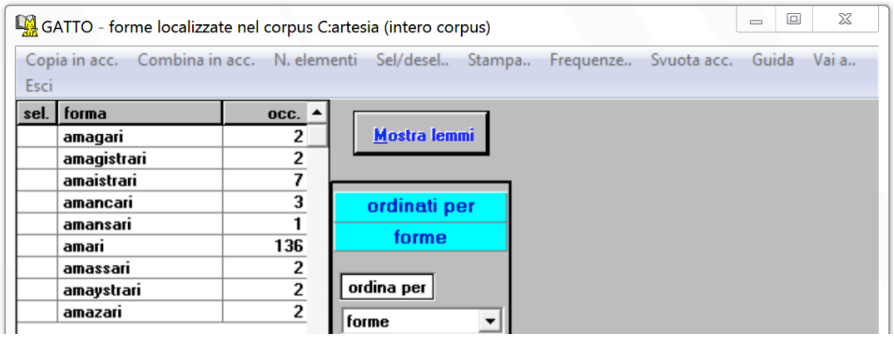


Figura 11 – Immissione di chiavi multiple

11.3 ALTRE OPZIONI DELLA FINESTRA *INPUT STRINGHE*

11.3.1 Opzione minore, uguale, maggiore

A sinistra di ciascun selettore (fig. 10) è presente un piccolo menu a tendina con 3 opzioni. Di *default* la ricerca è impostata su = e cerca tutte le forme uguali alla stringa inserita, o, più precisamente, alle forme che soddisfano le condizioni descritte dal contenuto del campo. Con le opzioni < e >, il programma utilizza la ricerca nel campo come limite alfabetico e trova tutte le forme a questo alfabeticamente precedenti o successive. Ad es. la ricerca di *abalcau*, una delle prime forme in ordine alfabetico nel *Corpus Artesia 2015*, restituisce *a*, *a·*, *a'*, *aaron*, *ab*, *abactiri*, *abactiu*, *abactuti*. Due campi possono essere utilizzati in combinazione per descrivere i due termini di un intervallo. Per es. con il

primo campo impostato su > di *a* e il secondo su < di *d*, si produrrà il formario di tutte le forme presenti sul *Corpus* che in un ordinamento alfabetico si collocano tra *a* e *d*. Come ovvio, in mancanza di un'opzione ≥, la forma *a* non verrà inclusa nell'intervallo. Per generare il formario completo delle parole che cominciano per *a*, con *a* compresa, si potrà usare un terzo selettore per includere anche quest'ultima forma (figura 12).

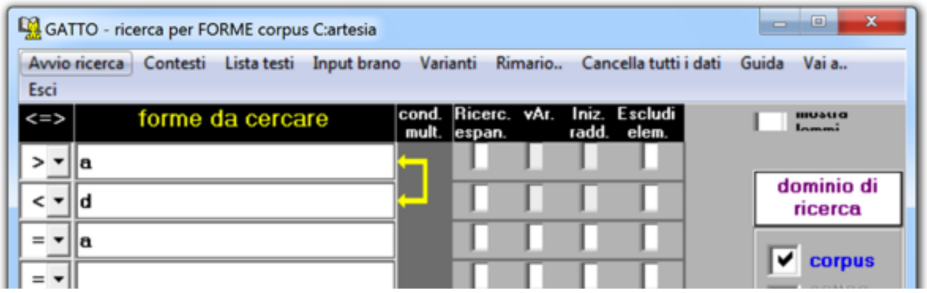


Figura 12 – Ricerca delle forme comprese tra *a* e *b*

11.3.2 Ricerca espansa / Iniziale raddoppiata / Escludi elemento

La ricerca espansa rende la ricerca insensibile ad accenti, cediglie e pseudocaratteri. Ad es. cercando *a* con la ricerca espansa, GATTO restituisce le occorrenze relative ad *a* (23630 occorrenze), *a·* (1 occorrenza), *a'* (3 occorrenze), à (393 occorrenze), à' (1 occorrenze) e á (1 occorrenze).

Con "iniziale raddoppiata" il programma cercherà, oltre alla forma indicata nel campo, anche l'eventuale forma con l'iniziale raddoppiata (es. *salvacioni* restituirà *salvacioni* (11) e *ssalvacioni* (2).

L'opzione "escludi elemento" sottrae dai risultati della ricerca il risultato restituito dal selettore.

11.4 RICERCA LIMITATA AL RIMARIO

Dalla barra dei menu selezionare *Rima*, quindi *Fine verso*. Il programma, nella finestra dei contesti localizzati, segnalerà le

occorrenze sull'intero *Corpus* ma nei contesti mostrerà solo le occorrenze a fine verso.
Es. immettendo la chiave di ricerca *avinimenti*, dopo aver attivato l'opzione *Fine verso*, GATTO segnala 8 occorrenze. Aprendo la finestra dei contesti si troverà una sola occorrenza in SonLibArbXIVC.

11.5 ESPORTAZIONE DEI RISULTATI E IMPAGINAZIONE DEI FORMARI

Nella barra dei menu della finestra di selezione dei risultati e nell'accumulatore, è presente un pulsante "Stampa" per mezzo del quale si possono stampare o esportare i risultati visualizzati nelle rispettive finestre. Dalla tendina, selezionando "lista forme" (per tutte le righe o per le righe selezionate) si otterrà un file *rtf* con le forme localizzate in corsivo e separate da virgola ("*acqua*, *focu*"). Selezionando dalla finestra delle forme localizzate *Tutte le righe* o *Righe selezionate* verrà creato un file con la seguente forma:

corpus ARTESIA Forme localizzate 01/01/2015

forma	occ.
acqua	354
focu	527

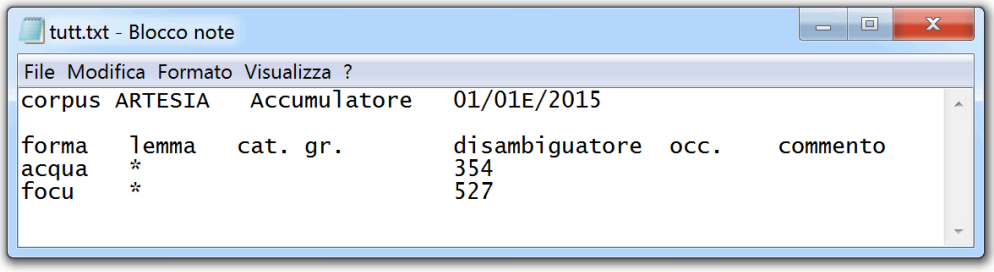
Diverso, invece, il risultato ottenuto eseguendo la stessa procedura dall'accumulatore:

corpus ARTESIA Accumulatore 01/01/2015

forma	lemma	cat. gr.	disambiguatore	occ.	commento
acqua	*		354		
focu	*		527		

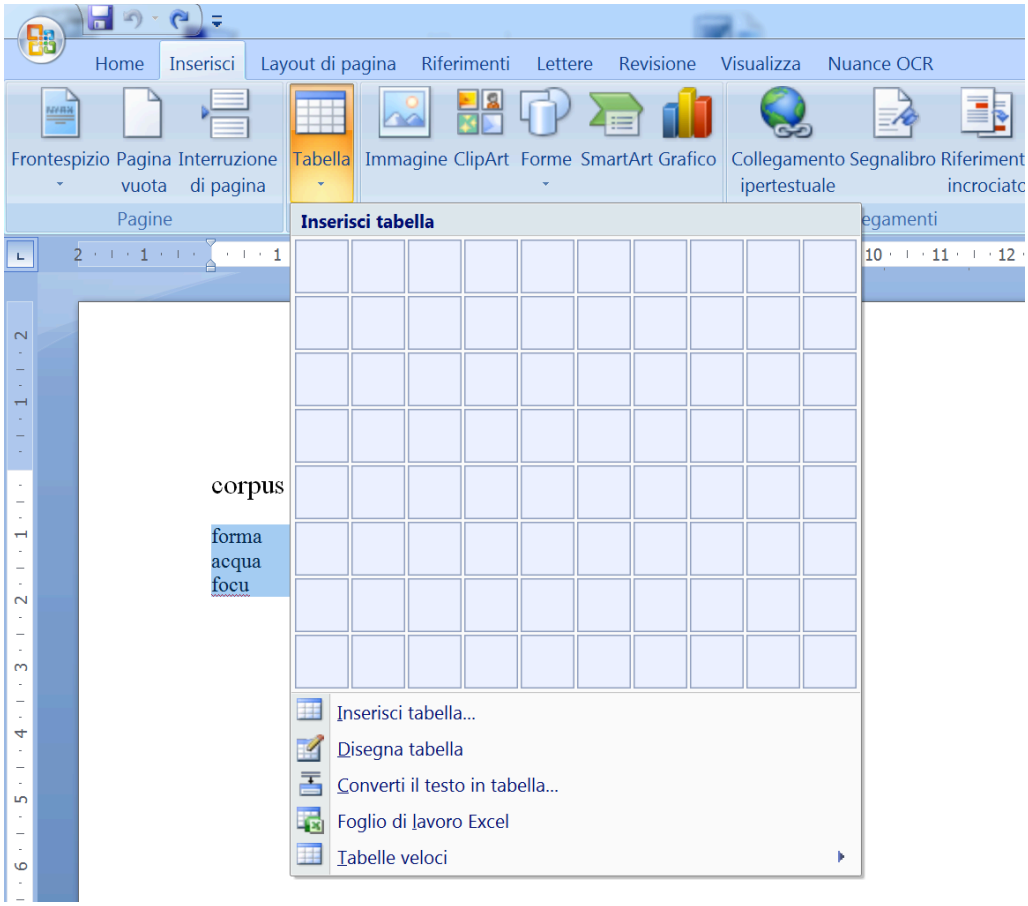
Come si può notare dall'esempio precedente, l'incolonnamento non è preciso e il numero di occorrenze viene visualizzato sotto *disambiguatore*. Al fine di renderne più agevole i trattamenti successivi, i dati esportati dal *Corpus* sono separati tramite tabulatori (Tab ⇆); la

tabulazione, nata sulle vecchie macchine da scrivere per agevolare l'incolonnamento del testo, viene gestita in modo differente in base ai programmi che aprono il file: i software destinati all'utilizzo di testo semplice (come Blocco note) li trattano in modo rigido, mentre i programmi Wordpad e Word hanno funzioni complesse che ne alterano la dimensione. Le informazioni sono tuttavia separate correttamente, lo si può verificare salvando in formato txt e aprendo con blocco note:

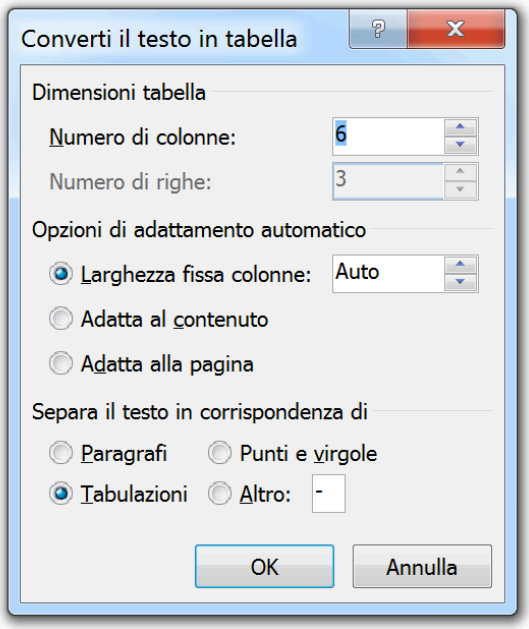


I dati separati da virgola e da tabulatori sono molto semplici da importare in altri programmi, come database o fogli di calcolo, e rendono i dati esportati da GATTO idonei per il trattamento su altri software o piattaforme.

Riformattare il documento Word per creare formari perfettamente incolonnati è estremamente semplice. Uno dei metodi consiste nel selezionare il testo da allineare, dalle intestazioni all'ultima voce, andare su *Inserisci* > *Tabella* > *Converti il testo in tabella*.



Nella finestra di dialogo spuntare, nella sezione *Separa il testo in corrispondenza di*, spuntare l'opzione *tabulazioni*.



11.6 DEFINIZIONE DI SOTTOCORPORA

Per accedere alla finestra per la definizione di sottocorpora in GATTO, occorre selezionare l'ambiente ricerche, aprire il *Corpus Artesia*, e quindi spostarsi su "Sottocorpus"; in GattoWeb la pagina corrispondente è raggiungibile dal menu "Altre funzioni...".

Il sistema permette di impostare un massimo di 6 raggruppamenti, identificati da una lettera da A a F. All'apertura della finestra il corpus A sarà selezionato di default, ma è possibile spostarsi su una qualsiasi delle sei posizioni per mezzo del riquadro in basso a destra.

La definizione del raggruppamento di testi viene effettuata impostando dei criteri di appartenenza all'insieme, specificando una o più condizioni sui metadati bibliografici. Per isolare tutti i testi in versi, ad esempio, basterà inserire "v" nel campo "Forma" e quindi cliccare su "Opera selezione..." > "nuova" (su GattoWeb il tipo "v" deve essere selezionato dal menu a tendina associato al campo "Forma").

Concluse queste operazioni, GATTO mostrerà una schermata contenente le informazioni bibliografiche sui testi selezionati e, da questo momento in poi, per ogni ricerca sarà possibile scegliere se interrogare l'intero corpus o restringere il dominio di ricerca ai soli testi in versi. Nel caso in cui siano stati definiti più sottocorpora, il sistema consentirà scelte multiple e, in presenza di testi inclusi in più di un gruppo, provvederà a evitare contesti duplicati.

Aprendo la finestra delle statistiche per il sottocorpus dei testi in versi così creato, si ricaveranno le seguenti informazioni:

Numero testi: 23

Numero occorrenze: 49.322

Numero forme: 7.921

11.7 ALTRE INFORMAZIONI

Si possono mettere in esecuzione contemporaneamente più istanze di GATTO per interrogare lo stesso corpus o corpora diversi. Per l'utilizzo di più istanze al di fuori dell'ambiente ricerche fare riferimento al manuale del programma.

11.8 ACCORDO DI LICENZA

GATTO è un programma registrato ed è protetto dalle vigenti leggi italiane ed internazionali sulla tutela del software.

GATTO può essere liberamente usato e riprodotto per scopi di studio e ricerca senza fini di lucro, con l'obbligo della citazione in tutte le pubblicazioni che ne derivino e in tutte le applicazioni nelle quali venga utilizzato.

Qualunque altro uso deve essere oggetto di un accordo scritto preventivo con l'Istituto Opera del Vocabolario Italiano.

(GATTO - Copyright 8-2-99 - Registrazione n. 001172.)