



# Verso una lessicografia elettronica semiautomatica? ChatGPT per il mark-up in XML-TEI: primi test su raccolte paremiologiche

Barbara Patella

L'articolo prende le mosse da una delle principali criticità che rendono piuttosto lenta e complessa l'informatizzazione delle risorse lessicali e lessicografiche: il mark-up manuale in XML-TEI. Accanto agli indubbi vantaggi offerti dalle versioni elettroniche – in primis l'accessibilità e la valorizzazione delle risorse – la conversione dei dizionari di seconda generazione (specialmente quelli di secondo livello, che prevedono l'informatizzazione del testo e della sua struttura) si rivela spesso onerosa in termini di tempo e risorse.

A partire da questa premessa, la ricerca propone un approccio sperimentale che prevede l'impiego dell'intelligenza artificiale – nello specifico ChatGPT – per marcare i testi in XML-TEI in modo semiautomatico: dopo una fase preliminare di formalizzazione del testo (il cosiddetto *file pre-markup*), orientata all'individuazione di pattern ricorrenti, il modello viene guidato attraverso istruzioni mirate per generare file .xml conformi alle linee guida TEI, con l'obiettivo di ottenere risultati di livello medio-alto pur riducendo significativamente i tempi di lavorazione, e favorire così la sostenibilità di progetti di lessicografia elettronica su larga scala.

*This paper starts from one of the main critical issues that slow down and complicate the computerizing lexical and lexicographic resources: manual XML-TEI markup. Alongside the undeniable advantages offered by electronic versions – notably in terms of accessibility and resources enhancement – the conversion of retro-digitized dictionaries (especially second-level dictionaries, which involve the digital encoding of both the text and its internal structure) often proves time-consuming and resource-intensive.*

*Building on this premise, the study proposes an experimental approach involving the use of artificial intelligence – specifically ChatGPT – to perform semi-automatic XML-TEI markup. Following a preliminary phase of text formalization (the so-called pre-markup file), aimed at identifying recurring structural patterns, the model is guided through targeted instructions to generate .xml files compliant with TEI guidelines. The ultimate goal is to produce medium-to-high quality outputs while significantly reducing processing time, thereby promoting the sustainability of large-scale electronic lexicography projects.*

**Parole chiave:** mark-up, XML-TEI, ChatGPT, lessicografia elettronica, dizionari

**Keywords:** mark-up, XML-TEI, ChatGPT, electronic lexicography, dictionaries

**Sommario:** 1. Introduzione: vantaggi e svantaggi della lessicografia elettronica (ad oggi) - 1.1 Una premessa terminologica e metodologica: *repertori elettronici/informatizzazione vs repertori digitali/digitalizzazione* - 2. Una nuova soluzione per il mark-up semiautomatico: l'intelligenza artificiale (IA) - 2.1 Il flusso di lavoro semiautomatico con ChatGPT: due casi di studio - 2.2 "Pre-marcatatura" del testo (automatica e non): *file pre-markup* e *simboli-marcatori* - 2.3 Il mark-up automatico con ChatGPT - 2.4 I siti pilota - 3. Conclusioni: verso una nuova velocità di crociera per i progetti di lessicografia elettronica?

---

#### Peer review

Submitted 14/05/2025

Accepted 28/05/2025

Published 13/06/2025

#### Open access

© 2025 | Attribution - Non commercial - Non derivatives (IT)

---

**DOI** 10.35948/DILEF/2025.4366

## 1. Introduzione: vantaggi e svantaggi della lessicografia elettronica (ad oggi)

Negli ultimi anni nel campo dell'umanistica digitale e, più precisamente, in quello relativo alla lessicografia elettronica<sup>1</sup>, si sono compiuti notevoli passi in avanti, e non solo in merito allo sviluppo di strumenti lessicografici ex novo, ma anche rispetto a dizionari non nativi digitali, cioè *dizionari elettronici di seconda generazione*<sup>2</sup> – quelli che, secondo un'etichetta ben sedimentata nel contesto accademico internazionale, sono definiti *retro-digitized* (per distinguerli dai dizionari elettronici *born-digital*). Si tratta di opere che, nate in versione cartacea, vengono trasformate in formato elettronico per diventare fruibili e interrogabili in rete (attraverso ricerche libere, avanzate o guidate); pensiamo ad esempio, nel panorama dei dizionari italiani, all'informatizzazione del *GDLI*<sup>3</sup>, del *Tommaseo-Bellini*<sup>4</sup>, delle cinque edizioni del *Vocabolario degli Accademici della Crusca*<sup>5</sup>, dei principali dizionari metodici ottocenteschi<sup>6</sup> o del *Vocabolario toscano dell'arte del disegno* di Filippo Baldinucci<sup>7</sup>.

I vantaggi che gli strumenti informatici sono in grado di apportare alla consultabilità di simili risorse testuali (in termini sia quantitativi sia qualitativi) sono numerosi e ben noti, e potrebbero essere sintetizzati in tre parole chiave: *valorizzazione*, *estendibilità* e *accessibilità* – più esplicitamente, *valorizzazione* intesa come implementazione dinamica dei dati (con un potenziamento delle modalità di ricerca e di visualizzazione rispetto alla staticità della consultazione cartacea); *estendibilità* come possibilità di aggiornare e allargare continuamente il bacino di raccolta dei dati senza limiti di spazio fisico; *accessibilità* come fruizione libera e immediata dei testi su piattaforme online.

Tuttavia, accanto a questi, si devono considerare anche alcuni svantaggi, primo fra tutti l'elevato dispendio di tempo, che è direttamente proporzionale alla dimensione della risorsa lessicale o lessicografica (dizionari, glossari, lessici, raccolte paremiologiche). Non a caso, il processo per trasformare il dato cartaceo in dato elettronico è tutt'altro che rapido e lineare; in particolare, l'attività di marcatura in XML-TEI continua a rappresentare una delle fasi più onerose, soprattutto nel caso di testi non nativi digitali, poiché il ventaglio di difficoltà con cui il linguista deve cimentarsi è piuttosto ampio. Basti pensare, in primo luogo, alla serie di variabili determinate dalla qualità e dalla conformazione del testo cartaceo disponibile: lo stato di conservazione, il grado di leggibilità del testo, l'impostazione tipografica (su una, due o più colonne), la distribuzione dei contenuti (talvolta irregolare o disomogenea), il secolo in cui l'opera è stata redatta (linguisticamente rilevante per il riconoscimento di forme e lemmi). Un ulteriore fattore è rappresentato dalla versione del software OCR utilizzato, che incide sulla precisione del riconoscimento automatico e, di conseguenza, sulla maggiore o minore affidabilità del testo elaborato. A ciò si aggiunge ancora un altro svantaggio, spesso sottovalutato: la scarsa sostenibilità di

simili progetti, la cui durata effettiva rimane incerta. Una volta completato il lavoro di informatizzazione e pubblicazione in rete, infatti, questi strumenti rischiano di restare attivi – e quindi effettivamente funzionanti e accessibili – solo per pochi anni, in assenza di una pianificazione adeguata per l'aggiornamento e la manutenzione, necessaria a prevenire l'obsolescenza tecnologica (server dismessi, domini non rinnovati e incompatibilità con i nuovi ambienti di visualizzazione e interrogazione). Ad oggi, dunque, la trasformazione di una raccolta lessicale/lessicografica in formato elettronico risulta lunga e faticosa, soprattutto per l'informatizzazione della struttura, in cui spetta all'umanista digitale disegnare un tagset adeguato, sulla base di campi ed elementi costitutivi degli articoli lessicografici (lemmi, definizioni, esempi, etimologie, citazioni, ecc.), che andranno annotati manualmente secondo tag, attributi e valori TEI più idonei<sup>8</sup> – bisogna inoltre considerare che si tratta di scelte talvolta controverse, poiché le linee guida del consorzio tei non contemplano tutti i casi possibili oppure presentano soluzioni ambigue<sup>9</sup>. Quest'ultimo aspetto mette in evidenza un altro svantaggio: l'eterogeneità delle risorse da informatizzare.

La presente ricerca, prendendo le mosse proprio da queste criticità (che rallentano e rendono piuttosto complessa l'informatizzazione di repertori lessicografici e lessicali, col rischio di compromettere non tanto la costruzione di *repertori elettronici* tout court, ma la sostenibilità di simili progetti), mira a ridurre considerevolmente le tempistiche necessarie al processo di marcatura in XML-TEI, attraverso l'uso sperimentale dell'intelligenza artificiale come strumento di conversione automatica – nel caso specifico, ChatGPT.

## 1.1 Una premessa terminologica e metodologica: *repertori elettronici/informatizzazione* vs *repertori digitali/digitalizzazione*

Prima di vedere in che misura ChatGPT, repertori lessicali/lessicografici e XML-TEI possono interagire fra loro, è opportuno fare una duplice premessa, di natura terminologica e metodologica. In questa sede il focus sarà dedicato esclusivamente ai *repertori elettronici*: d'ora in poi ci serviremo del binomio-iponimo *repertorio elettronico* (o *dizionario elettronico*) per indicare soltanto quelle risorse sottoposte a un processo di *informatizzazione* superiore (che prevede l'elaborazione del testo e della sua struttura), in contrapposizione al binomio-iperonimo *repertorio digitale* – concettualmente vago, in quanto implica un'elaborazione informatica non definita (rischiando così di denotare elaborazioni generiche o limitate alla sola digitalizzazione delle immagini ed, eventualmente, del testo). In tal senso, seguiamo la proposta terminologica avanzata da Biffi<sup>10</sup>, secondo cui

*digitale* sembra configurarsi piuttosto come una sorta di iperonimo, per indicare dizionari composti da materiale digitale vario (immagini, ma anche i suoni che riportano in certi casi le pronunce), o che nelle versioni digitali trovano la possibilità di un terreno ampliato

di multimedialità. Ma *elettronico* rimane caratterizzante quando ci si voglia riferire a quelli che sono gli elementi realmente identificatorii in ottica propriamente lessicografica, quei fattori che hanno cambiato la fisionomia del dizionario e le sue potenzialità, vale a dire il recupero della natura di *database* e l'apertura a un'interrogazione sistematica del testo nella sua integralità. Caratteristiche per le quali non è sufficiente la registrazione in più *digit*, ma serve anche un'elaborazione più complessa attraverso procedure informatiche che da sempre vengono indicate come elettroniche.<sup>11</sup>

In linea con ciò, nella nostra trattazione – e, più in generale, nell'ambito della lessicografia elettronica italiana<sup>12</sup> – potrebbe essere utile proporre un'ulteriore polarizzazione terminologica: adoperare in maniera differenziata i vocaboli *informatizzazione* e *digitalizzazione*, eleggendo il primo come veicolo semantico più ristretto, specifico e tecnico, atto a designare una 'elaborazione informatica di livello superiore, con alto grado di granularità', rispetto al secondo, che – tendendo ad appiattirsi sul ruolo di iperonimo – finirebbe per denotare semplicisticamente un 'qualsiasi processo di conversione in digitale' (senza altri corollari informatici). Riepilogando, dunque, si propone – e si preferisce – adottare il seguente pacchetto terminologico: *dizionario/repertorio elettronico, informatizzazione, informatizzare*.

Il campo della nostra ricerca, però, non riguarda risorse elettroniche lato sensu, ma è circoscritto a quelle *di seconda generazione* (prevalentemente pre-novecentesche) *di secondo livello*, quindi a opere nate come cartacee e poi trasformate in prodotti elettronici «in cui l'informatizzazione coinvolge la struttura, attraverso l'individuazione e la marcatura dei campi della voce e la costruzione di un motore di ricerca che consente di fare ricerche mirate nei singoli campi»<sup>13</sup>. Non si farà perciò riferimento ai repertori nei quali «ci si ferma all'acquisizione del testo elettronico, possibilmente collazionato, per cui conseguentemente ci si deve accontentare di una ricerca libera, per quanto assistita da caratteri *jolly*, operatori booleani ed eventualmente strumenti statistici»<sup>14</sup> (*di primo livello*). Di conseguenza, sarà esclusa tutta quella gamma di progetti e strumenti – comunque validi e preziosi – che offre soltanto una riproduzione in facsimile secondo tre diverse modalità di consultazione, dipendenti dal livello di *digitalizzazione* applicato:

- immagini sfogliabili senza possibilità di interrogare il testo (livello zero: *digitalizzazione delle immagini senza OCR disponibile*);
- immagini sfogliabili con possibilità di fare ricerche (solo full-text) su testo non corretto (livello uno: *digitalizzazione con OCR automatico grezzo/senza controllo*), come accade nelle grandi collezioni di Google Libri e Internet Archive, in continua espansione;
- immagini sfogliabili con possibilità di ricerche (solo full-text) su testo interrogabile revisionato (livello due: *digitalizzazione con OCR corretto*).

---

## 2. Una nuova soluzione per il mark-up semiautomatico: l'intelligenza artificiale (IA)

Sebbene esistano strumenti, tecniche e strategie in grado di eseguire procedure di mark-up semiautomatico, tuttavia questi richiedono generalmente competenze informatiche avanzate e specifiche sessioni di addestramento<sup>15</sup>. Questo studio propone una soluzione alternativa, più accessibile alla figura dell'umanista digitale – che non può né deve sostituire il ruolo dell'informatico, ma può operare efficacemente in autonomia, a condizione di conoscere il linguaggio XML-TEI – e al tempo stesso efficiente, sia nella riduzione dei tempi di lavoro sia nella qualità dei risultati raggiungibili: la realizzazione di versioni elettroniche di opere lessicografiche in modo semiautomatico mediante il supporto dell'intelligenza artificiale.

### 2.1 Il flusso di lavoro semiautomatico con ChatGPT: due casi di studio

Con l'obiettivo di testare il modello su raccolte paremiologiche differenti, la ricerca ha previsto l'analisi di due casi di studio (nell'ottica di estendere il procedimento e renderlo applicabile ad altre tipologie lessicografiche): come banco di prova sono stati scelti due testi ottocenteschi, *I proverbi e le sentenze proverbiali* di Giovanni Rampoldi (Milano, 1852<sup>3</sup>) e i *Proverbi e modi proverbiali* di Augusto Alfani (Torino, 1882) – specifichiamo che i test con ChatGPT<sup>16</sup> sono stati effettuati solo sul corpo delle raccolte lessicografiche (quindi su tutto ciò che è contenuto all'interno dell'elemento <body>) e non anche sugli apparati (prefazioni, postfazioni e simili), dal momento che la parte più impegnativa riguarda proprio il mark-up manuale dell'endotesto, non quello degli esotesti<sup>17</sup>.

Si illustrerà l'intero flusso di lavoro a partire dai file .txt (derivati da OCR post-processed)<sup>18</sup>:





Fig. 1: Flusso di lavoro con ChatGPT per il mark-up automatico di raccolte lessicografiche.

## 2.2 “Pre-marcatura” del testo (automatica e non): *file pre-markup* e *simboli-marcatori*

Prima di procedere alla conversione automatica vera e propria – cioè al mark-up in XML-TEI mediante ChatGPT – è essenziale compiere due passaggi propedeutici. Il primo consiste in un’analisi visiva della struttura del dizionario (cartaceo), finalizzata a individuare coordinate di riferimento precise, cioè elementi testuali ricorrenti e regolarità nel layout: l’individuazione di tali pattern consentirà di predisporre adeguatamente il testo che ChatGPT dovrà elaborare e che definiamo, sinteticamente, *file pre-markup*. L’allestimento di questo file costituisce il secondo passaggio preliminare: si tratta di una versione formalizzata del file .txt originale, ottenuta mediante l’inserimento sistematico di *simboli-marcatori*<sup>19</sup>, impiegati per contrassegnare i segmenti testuali da isolare (singoli elementi o porzioni di testo). L’operazione di formalizzazione – che può essere svolta automaticamente (a.) o manualmente (b.) – è fondamentale per garantire un processo automatico ottimale: fornire un file nella veste più formalizzata possibile<sup>20</sup> consente, infatti, a ChatGPT di individuare più agevolmente gli elementi strutturali da marcare in modo automatico. Per esemplificare il primo step (quello relativo all’analisi visiva), si osservi la riproduzione tratta dalla raccolta di Rampoldi:

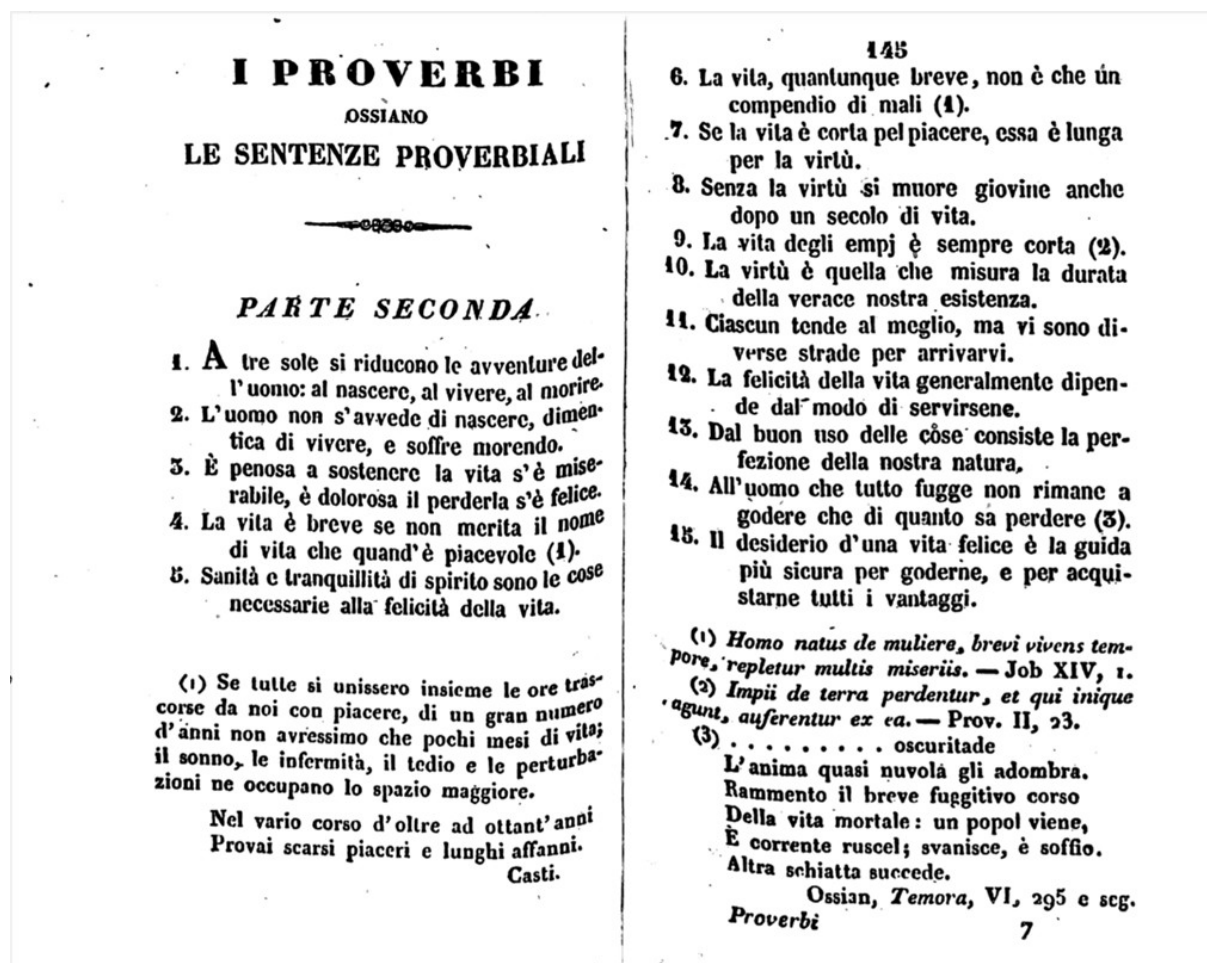


Fig. 2: Riproduzione di Rampoldi 1852, pp. 144-145.

Dal punto di vista strutturale, il testo presenta alcuni pattern ricorrenti, che in questo caso si rivelano sufficientemente regolari e omogenei da permettere la conversione automatica del file pre-markup:

- 4 sezioni: *Sentenze proverbiali* (Parte Prima e Parte Seconda); *Precetti proverbiali*; *Dialoghi proverbiali*
- proverbi numerati progressivamente:
 

*A tre sole si riducono le avventure dell'uomo: al nascere, al vivere, al morire.*  
*L'uomo non s'avvede di nascere, dimentica di vivere, e soffre morendo.*  
*È penosa a sostenere la vita s'è mirabile, è dolorosa il perderla s'è felice [...]*
- numeri di aggancio alle note a piè di pagina (posti fra parentesi tonde alla fine del proverbio):

*proverbio (1)*

*proverbio (2)*



## proverbio (3)

- note a piè di pagina (precedute da numeri fra parentesi tonde):

(1) *Se tutte si unissero insieme le ore trascorse da noi con piacere, di un gran numero d'anni non avremmo che pochi mesi di vita; il sonno, le infermità, il tedio e le perturbazioni ne occupano lo spazio [...].*

(2) *Impii de terra perdentur; et qui inique agunt, auferentur ex ea – Prov. II, 23 [...].*

(3) ... *oscuritade/ L'anima quasi nuvola gli adombra./ Rammento il breve fuggitivo corso/ Della vita mortale: un popol viene,/ È corrente ruscel; svanisce, è soffio./ Altra schiatta succede./ Ossian, Temora, VI, 295 e seg.*

Nella raccolta di Alfani, rispetto al testo di Rampoldi, si riscontrano sequenze altrettanto regolari, ma l'organizzazione strutturale è più complessa:

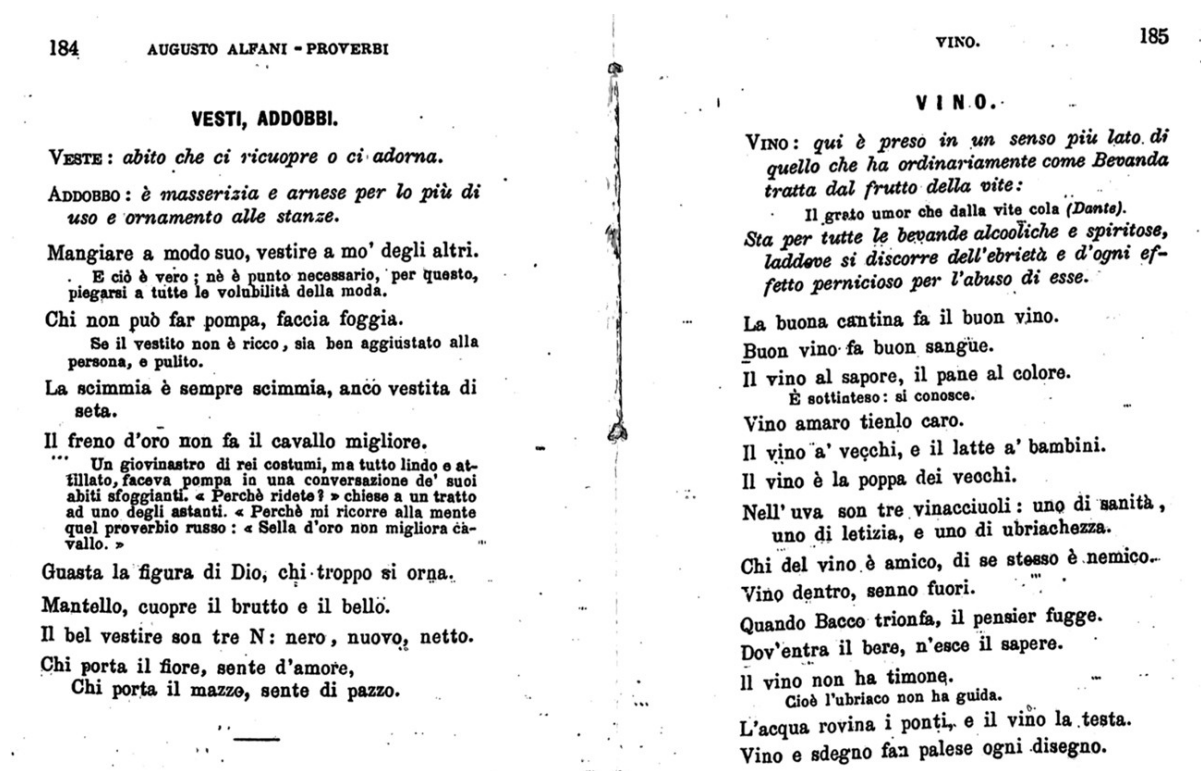


Fig. 3: Riproduzione di Alfani 1882, pp. 184-185.

Nonostante, infatti, ci siano elementi che figurano sistematicamente, tuttavia non ci sono espedienti tali (Fig. 3) da poter sovrascrivere il file pre-markup automaticamente (in altre parole, gli elementi testuali da segmentare possono essere riconosciuti solo da una lettura umana ed essere contraddistinti manualmente, non automaticamente):

- 88 sezioni tematiche:

*Agricoltura, Meteorologia, Stagioni, Economia rurale*  
*Fallacia dei giudizi*  
*Giuoco*  
*Parlare, Tacere*  
*Vesti, addobbi*  
*Tavola, Cucina [...]*

- parole chiave delle sezioni tematiche accompagnate da spiegazione ed eventuali citazioni:

*VESTE: abito che ci ricuopre o ci adorna.*

*ADDOBBO: è masserizia e arnese per lo più di uso e ornamento alle stanze.*

- serie di proverbi, di cui alcuni accompagnati da definizioni, commenti e/o citazioni:

*Mangiare a modo suo, vestire a mo' degli altri.*

*E ciò è vero; né è punto necessario, per questo, piegarsi a tutte le volubilità della moda.*

*Chi non può far pompa, faccia foggia.*

*Se il vestito non è ricco, sia ben aggiustato alla persona, e pulito.*

*La scimmia è sempre scimmia, anco vestita di seta.*

*Il freno d'oro non fa il cavallo migliore.*

*Un giovinastro di rei costumi, ma tutto lindo e attillato, faceva pampa in una conversazione de' suoi abiti sfoggianti. «Perchè ridete?» chiese a un tratto ad uno degli astanti. «Perchè mi ricorre alla mente quel proverbio russo: «Sella d'oro non migliora cavallo.»*

Una volta concluso l'esame visivo – e di conseguenza definite le opzioni di ricerca in funzione degli elementi che si intende rendere interrogabili (proverbi, definizioni, note, parole chiave, ecc.) –, il testo formalizzato può essere predisposto attraverso due modalità: una automatica (a.) e una manuale (b.), a seconda dalla configurazione grafica e strutturale della raccolta paremiologica.

a. Qualora il testo presenti elementi ricorrenti e facilmente individuabili (Fig. 2), il file pre-markup può essere generato automaticamente (ad esempio tramite uno script Python), come nel caso di Rampoldi 1852. Per segmentare il testo ci siamo serviti di uno script Python realizzato con ChatGPT stesso; nella versione modificata del file .txt (Fig. 6) vengono introdotti due simboli-marcatori (\* e #), posizionati sistematicamente a inizio riga – e che saranno funzionali all'IA per identificare i diversi elementi testuali. Lo script ha quindi permesso di eseguire una trasformazione automatica in base alle seguenti regole (estrapolate sulla scorta delle caratteristiche strutturali della raccolta paremiologica di Rampoldi):

- le righe che iniziano con un numero seguito da un punto (es. 1., 23., 150.), contenenti i proverbi, devono essere marcate con il simbolo \*;

- le righe che iniziano con un numero tra parentesi tonde (es. (1), (2) e simili), corrispondenti alle note, devono essere contrassegnate con il simbolo #;
- le righe che non corrispondono a nessuno dei due pattern restano invariate.

Di seguito il codice completo (generato tramite ChatGPT):

```
python

import re

def trasforma_file(percorso_file):
    with open(percorso_file, 'r', encoding='utf-8') as file:
        righe = file.readlines()

    righe_modificate = []
    for riga in righe:
        # Rimuove eventuali spazi iniziali per un match preciso
        riga_stripped = riga.lstrip()

        # Controlla se inizia con numero puntato (es. 1.)
        if re.match(r'^\d+\.', riga_stripped):
            nuova_riga = '* ' + riga_stripped
        # Controlla se inizia con numero tra parentesi (es. (1))
        elif re.match(r'^\(\d+\)', riga_stripped):
            nuova_riga = '# ' + riga_stripped
        else:
            nuova_riga = riga
        righe_modificate.append(nuova_riga)

    # Sovrascrive il file originale con le righe modificate
    with open(percorso_file, 'w', encoding='utf-8') as file:
        file.writelines(righe_modificate)
```

Fig. 4: Script Python per sovrascrivere automaticamente il file .txt *pre-markup*.

Il risultato della conversione del file .txt originale (Fig. 5) nel corrispondente file pre-markup (Fig. 6) è il seguente:

1. A tre sole si riducono le avventure dell'uomo: al nascere, al vivere, al morire.  
 2. L'uomo non s'avvede di nascere, dimentica di vivere, e soffre morendo.  
 3. È penosa a sostenere la vita s'è mirabile, è dolorosa il perderla s'è felice.  
 4. La vita è breve se non merita il nome di vita che quand'è piacevole (1).  
 5. Sanità e tranquillità di spirito sono le cose necessarie alla felicità della vita.

(1) Se tutte si unissero insieme le ore trascorse da noi con piacere, di un gran numero d'anni non avremmo che pochi mesi di vita; il sonno, le infermità, il tedio e le perturbazioni ne occupano lo spazio maggiore.  
 Nel vario corso d'oltre ad ottant'anni  
 Provai scarsi piaceri e lunghi affanni.  
 Casti.

6. La vita, quantunque breve, non è che un compendio di mali (1).  
 7. Se la vita è corta pel piacere, essa è lunga per la virtù.  
 8. Senza la virtù si muore giovine anche dopo un secolo di vita.  
 9. La vita degli empj è sempre corta (2).  
 10. La virtù è quella che misura la durata della verace nostra esistenza.  
 11. Ciascun tende al meglio, ma vi sono diverse strade per arrivarvi.  
 12. La felicità della vita generalmente dipende dal modo di servirsene.  
 13. Dal buon uso delle cose consiste la perfezione della nostra natura.  
 14. All'uomo che tutto fugge non rimane a godere che di quanto sa perdere (3).  
 15. Il desiderio d'una vita felice è la guida più sicura per goderne, e per acquistarne tutti i vantaggi.

(1) Homo natus de muliere, brevi vivens tempore, repletur multis miseriis. – Job XIV, I.  
 (2) Impii de terra perdentur, et qui inique agunt, auferentur ex ea. – Prov. II, 23.  
 (3) ..... oscuritade  
 L'anima quasi nuvola gli adombra.  
 Rammento il breve fuggitivo corso  
 Della vita mortale: un popol viene  
 è corrente ruscel; svanisce, è soffio.  
 Altra schiatta succede.  
 Ossian, Temora, VI, 295 e seg.

Fig. 5: File .txt derivato dall'OCR e collazionato (Rampoldi 1852, pp. 144-145).

\* 1. A tre sole si riducono le avventure dell'uomo: al nascere, al vivere, al morire.  
 \* 2. L'uomo non s'avvede di nascere, dimentica di vivere, e soffre morendo.  
 \* 3. È penosa a sostenere la vita s'è mirabile, è dolorosa il perderla s'è felice.  
 \* 4. La vita è breve se non merita il nome di vita che quand'è piacevole (1).  
 \* 5. Sanità e tranquillità di spirito sono le cose necessarie alla felicità della vita.

# (1) Se tutte si unissero insieme le ore trascorse da noi con piacere, di un gran numero d'anni non avremmo che pochi mesi di vita; il sonno, le infermità, il tedio e le perturbazioni ne occupano lo spazio maggiore.  
 Nel vario corso d'oltre ad ottant'anni  
 Provai scarsi piaceri e lunghi affanni.  
 Casti.

\* 6. La vita, quantunque breve, non è che un compendio di mali (1).  
 \* 7. Se la vita è corta pel piacere, essa è lunga per la virtù.  
 \* 8. Senza la virtù si muore giovine anche dopo un secolo di vita.  
 \* 9. La vita degli empj è sempre corta (2).  
 \* 10. La virtù è quella che misura la durata della verace nostra esistenza.  
 \* 11. Ciascun tende al meglio, ma vi sono diverse strade per arrivarvi.  
 \* 12. La felicità della vita generalmente dipende dal modo di servirsene.  
 \* 13. Dal buon uso delle cose consiste la perfezione della nostra natura.  
 \* 14. All'uomo che tutto fugge non rimane a godere che di quanto sa perdere (3).  
 \* 15. Il desiderio d'una vita felice è la guida più sicura per goderne, e per acquistarne tutti i vantaggi.

# (1) Homo natus de muliere, brevi vivens tempore, repletur multis miseriis. – Job XIV, I.  
 # (2) Impii de terra perdentur, et qui inique agunt, auferentur ex ea. – Prov. II, 23.  
 # (3) ..... oscuritade  
 L'anima quasi nuvola gli adombra.  
 Rammento il breve fuggitivo corso  
 Della vita mortale: un popol viene  
 è corrente ruscel; svanisce, è soffio.  
 Altra schiatta succede.  
 Ossian, Temora, VI, 295 e seg.

Fig. 6: File .txt trasformato in testo *pre-markup* in modo automatico (Rampoldi 1852, pp. 144-145).

**b.** Diversamente, laddove non sia possibile intervenire automaticamente (non per mancanza di pattern regolari in sé, ma per mancanza di pattern regolari riconoscibili in modo automatico), si potrà procedere manualmente alla segmentazione del testo prima di elaborarlo con ChatGPT (Figg. 7-8), introducendo i simboli-marcatori \* e #



(processo naturalmente più lento, ma comunque vantaggioso e sostenibile ai fini del mark-up automatico in XML-TEI). È il caso di Alfani 1882, in cui non ci sono caratteristiche tali da poter distinguere e formalizzare automaticamente i vari elementi (proverbi, definizioni, citazioni, parole chiave delle sezioni tematiche):

#### VESTI, ADDOBBI.

Veste: abito che ci ricuopre o ci adorna.

Addobbo: è masserizia e arnese per lo più di uso e ornamento alle stanze.

Mangiare a modo suo, vestire a mo' degli altri.

E ciò è vero; né è punto necessario, per questo, piegarsi a tutte le volubilità della moda.

Chi non può far pompa, faccia foggia.

Se il vestito non è ricco, sia ben aggiustato alla persona, e pulito.

La scimmia è sempre scimmia, anco vestita di seta.

Il freno d'oro non fa il cavallo migliore.

Un giovinastro di rei costumi, ma tutto lindo e attillato, faceva pampa in una conversazione de' suoi abiti sfoggianti. «Perché ridete?» chiese a un tratto ad uno degli astanti. «Perché mi ricorre alla mente quel proverbio russo: «Sella d'oro non migliora cavallo.»

Guasta la figura di Dio, chi troppo si orna.

Mantello, cuopre il brutto e il bello.

Il bel vestire son tre N: nero, nuovo, netto.

Chi porta il fiore, sente d'amore, Chi porta il mazzo, sente di pazzo.

#### VINO.

VINO: qui è preso in un senso più lato, di quello che ha ordinariamente come Bevanda tratta dal frutto della vite:

Il grato umor che dalla vite cola (Dante).

Sta per tutte le bevande alcoliche e spiritose, laddove si discorre dell'ebrietà e d'ogni effetto pernicioso per l'abuso di esse.

La buona cantina fa il buon vino.

Buon vino fa buon sangue.

Il vino al sapore, il pane al colore.

È sottinteso: si conosce.

Vino amaro tienlo caro.

Il vino a' vecchi, e il latte a' bambini.

Il vino è la poppa dei vecchi.

Nell'uva son tre vinacciuoli: uno di sanità, uno di letizia, e uno di ubriachezza.

Chi del vino è amico, di se stesso è nemico.

Vino dentro, senno fuori.

Quando Bacco trionfa, il pensier fugge.

Dov'entra il bere, n'esce il sapere.

Il vino non ha timone.

Cioè l'ubriaco non ha guida.

L'acqua rovina i ponti, e il vino la testa.

Vino e sdegno fan palese ogni disegno.

Fig. 7: File .txt derivato dall'OCR e collazionato (Alfani 1882, pp. 184-185).

---

\* VESTI, ADDOBBI.

Veste: abito che ci ricuopre o ci adorna.

Addobbo: è masserizia e arnese per lo più di uso e ornamento alle stanze.

# Mangiare a modo suo, vestire a mo' degli altri.

E ciò è vero; né è punto necessario, per questo, piegarsi a tutte le volubilità della moda.

# Chi non può far pompa, faccia foggia.

Se il vestito non è ricco, sia ben aggiustato alla persona, e pulito.

# La scimmia è sempre scimmia, anco vestita di seta.

# Il freno d'oro non fa il cavallo migliore.

Un giovinastro di rei costumi, ma tutto lindo e attillato, faceva pampa in una conversazione de' suoi abiti sfoggianti. «Perché ridete?» chiese a un tratto ad uno degli astanti. «Perché mi ricorre alla mente quel proverbio russo: «Sella d'oro non migliora cavallo.»

# Guasta la figura di Dio, chi troppo si orna.

# Mantello, cuopre il brutto e il bello.

# Il bel vestire son tre N: nero, nuovo, netto.

# Chi porta il fiore, sente d'amore, Chi porta il mazzo, sente di pazzo.

\* VINO.

VINO: qui è preso in un senso più lato, di quello che ha ordinariamente come Bevanda tratta dal frutto della vite:

Il grato umor che dalla vite cola (Dante).

Sta per tutte le bevande alcoliche e spiritose, laddove si discorre dell'ebrietà e d'ogni effetto pernicioso per l'abuso di esse.

# La buona cantina fa il buon vino.

# Buon vino fa buon sangue.

# Il vino al sapore, il pane al colore.

È sottinteso: si conosce.

# Vino amaro tienlo caro.

# Il vino a' vecchi, e il latte a' bambini.

# Il vino è la poppa dei vecchi.

# Nell'uva son tre vinaccioli: uno di sanità, uno di letizia, e uno di ubriachezza.

# Chi del vino è amico, di se stesso è nemico.

# Vino dentro, senno fuori.

# Quando Bacco trionfa, il pensier fugge.

# Dov'entra il bere, n'esce il sapere.

# Il vino non ha timone.

Cioè l'ubriaco non ha guida.

# L'acqua rovina i ponti, e il vino la testa.

# Vino e sdegno fan palese ogni disegno.

Fig. 8: File .txt trasformato in testo *pre-markup* in modo manuale (Alfani 1882, pp. 184-185).

Oltre alle modalità *a.* (automatica) e *b.* (manuale), può essere adottata una terza opzione, semiautomatica, qualora nel testo siano individuabili, anche solo parzialmente, pattern ricorrenti: in questi casi è possibile avviare una marcatura preliminare automatica, da revisionare successivamente con interventi manuali (ad esempio per rimuovere marcatori in eccesso).

Simili operazioni di pre-markup consentono di trasformare un testo continuo (non ancora segmentato) in una serie di record autonomi, processabili singolarmente: ciò facilita, nelle fasi successive, l'intervento dell'intelligenza artificiale, che può riconoscere, estrarre e modificare singoli elementi o porzioni testuali in modo molto più preciso e flessibile rispetto a un semplice confronto "tradizionale". Quest'ultimo, infatti, basato su corrispondenze esatte tra stringhe di testo, è in grado di identificare un segmento solo se questo coincide esattamente con la sequenza specificata – ciò implicherebbe la necessità di elencare manualmente tutte le varianti possibili (operazione antieconomica e poco efficiente), che può appesantire e compromettere l'efficacia dell'elaborazione automatica da parte di ChatGPT. Si consideri, per esempio,



un’istruzione come la seguente: “considera come entrate (quindi marca con <entry>) tutte le righe che iniziano con un numero seguito da un punto, a cui segue una parola con iniziale maiuscola, escludendo le righe che iniziano con numeri fra parentesi o quelle prive di numerazione...”. Indicazioni di questo tipo, se fornite direttamente su un testo continuo, non solo rischiano di essere incomplete (cioè di non contemplare tutte le combinazioni effettivamente presenti nel testo), ma tendono anche a confondere il modello: ChatGPT può facilmente sbagliare la segmentazione, interrompersi prima del previsto o accorpare elementi che dovrebbero restare distinti (come nel caso di abbreviazioni, forme seguite da punteggiatura e simili). Questo accade in assenza di un pre-markup che isoli in modo netto e strutturato le singole unità: sebbene ChatGPT sia in grado di gestire variazioni e riconoscere schemi complessi, ha comunque bisogno che il testo sia suddiviso logicamente in blocchi coerenti<sup>21</sup>.

Grazie alla formalizzazione del testo, la conversione automatica in XML-TEI da parte dell’IA non avviene – almeno nei primi step – direttamente sul contenuto del testo, ma sulla base della sua organizzazione strutturale; questo rende lo schema operativo più agile e facilmente adattabile, offrendo un ulteriore vantaggio: nella prospettiva di applicarlo ad altre risorse, una volta individuati i pattern fondamentali, lo stesso metodo può essere riutilizzato e personalizzato in funzione della struttura dei repertori lessicali/lessicografici da informatizzare.

## 2.3 Il mark-up automatico con ChatGPT

Una volta completata la preparazione del testo pre-markup – ottenuta in modo automatico o manuale – si potrà avviare la seconda fase, fornendo a ChatGPT il file in input: grazie a una serie di istruzioni mirate (modulabili secondo la tipologia e la strutturale testuale), l’IA sarà in grado di riconoscere in maniera apprezzabile i diversi campi da marcare automaticamente in XML-TEI.

a. Nel caso del testo di Rampoldi, per istruire correttamente ChatGPT, occorre specificare in prima istanza alcune caratteristiche:

- è una raccolta paremiologica, composta da due tipologie di unità autonome: proverbi e note;
- ogni riga che inizia con asterisco (\*) introduce un nuovo record, da marcare come <entry>, ovvero come un’unità autonoma contenente un proverbio, che deve essere racchiuso all’interno dei tag <form><orth>*proverbio*</orth></form>;
- i proverbi sono numerati, quindi le <entry> devono riportare una numerazione progressiva;

- ogni riga che inizia con # corrisponde a un nuovo record, da marcare come `<note xml:id="n1" type="footnote">testo della nota</note>`: si tratta di un'unità autonoma contenente il testo di una nota a piè di pagina;
- ogni `<entry>` e ogni `<note>` si chiudono prima di un record successivo.

A questo punto, ChatGPT elabora e riepiloga le istruzioni fornite (Fig. 9) e genera il file .xml (Fig. 10):

#### Struttura del file .txt :

- \* indica l'inizio di una nuova `<entry>` .
- Il contenuto della riga è un proverbio, da marcare come:

```
xml Copia Modifica  
  
<form><orth>PROVERBIO</orth></form>
```

- L' `<entry>` si chiude prima di un nuovo \* o # .
- # indica l'inizio di una **nota**, da marcare come:

```
xml Copia Modifica  
  
<note xml:id="n1" type="footnote">TESTO</note>
```

- Le note si chiudono prima di un nuovo \* o #
- L' `xml:id` va numerato progressivamente ( `n1` , `n2` , `n3` , ... )

Fig. 9: Schermata tratta da ChatGPT con il riepilogo delle istruzioni necessarie per marcare il file pre-markup.

```

<entry xml:id="e248">
  <form><orth>248. Non v'è più facil cosa che ingannar sé stesso.</orth></form>
</entry>

<entry xml:id="e249">
  <form><orth>249. Col fuoco si prova l'oro, coll'oro la donna, colla donna l'uomo.</orth></form>
</entry>

<entry xml:id="e250">
  <form><orth>250. L'uom vantasi padron di tutto, ma sovente della donna è schiavo (1).</orth></form>
</entry>

<entry xml:id="e251">
  <form><orth>251. Dominano sempre le femine quando gli uomini sono voluttuosi (2).</orth></form>
</entry>

<note xml:id="n40" type="footnote">(1) Nos imperamus omnibus, uxores nobis. — Plutarc. in Rom. Apophth.</note>

<note xml:id="n41" type="footnote">(2) Son padrone del mondo oggi le donne, E cedon toghe e spade a cuffie e gonne. Cent. Piacent. VII, 76. La donna ha in oggi la maschil fierezza, E l'uomo della donna ha la mollezza. id. IX, 92.</note>

<entry xml:id="e252">
  <form><orth>252. Non comanda mai la donna per sua fortezza, ma bensì per debolezza dell'uomo.</orth></form>
</entry>

```

Fig. 10: Schermata tratta da Oxygen XML Editor con il risultato del mark-up automatico dopo la prima sessione.

Dopo aver controllato la conversione del file dal formato .txt a quello .xml prodotta automaticamente da ChatGPT, è quasi sempre necessario eseguire ulteriori operazioni per correggere e integrare il testo XML-TEI. Per il file Rampoldi, ad esempio, è indispensabile un'istruzione aggiuntiva per marcare gli agganci alle note (che durante la prima sessione non erano stati inclusi), affinché ChatGPT possa aggiornare le istruzioni (Fig. 11) e, sulla base di queste ultime, generare un nuovo file .xml (Fig. 12): i numeri fra parentesi tonde contenuti in <orth> devono essere marcati con <ref target="#nX"> ed essere collegati alla prima nota successiva corrispondente (e mai a una precedente), secondo una logica ben precisa: se in <orth> c'è (1), questo dovrà essere collegato al testo della nota successiva che inizia con (1), quindi essere collegato al numero fra parentesi (n) che precede il testo della nota (e non all'xml:id di <note>).

## 🎯 Obiettivo:

Marcare ogni numero tra parentesi tonde nel contenuto di `<orth>` con un tag `<ref>` e collegarlo alla prima `<note>` successiva che contiene lo stesso numero.

## ✅ Regole operative:

1. Solo nei tag `<orth>` :  
Cerca (1), (2), ecc. solo dentro `<orth>`.
2. Identifica la prima `<note>` successiva che contiene esattamente lo stesso numero tra parentesi (n) nel testo (non nell'xml:id!).
3. Trasforma (n) in:

```
xml
<ref target="#nX">(n)</ref>
```

Copia Modifica

Dove nX è l' xml:id della nota trovata (es. #n23 ).

Fig. 11: Schermata tratta da ChatGPT con il riepilogo delle istruzioni aggiuntive di mark-up.

```
<entry xml:id="e248">
  <form><orth>248. Non v'è più facil cosa che ingannar sé stesso.</orth></form>
</entry>

<entry xml:id="e249">
  <form><orth>249. Col fuoco si prova l'oro, coll'oro la donna, colla donna l'uomo.</orth></form>
</entry>

<entry xml:id="e250">
  <form><orth>250. L'uom vantasi padron di tutto, ma sovente della donna è schiavo <ref target="#n40">(1)</ref>.</orth></form>
</entry>

<entry xml:id="e251">
  <form><orth>251. Dominano sempre le femine quando gli uomini sono voluttuosi <ref target="#n41">(2)</ref>.</orth></form>
</entry>

<note xml:id="n40" type="footnote">(1) Nos imperamus omnibus, uxores nobis. — Plutarc. in Rom. Apopht.</note>

<note xml:id="n41" type="footnote">(2) Son padrone del mondo oggi le donne, E cedon toghe e spade a cuffie e gonne. Cent. Piacent. VII, 76. La donna ha in oggi la maschil fierezza, E l'uomo della donna ha la mollezza. id. IX, 92.</note>

<entry xml:id="e252">
  <form><orth>252. Non comanda mai la donna per sua fortezza, ma bensì per debolezza dell'uomo.</orth></form>
</entry>
```

Fig. 12: Schermata tratta da Oxygen XML Editor con il risultato del mark-up automatico dopo la seconda sessione.

Bisogna considerare che fra i primi input e il risultato finale ci sono diversi passaggi “di rodaggio”, cioè piccole sessioni provvisorie e transitorie (che qui non riproduciamo), necessari per controllare, correggere e istruire ChatGPT, affinché

svolga tutto correttamente. Come emerge dal test sulla raccolta di Rampoldi, è opportuno (specialmente per strutture complesse) suddividere l'attività di mark-up automatico in più fasi, guidando ChatGPT attraverso sessioni progressive per evitare errori di annotazione dovuti a elaborazioni simultanee (Fig. 13). Un esempio è la gestione dei riferimenti univoci: in questo caso, infatti, è opportuno far marcare in prima battuta gli `@xml:id` e i relativi valori (assegnati agli elementi `<entry>` e `<note>`) e, solo in un secondo momento, procedere con il mark-up dei riferimenti incrociati, quindi dell'elemento `<ref>` e dell'attributo `@target` per richiamare le singole entità (Fig. 12).

### Ecco il piano operativo Rampoldi:

#### STEP 1 — Parsing del file txt

- Leggo riga per riga
- Identifico `*` (proverbi) e `#` (note)

#### STEP 2 — Generazione `<entry>` con numerazione corretta

#### STEP 3 — Generazione `<note>` con `xml:id="nX"`

#### STEP 4 — Inserimento dei `<ref>` (solo dentro `<orth>`) puntati alla prima `<note>` successiva con lo stesso (n)

#### STEP 5 — Pulizia finale:

- Righe bianche corrette
- Numerazione coerente
- Nessun ref duplicato

Fig. 13: Schermata tratta da ChatGPT con il riepilogo finale delle istruzioni di mark-up.

**b.** Per eseguire il mark-up automatico con ChatGPT sulla raccolta paremiologica di Alfani occorre partire da una serie di istruzioni:

- è una raccolta di proverbi;
- ogni sezione comincia quando a inizio riga c'è un asterisco (\*) e finisce prima di un nuovo \*; le sezioni devono essere marcate con `<div type="section">sezione</div>`;
- ogni sezione comprende blocchi di proverbi: ogni proverbio comincia quando a inizio riga c'è un cancelletto (#) e finisce prima di un nuovo #;
- i proverbi devono essere racchiusi dentro `<entry><form><orth>proverbio</orth></form></entry>`;
- all'interno di `<orth>` deve essere racchiuso il proverbio vero e proprio, cioè le parole che vanno da # fino al primo punto fermo.

ChatGPT avrà allora le informazioni necessarie per applicare il mark-up automatico in XML-TEI (Fig. 14) e generare, dopo la prima sessione, un file .xml (Fig. 15) – che, come si può vedere, risulta incompleto: sono stati esclusi dal mark-up (quindi non sono racchiusi in alcun tag) sia le parole chiave delle sezioni tematiche sia l'eventuale testo dopo i proverbi racchiusi in <orth>; all'elemento <entry> non è stato assegnato alcun @xml:id. Di conseguenza, bisognerà fornire ulteriori comandi all'IA.

### Struttura del file .txt

#### ◆ Sezioni ( <div type="section"> )

- Ogni sezione inizia con \* a inizio riga
- Termina prima del successivo \*
- Viene convertita in:

```
xml Copia Modifica  
  
<div type="section"> ... </div>
```

#### ◆ Proverbi ( <entry> )

- Ogni proverbio inizia con # a inizio riga
- Termina prima del prossimo #
- Ogni proverbio viene racchiuso in:

```
xml Copia Modifica  
  
<entry>  
  <form>  
    <orth>TESTO DEL PROVERBIO</orth>  
  </form>  
</entry>
```

- Solo la parte dal # fino al primo punto ( . ) va in <orth>

Fig. 14: Schermata tratta da ChatGPT con il riepilogo delle istruzioni iniziali per il mark-up automatico.



```

<div type="section">
  <head>ABITI (Abitudini), USANZE.</head>
  Abito (o come si dice comunemente, Abitudine) in generale è: Abilità acquistata per uso frequente di operazioni simili. Più ristrettamente poi, come qui, è: Consuetudine diventata quasi natura, la quale spesso è difficile a. vincersi.
  Usanza è maniera di vivere e di procedere in alcune cose, accettata e frequentata dai più.
  L'uomo è tratto da due cose perfettamente contrarie, dall'abito e dalla novità (La Bruyère).

  <entry>
    <form>
      <orth>A usanza nuova non correre.</orth>
    </form>
  </entry>
  Prudenza conservativa che risiede massimamente nel popolo, quando egli segue suo proprio istinto e sua ragione (G.).

  <entry>
    <form>
      <orth>Ciò che s'usa non fa scusa.</orth>
    </form>
  </entry>
  Non tutte le cose sono scusabili per dire: così s'usa, e così fanno gli altri (Serdonati).

  <entry>
    <form>
      <orth>Consuetudine è una seconda natura.</orth>
    </form>
  </entry>

```

Fig. 15: Schermata tratta da Oxygen XML Editor con il risultato del mark-up automatico dopo la prima sessione.

Servirà allora un ulteriore step per migliorare e completare il mark-up, integrando le seguenti istruzioni (Fig. 16):

- l'elemento <entry> deve contenere tutto il testo compreso tra il simbolo # e il successivo #;
- il testo compreso tra \* e # deve essere racchiuso all'interno di <p>;
- all'interno di <orth>, il contenuto deve estendersi a partire dal simbolo # fino al primo punto fermo, a eccezione delle abbreviazioni o S. (presenti nei proverbi che citano nomi di santi), nel qual caso il segmento si estenderà fino al secondo punto fermo;
- all'interno di <entry> tutto il testo successivo a <orth> (quindi definizioni, citazioni e commenti) deve essere racchiuso dentro il tag <sense>.

## 1. Sezioni

- Ogni riga che inizia con **\*** apre una nuova sezione:

xml

Copia

Modifica

```
<div type="section">
```

- Il blocco di testo subito dopo l'asterisco **\*** (spiegazioni introduttive) viene incluso in:

xml

Copia

Modifica

```
<p>Testo introduttivo...</p>
```

- La sezione si chiude **prima del prossimo \***

## 2. Proverbi (entry)

- Ogni riga che inizia con **#** è un nuovo proverbio, marcato con:

xml

Copia

Modifica

```
<entry>
  <form><orth>TESTO DEL PROVERBIO FINO AL PRIMO PUNTO FERMO</orth></form>
  <sense>TESTO RIMANENTE (se c'è)</sense>
</entry>
```

### Regole per **<orth>**:

- Si prende il **testo dalla #** fino al **primo punto fermo (.)**
- Se il punto fermo fa parte di **abbreviazione tipo s. o S.**, si ignora quel punto e si cerca il **prossimo vero punto** che chiude la frase.
- Il resto del testo (dopo il punto fermo valido) va dentro **<sense>**

Fig. 16: Schermata tratta da ChatGPT con il riepilogo delle istruzioni aggiornate per il mark-up automatico.

In vista della progettazione di ricerche avanzate o guidate (per il sito che ospiterà la raccolta in versione interrogabile), sarà utile introdurre un'ulteriore istruzione: associare a ciascun proverbio una serie di parole chiave (estraendone da una a un massimo di quattro), da inserire come valori dell'attributo @ana (che viene assegnato a ogni elemento <entry>). Sulla base di queste nuove indicazioni – tra cui il criterio di escludere parole semanticamente vuote nella selezione delle *keywords* – ChatGPT aggiorna le regole per il mark-up automatico (Fig. 17):

### ✓ Regole per l'assegnazione delle parole chiave (@ana) alle <entry>

#### 1. Analizzare il contenuto di ogni <orth>

→ estrarre **da 1 a 4 parole chiave** significative.

#### 2. Assegnare all'elemento <entry>

→ l'attributo @ana con i valori corrispondenti alle **keywords estratte**.

#### 3. Criteri per la scelta delle parole chiave:

- privilegiare **sostantivi**;
- **escludere parole semanticamente vuote**, in particolare:
  - articoli (il, la, un, ecc.),
  - preposizioni (di, a, con, su, ecc.),
  - congiunzioni (e, ma, o, ecc.),
  - pronomi (che, cui, essi, lui, ecc.),
  - avverbi generici (già, poi, ancora, qui, ecc.).

Fig. 17: Schermata tratta da ChatGPT con il riepilogo delle istruzioni aggiornate per il mark-up delle parole chiave.

ChatGPT sarà in grado di produrre un nuovo file .xml contenente le parole chiave:

```
<div type="section">
  <head>POCHEZZA D'ANIMO.</head>
<p>Pochezza d'animo è contrario a fortezza e coraggio, segnatamente morale; quando, cioè, l'uomo a ogni lieve ostacolo si smarrisce e
avvilisce; ad ogni difficoltà o intimidimento che altri gli faccia, si turba e cede.</p>

<entry xml:id="pr1307" ana="piedi collo">
  <form>
    <orth>Chi non vuol piedi sul collo, non s'inchini.</orth>
  </form>
  <sense>Non sia vile d' animo, senta la propria dignità d'uomo, chi vuole che altri lo rispetti e non gli faccia soverchierie.</sense>
</entry>

<entry xml:id="pr1308" ana="pecora lupo">
  <form>
    <orth>Chi pecora si fa, lupo la mangia.</orth>
  </form>
</entry>

<entry xml:id="pr1309" ana="stimare">
  <form>
    <orth>Chi vuol essere stimato, stimi se stesso.</orth>
  </form>
</entry>

<entry xml:id="pr1310" ana="uomo stimare">
  <form>
    <orth>Tanto vale l'uomo, quanto si stima.</orth>
  </form>
</entry>
```

Fig. 18: Schermata tratta da Oxygen xml Editor con il risultato del mark-up automatico dopo la terza sessione.

## 2.4 I siti pilota

Per osservare ancora più concretamente i risultati dei primi test condotti sulle due raccolte paremiologiche – quella di Rampoldi e quella di Alfani – sono stati allestiti dei siti pilota<sup>22</sup>, concepiti come prototipi leggeri per testare visivamente l'intero output marcato, ovvero il file .xml prodotto in modo semiautomatico con ChatGPT, e verificarne la struttura e gli elementi annotati (nella fattispecie sezioni, proverbi, commenti/spiegazioni, note, parole chiave). Sebbene non integrino un motore di ricerca vero e proprio basato su back-end (ossia gestito da un server che interroga un database), questi prototipi offrono funzioni di ricerca e filtri lato client, eseguite direttamente nel browser tramite JavaScript: si tratta di una soluzione tecnicamente più semplice, ma comunque sufficiente per un'esplorazione interattiva e una prima valutazione della marcatura.

Di seguito, la schermata principale che ospita la raccolta di Rampoldi (Fig. 19) e quella relativa alla raccolta di Alfani (Fig. 20) – si noti la possibilità di filtrare “tutte le sezioni” o di selezionarne soltanto una dal menù a tendina (opzione di ricerca permessa grazie alla presenza dei <div> tematici, marcati nel file .xml all'interno del

<body> di ciascun repertorio):

***I proverbi e le sentenze proverbiali.  
Raccolta di Giovanni Rampoldi (1852)***

Tutte le sezioni

Cerca nei proverbi...

4973 proverbi trovati

✓ Tutte le sezioni  
Sentenze proverbiali – Parte prima  
Sentenze proverbiali – Parte seconda  
Precetti proverbiali – Parte terza  
Dialoghi proverbiali – Parte quarta

*Sentenze proverbiali – Parte prima*

1. L'uomo nasce, vive e muore in un solo istante <sup>(1)</sup>.

*Nota: (1) È un sogno passaggér la nostra vita, E allor che ci svegliamo è già finita. Cent, Piacent. I, 29.*

*Sentenze proverbiali – Parte prima*

2. Soltanto per la società è fatto l'uomo.

*Sentenze proverbiali – Parte prima*

3. Previdenza e speranza fan l'uomo superiore agli altri animali.

*Sentenze proverbiali – Parte prima*

4. L'uomo è il re degli animali <sup>(2)</sup>.

*Nota: (2) Meschina però è tale dignità per l'uomo; le torbide cure lo infestano, i dolori lo straziano, le passioni lo agitano e lo tormentano, i flagelli e le calamità lo divorano, ed il baratro della morte si spalanca ad ogni stante sott'i suoi passi e minaccia d'ingojarlo .... l'homme, Ce rot des animaux, combien a-t'il de rois, L'ambition, l'amour, l'avarice, ou la haine, Tiennet comme un forcat son esprit à la chaine, Boileau, Sat. VIII.*

Fig. 19: Schermata principale del sito pilota con la raccolta paremiologica di Rampoldi.

## Proverbi e modi proverbiali scelti ed annotati da Augusto Alfani (1882)

Esplora i proverbi tramite parole chiave, sezioni o ricerca full-text

[Torna alla pagina principale](#)

Cerca per testo libero...

Filtra per sezione: Tutte le sezioni

**A usanza nuova non correre.**  
Prudenza conservativa che risiede massimamente nel popolo, quando egli segue  
Parole chiave: usanza correre

**Ciò che s'usa non fa scusa.**  
Non tutte le cose sono scusabili per dire: così s'usa, e così fanno gli altri (Serdona  
Parole chiave: usare scusa

**Consuetudine è una seconda natura.**  
Quando abbiām fatto tanto di prendere un abito, segnatamente cattivo, è difficile a  
Parole chiave: consuetudine natura

**È difficile condurre il can vecchio a mano.**  
Far perder cioè un abito inveterato.  
Parole chiave: condurre cane

✓ Tutte le sezioni

ABITI (Abitudini), USANZE.  
ADULAZIONE, LODI, LUSINGHE.  
AFFETTI, AMORE, PASSIONI, VOGLIE.  
AGRICOLTURA, METEOROLOGIA, STAGIONI ECONOMIA RURALE.  
ALLEGRIA, DARSİ BEL TEMPO.  
AMBIZIONE, SIGNORIA, CORTE.  
AMICIZIA.  
ASTUZIA, FRODE, INGANNO, RAPINA.  
AVARIZIA.  
BELLEZZA E SUO CONTRARIO, FATTEZZE DEL CORPO.  
BENEFICENZA, SOCCORRERSI.  
BENIGNITÀ, PERDONO.  
BISOGNO, NECESSITÀ.  
BUONA E MALA FAMA.  
BUONI E MALVAGI.  
CASA.  
COMPAGNIA BUONA E CATTIVA.  
CONDIZIONI E SORTI DISUGUALI.  
CONFORTI NE' MALI.  
CONSIGLIO, RIPRENSIONE, ESEMPIO.  
CONTENTARSI DELLA PROPRIA SORTE.  
CONTRATTAZIONI, MERCATURA.  
COSCIENZA, GASTIGO DEI FALLI.  
COSTANZA, FERMEZZA.

Fig. 20: Schermata principale del sito pilota con la raccolta paremiologica di Alfani.



Infine, per quanto riguarda la funzionalità delle parole chiave, questa si rivela particolarmente utile nelle ricerche sui repertori paremiologici: cliccando sui pulsanti associati alle *keywords* che sono state marcate, si potranno visualizzare tutti i proverbi della raccolta marcati per la parola selezionata. Ad esempio, selezionando “lode” oppure “adulatore”, si otterrà l’elenco completo di tutti i proverbi in cui tali parole chiave compaiono:

The screenshot displays two columns of search results from the DILEF interface. Each result consists of a proverb, its explanation, and a list of associated keywords with interactive buttons.

Proverbio	Spiegazione	Parole chiave
Chi ci loda si dee fuggire, e chi c'ingiuria si dee soffrire.	Perchè spesso le lodi han per cagione il tornaconto di chi le profonde. Il sopportare, poi, le ingiurie e perdonare l'offensore, è precetto sublime di religione.	lode, ingiuria
La lode giova al savio o nuoce al matto.	Perchè il matto ci si affida alla cieca, e il savio si mette in guardia o si dà sempre più cura di meritarsela.	lode, savio, matto
La vera lode adorna, la non vera riprende.	Almeno le anime gentili, e quelle che sentono giustamente la propria dignità (G.).	vera, lode
Gola degli adulatori, sepolcro aperto.		gola, adulatore, sepolcro
La vita dell'adulatore poco tempo sta in fiore.		vita, adulatore, fiore
Adulatori e parassiti son come i pidocchi.	Campano sulla pelle altrui.	adulatore, parassita, pidocchio

Fig. 21: Risultato della ricerca delle parole chiave “lode” e “adulatore” nel sito con la raccolta paremiologica di Alfani.

### 3. Conclusioni: verso una nuova velocità di crociera per i progetti di lessicografia elettronica?

Se finora la maggior parte dei progetti di lessicografia elettronica – già realizzati o tuttora in corso – ha comportato un notevole dispendio di tempo e di energie (soprattutto a causa di procedure prevalentemente manuali), l’impiego di strumenti basati sull’intelligenza artificiale – come ChatGPT (ma anche Claude, Gemini e altri modelli linguistici avanzati) – offre la possibilità di razionalizzare l’intero flusso di lavoro, automatizzando in larga parte il processo di marcatura. L’uso di questi modelli linguistici per il mark-up automatico in XML-TEI, basati su istruzioni ben definite e modellabili secondo la tipologia di testo da elaborare, consente un risparmio significativo: in termini quantitativi si può stimare una riduzione del lavoro tra il 75% e il 90% (variabile a seconda della conformità del testo sorgente, della complessità delle singole voci e del livello di granularità imposto dal progetto). Questa stima deriva da un confronto tra i tempi medi richiesti dalla marcatura tradizionale (calcolati sulla base di dati empirici ricavati da progetti già realizzati) e quelli ottenibili con ChatGPT mediante generazione automatica in batch (cioè insieme di voci elaborate in blocco); trattandosi di una metodologia ancora in fase sperimentale, le stime relative all’automatizzazione poggiano, da un lato, sui primi test effettuati

sulle due raccolte di proverbi (Rampoldi 1852 e Alfani 1882) e, dall'altro, su proiezioni teoriche riferite ad altre tipologie lessicografiche. Partiamo da queste ultime.

Se consideriamo un dizionario di circa 5.000 voci, la stima minima del tempo necessario per la marcatura manuale si aggira intorno alle 375 ore complessive (calcolando una media di 2-3 minuti per una voce semplice, a cui va aggiunto circa 1 minuto per le operazioni di revisione e validazione). La stima massima, invece, può arrivare fino a 540 ore (equivalenti a circa 3 mesi di lavoro continuativo), laddove il dizionario presenti un'elevata complessità strutturale: voci più articolate (comprendenti, ad esempio, lemma, categoria grammaticale, definizioni, esempi, etimologie, citazioni, indicatori d'uso, sottolemmi, note, commenti, ecc.) possono richiedere fino a 4-5 minuti ciascuna – se si tiene conto anche di costanti tecniche (come la verifica formale della sintassi XML-TEI) o di variabili a livello di mediostruttura (es. rinvii e riferimenti interni) e di macrostruttura (es. gruppi di voci semanticamente correlate). ChatGPT, invece, opportunamente guidato e istruito (con prompt mirati, calibrati e personalizzati in base alla risorsa), è in grado di restituire automaticamente l'intero dizionario in formato .xml in un arco che va da pochi minuti a 1-1,5 ore; a questa fase automatica va aggiunta una revisione da parte del linguista, che richiede in media 0,5-1,5 minuti per voce: il tempo totale stimato per la procedura semiautomatica si colloca così tra le 43 e le 127 ore complessive (comprendente sia della generazione del file .xml con l'IA sia del controllo umano).

I dati rappresentati nel grafico (Fig. 22) mostrano il divario tra le due modalità operative, quella manuale e quella semiautomatica – si noti come la marcatura generata da ChatGPT, anche includendo la fase di revisione del mark-up (il cui peso varierà in funzione delle correzioni e delle eventuali integrazioni), consente un risparmio di ore notevole, rendendo possibile la gestione di progetti di ampia portata, in tempi finora considerati proibitivi.

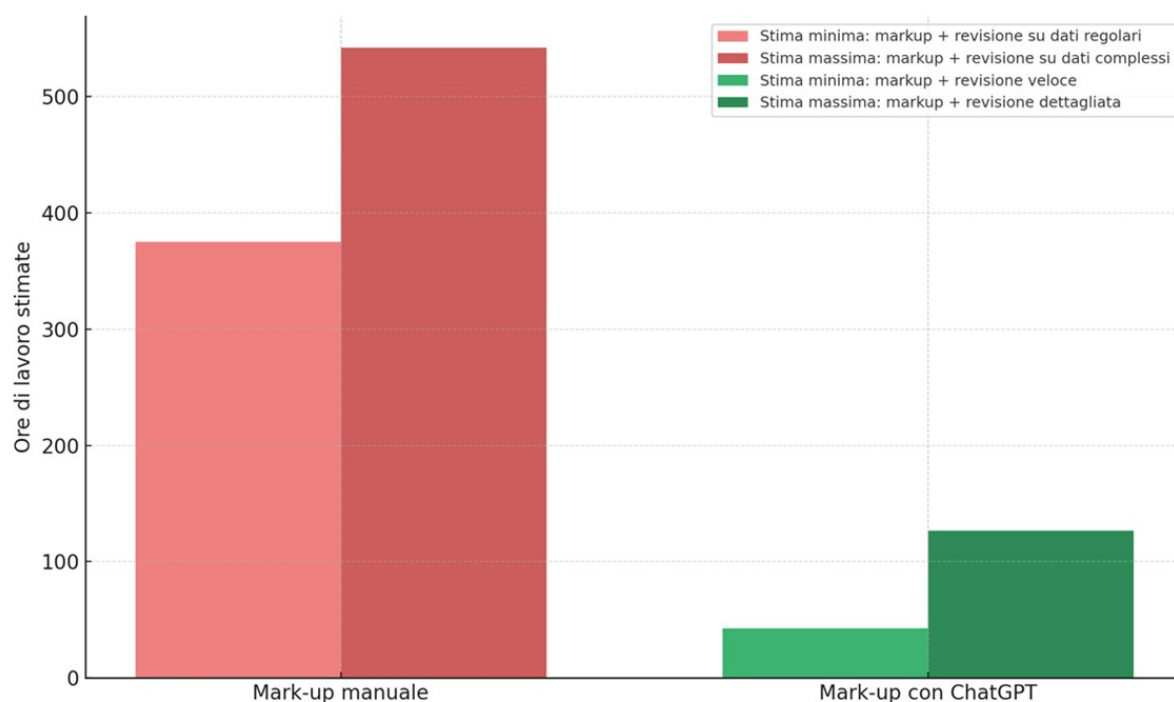


Fig. 22: Tempi di mark-up (manuale vs semiautomatico) su un dizionario di 5.000 voci.

Nel caso delle raccolte paremiologiche, caratterizzate da una struttura testuale mediamente non troppo articolata rispetto a quella dei dizionari, il risparmio temporale risulta ancora più vistoso. In scenari come questi, l'intelligenza artificiale non solo assiste, ma arriva a sostituire quasi totalmente l'operazione di marcatura: non contando il tempo di revisione, il risparmio rispetto al mark-up manuale è quasi del 100%; se invece si tiene conto delle ore di revisione, la percentuale si aggira attorno all'80% (restando comunque molto alta). Provando, dunque, a considerare solo l'automatizzazione pura (senza aggiungere il tempo di revisione, che richiederà meno di 50 ore) su una raccolta di 5.000 proverbi, il tempo del mark-up automatico sarà inferiore a 5 minuti, a fronte di circa 250 ore di marcatura manuale:

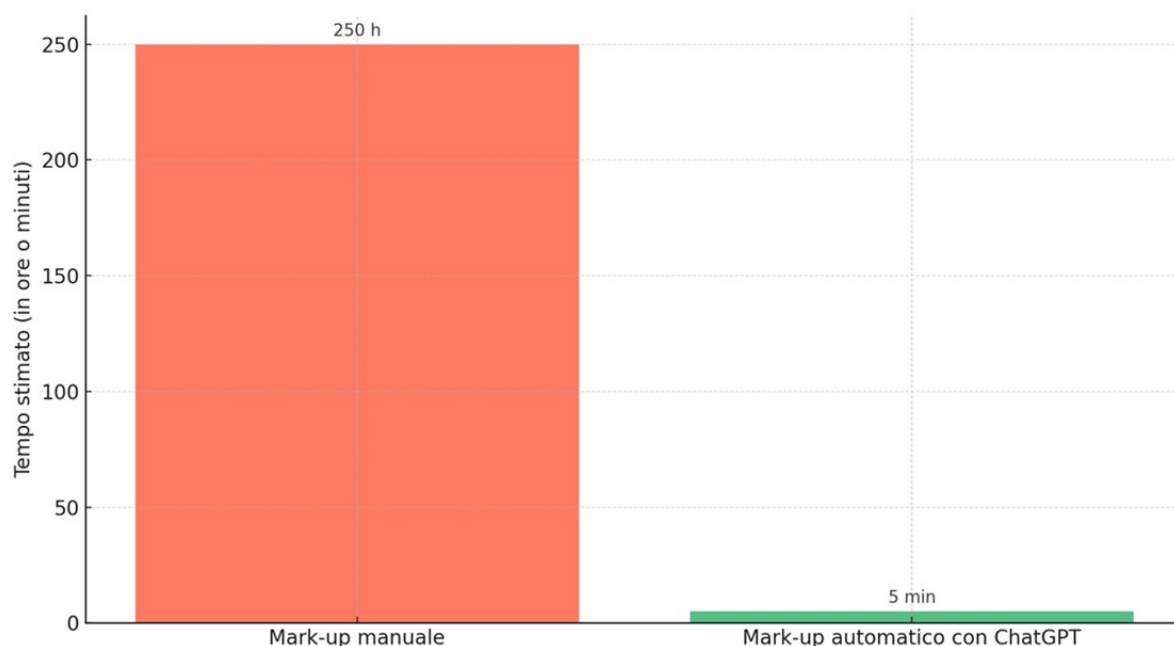


Fig. 23: Tempi di mark-up (manuale vs automatico) su una raccolta paremiologica di 5.000 proverbi.

In sintesi, rispetto a una costruzione manuale dell'intera struttura XML, con il supporto di ChatGPT il ruolo del linguista si trasforma e si sposta su un piano più “gestionale”, di supervisione, concentrandosi in quattro passaggi principali: preparazione del file .txt da elaborare tramite IA; somministrazione delle istruzioni di marcatura; guida e verifica del flusso semiautomatico (che porta dal file sorgente al file marcato); revisione e arricchimento del mark-up prodotto con ChatGPT.

Come si è tentato di dimostrare, l'avvento dell'intelligenza artificiale ha buone probabilità di segnare una nuova era per la lessicografia elettronica. In particolare, ChatGPT potrebbe rappresentare una chiave di volta – e di svolta – nella creazione su larga scala di versioni elettroniche di repertori lessicali o lessicografici (ma anche di altri strumenti affini), grazie ai seguenti vantaggi:

- ottimizzazione dei tempi e delle risorse per la marcatura in XML-TEI;
- riduzione significativa degli interventi manuali a favore di procedure semiautomatiche;
- accelerazione del processo di informatizzazione;
- possibilità di operare su un numero più ampio di testi;
- maggiore efficienza operativa e produzione di risultati di livello medio-alto;
- flessibilità nella configurazione dei flussi di lavoro, personalizzabili in base alla complessità e alla struttura della singola risorsa;
- buon rapporto costo-efficacia nella realizzazione di progetti di lessicografia elettronica.

In un recente contributo di Salgado *et al.* (2024), a proposito dell'informatizzazione di dizionari di seconda generazione – e, in particolare, a proposito del progetto MORDigital, volto alla creazione di una versione elettronica delle prime tre edizioni del *Diccionario da Lingua Portuguesa* (1789) di António de Morais Silva (noto anche come *Morais Dictionary*) –, si sottolinea l'importanza delle *best practies*, definite come «well-researched procedures, backed by evidence and experience, designed for broad adoption to ensure optimal working methods»<sup>23</sup>. Quelle proposte in questa ricerca potrebbero allora configurarsi come potenziali 'buone pratiche' che, seppur ancora in fase sperimentale, appaiono già orientate verso la definizione di linee guida e strategie operative di riferimento, avallate da metodi facilmente replicabili, adattabili a diverse tipologie di repertori lessicografici/lessicali e fondate su una sinergia ben bilanciata tra intervento umano – prevalentemente “gestionale” – e supporto dell'intelligenza artificiale per il mark-up automatico in XML-TEI.

Nell'introduzione, per sintetizzare le potenzialità dell'informatizzazione di risorse lessicografiche e lessicali, si è fatto riferimento a tre parole chiave: *valorizzazione*, *estendibilità* e *accessibilità*. Alla luce dei risultati emersi in questo studio – considerando l'intelligenza artificiale non solo come un potente acceleratore per l'attività di marcatura, ma anche come supporto facilmente accessibile e gestibile da parte dell'umanista digitale (senza richiedere competenze informatiche avanzate, se non la conoscenza del linguaggio XML-TEI) – si possono aggiungere, a buon diritto, altre due parole chiave: *velocità* e *sostenibilità*.

---

## Note

1. Si parla di un'era «marked by the emergence of mass digitization projects, revolutionizing the field of lexicography. These initiatives aim to preserve heritage materials and simplify access to historical documents, emphasizing the importance of best practices in ensuring data

quality and integrity. Technological progress has taken us beyond mere scanned texts [...] to transform printed dictionaries into computational, readily exploitable lexical resources, enhancing searchability. [...] This shift underscores the increasing significance of using advanced technologies and standards to enhance the accessibility of digitized dictionaries for research and academic purposes» (Salgado *et al.* 2024, p. 126). Questa nostra ricerca, infatti, rappresenta un piccolo tassello «of a broader landscape, where numerous [...] dictionaries are undergoing digitization processes and becoming accessible online» (*ibidem*).

2. In un recente contributo Marco Biffi, gettando luce su un problema che da anni serpeggia in seno alla lessicografia elettronica italiana, vale a dire quello di una terminologia instabile e non ancora condivisa fra gli esperti del settore – linguisti e informatici –, ha avanzato una serie di proposte, tentando di risolvere una lotta ormai ventennale fra le polirematiche *dizionario elettronico* e *dizionario digitale* (a favore della prima) e di fissare dei puntelli terminologici per i *dizionari elettronici* (di *prima generazione* e di *seconda generazione*; di *primo livello* e di *secondo livello*). Per un approfondimento sulla questione si rimanda a Biffi 2024.

3. <https://www.gdli.it>.

4. <https://www.tommaseobellini.it/>.

5. <http://new.lessicografia.it>.

6. La banca dati è in fase di allestimento (cfr. Patella 2023a, Ead. 2023b e 2024a).

7. <https://baldinucci.accademiadellacrusca.org/testo-del-vocabolario> (cfr. Patella 2024b).

8. Tra le possibili soluzioni di mark-up in XML, le linee guida della Text Encoding Initiative (TEI) risultano particolarmente adatte a soddisfare l'esigenza di una codifica standard, ma al tempo stesso aperta, flessibile e ampiamente condivisa nel campo delle Digital Humanities, in grado di rappresentare strutture lessicografiche o lessicali complesse e/o non ancora completamente uniformate. In particolare, si è fatto riferimento alla sezione *Dictionaries* (cap. 10 delle *Guidelines*), che mette a disposizione un'ampia gamma di elementi e attributi, facilmente adattabili a progetti in fase esplorativa e non ancora vincolati a schemi predefiniti. Non è stato invece adottato, almeno in questa prima fase, il profilo TEI Lex-0, che propone un sottoinsieme ridotto e standardizzato degli elementi TEI, concepito per favorire l'interoperabilità tra risorse già codificate secondo standard comuni. Lex-0 è infatti orientato alla standardizzazione e all'allineamento di dizionari elettronici, ma – come ricordano gli stessi autori – non deve essere inteso come un «replacement of the Dictionaries Chapter in the TEI Guidelines or as the format that must be necessarily used for editing or managing individual resources, especially in those projects and/or institutions that already have established workflows based on their own flavors of TEI» (Tasovac *et al.* 2018). La sezione *Dictionaries* della TEI, al contrario, privilegia una rappresentazione più flessibile e dettagliata, che può essere costruita ad hoc, in base alle specificità di ciascuna risorsa, risultando maggiormente adatta al mark-up di strutture eterogenee e ancora in via di formalizzazione, come nel caso delle raccolte paremiologiche qui considerate (cfr. TEI P5).

9. Cfr. Tasovac *et al.* 2018.

10. Cfr. nota 2.

11. Biffi 2024, p. 66.

12. Questo bisogno fisiologico di accordo monosemico non può non emergere nel momento in cui gli addetti ai lavori avvertono la necessità di esprimere in maniera immediata e inequivocabile un concetto all'interno del proprio campo d'azione, specialmente quando vi



sono particolari progressi e avanzamenti tecnologici, come nel caso dell'umanistica digitale e delle sue branche, che vedono crescere e irrobustirsi tecniche, tecnologie e progetti; di conseguenza, anche il relativo bagaglio terminologico sarà suscettibile di consolidamento e perfezionamento.

13. Biffi 2024, p. 56.

14. Ivi, pp. 55-56. Va inoltre considerato che i repertori *di primo livello* e quelli *di secondo livello* comportano soluzioni molto diverse «sia dal punto di vista delle risorse da investire, sia, e conseguentemente, dal punto di vista della potenza dello strumento. Nel primo caso si potranno fare soltanto ricerche libere, come se il dizionario fosse un romanzo o un trattato (e comunque anche così il dizionario diventa un corpus rappresentativo della lingua: non è un aspetto da sottovalutare), nel secondo caso si potranno fare anche ricerche mirate tenendo conto della struttura della voce» (ivi, p. 55).

15. Un esempio è *GROBID-Dictionaries*, un software sviluppato da Mohamed Khemakhem come estensione di *GROBID* (acronimo di *GeneRation Of Bibliographic Data*) e ideato per marcare in modo semiautomatico dizionari *retro-digitized*. Il sistema, basato su tecniche di *machine learning* (in particolare modelli CRF, cioè *Conditional Random Fields*), permette di convertire documenti PDF in file .xml – dopo aver addestrato il modello su un campione di dati annotati – con tempi inferiori rispetto a quelli richiesti dal mark-up manuale. Tuttavia, la qualità dei PDF rappresenta un requisito essenziale: per assicurare un'elaborazione soddisfacente è necessario disporre di digitalizzazioni ad alta risoluzione e di un riconoscimento OCR accurato (naturalmente tale presupposto viene meno in presenza di scansioni di scarsa qualità o di testi con strutture particolarmente complesse e poco omogenee). Al di là di queste difficoltà, bisogna sottolineare che *Grobid-Dictionaries*, pur essendo applicabile a tipologie lessicografiche differenti, è progettato per dizionari semasiologici, costituendo così un ulteriore limite in termini di flessibilità e adattabilità (per approfondimenti si rimanda a Khemakhem *et al.* 2017 e Id. *et al.* 2018). Guardando al panorama italiano, a proposito delle procedure semiautomatiche applicate a dizionari di seconda generazione, si possono citare i lavori di informatizzazione legati al *Grande Dizionario della Lingua italiana* (GDLI), progetto promosso dall'Accademia della Crusca in collaborazione con l'Istituto di Linguistica Computazionale “A. Zampolli” del CNR, per cui si rimanda a Sassolini *et al.* 2019, Sassolini-Biffi 2020, Sassolini *et al.* 2021, Sassolini *et al.* 2024.

16. I test sono stati condotti utilizzando la versione *Plus* di ChatGPT (GPT-4), che consente funzionalità estese rispetto alla versione base.

17. Cfr. Riccio 2016, p. 64.

18. A un flusso di lavoro più ampio, che comprenda anche la fase di acquisizione del testo tramite OCR (e relativo post-processing), riserveremo uno studio e una trattazione a parte, in altra sede. L'intero processo è riassumibile in sette passaggi principali (che variano in base alla modalità di acquisizione del testo sorgente):

- *ipotesi A*: digitalizzazione del testo con OCR di alta qualità > post-processing e minima collazione > progettazione delle opzioni di ricerca > file .txt formalizzato (automatico, semiautomatico o manuale) > mark-up XML-TEI eseguito con ChatGPT > controllo ed eventuale arricchimento del mark-up > costruzione del sito;
- *ipotesi B*: file .txt ricavati da Google Libri/Archive/altro > collazione sorvegliata > progettazione delle opzioni di ricerca > file .txt formalizzato (automatico, semiautomatico o manuale) > mark-up XML-TEI eseguito con ChatGPT > controllo ed eventuale arricchimento del mark-up > costruzione del sito.

19. Con l'espressione *simboli-marcatori* ci riferiamo a dei simboli grafici (ad esempio >, \*, [, #) che, scelti arbitrariamente nella fase pre-markup per formalizzare e strutturare il testo in modo omogeneo, nel passaggio successivo (cioè quando il file formalizzato viene dato in input a ChatGPT per la marcatura automatica) assumono la funzione di marcatori, quindi di elementi facilmente riconoscibili dall'intelligenza artificiale per marcare automaticamente righe o blocchi di testo secondo le regole prestabilite.

20. Per questo aspetto voglio ringraziare la consulenza di Lorenzo Gregori, che mi ha suggerito i primi criteri per formalizzare i testi, necessari per farli elaborare correttamente da ChatGPT.

21. In questa prospettiva, abbiamo condotto alcuni test su file non formalizzati, quindi privi di simboli-marcatori utili a segmentare le unità testuali. I risultati sono stati fortemente negativi: ChatGPT, in assenza di una struttura formale esplicita, non è in grado di interpretare correttamente le istruzioni, anche quando queste sono dettagliate e ben formulate. Anche su testi brevi, il modello tende a commettere errori di segmentazione, saltare porzioni di testo o attribuire in modo errato il mark-up, poiché non riesce a distinguere in modo affidabile i diversi elementi strutturali; questo dimostra quanto sia fondamentale la fase preliminare di formalizzazione del file pre-markup, per guidare in modo efficace l'elaborazione automatica.

22. Cfr. <https://proverbi-rampoldi-prova.netlify.app> e <https://proverbi-alfani-prova.netlify.app>. Poiché in questa sede l'obiettivo principale è descrivere il flusso di lavoro, il metodo sperimentale e mostrare il livello dei risultati in una forma quasi grezza, i file .xml caricati nei siti pilota non sono stati sottoposti a una revisione completa (ad esempio per le parole chiave), e conterranno quindi alcuni errori.

23. Salgado *et al.* 2024, p. 126.

## Bibliografia

- Biffi 2024 = Marco Biffi, *Per una terminologia condivisa dei dizionari elettronici/digitali*. «Ladinia», vol. XLVIII, pp. 53-70, ISSN:1124-1004 DOI: <https://doi.org/10.54218/ladinia.48.53-70>.
- Khemakhem *et al.* 2017 = Mohamed Khemakhem, Luca Foppiano and Laurent Romary, *Automatic Extraction of TEI Structures in Digitized Lexical Resources using Conditional Random Fields*, in *Electronic lexicography*, eLex 2017, Leiden, Netherlands, September.
- Khemakhem *et al.* 2018 = Mohamed Khemakhem, Axel Herold and Laurent Romary, *Enhancing Usability for Automatically Structuring Digitised Dictionaries*, GLOBALEX Workshop at LREC 2018 (May 2018), Miyazaki, Japan.
- Patella 2023a = Barbara Patella, *Dizionari metodici dell'Ottocento: verso una piattaforma interrogabile*, in Emiliano Picchiorri e Francesco Montuori (a cura di), *In fieri, 4. Ricerche di linguistica italiana*, Atti della IV Giornata dell'ASLI per i dottorandi (Firenze, Accademia della Crusca, 2-4 dicembre 2021), Firenze, Cesati, 2023, pp. 241-247. ISBN: 979-12-5496-088-2
- Patella 2023b = Barbara Patella, *Versione elettronica dei principali dizionari metodici ottocenteschi della lingua italiana*, tesi di Dottorato in "Filologia, Letteratura italiana, Linguistica" (curriculum: "Umanistica Digitale"), Università degli Studi di Firenze, a.a. 2022/2023.
- Patella 2024a = Barbara Patella, *Codifica XML-TEI: proposta di mark-up per i dizionari metodici*, in AA., *I dati linguistici. Metodologie e strumenti della ricerca*, Studi e ricerche del Dipartimento di Lettere e Filosofia, Società Editrice Fiorentina, pp. 321-338, DOI: [10.35948/DILEF/978-88-6032-750-5.17](https://doi.org/10.35948/DILEF/978-88-6032-750-5.17).
- Patella 2024b = Barbara Patella, *La versione elettronica del Vocabolario toscano dell'arte del disegno (1681) di Filippo Baldinucci*. «Studi di Memofonte», vol. 33, pp. 153-180, ISSN:2038-0488 DOI: <https://doi.org/10.25433/2038-0488/15t9-7w77>.
- Riccio 2016 = Anna Riccio, *Gli strumenti per la ricerca linguistica. Corpora, dizionari e database*, Roma, Carocci.
- Salgado *et al.* 2024 = Ana Salgado, Laurent Romary, Rute Costa, Toma Tasovac, Anas Fahad Khan, Margarida Ramos, Bruno Almeida, Sara Carvalho, Mohamed Khemakhem, Raquel Silva e Boris Lehečka, *The Morais Dictionary: Following Best Practices in a Retro-digitized Dictionary Project*, *International Journal of Humanities and Arts Computing*. 18(1), pp. 125-147 (<https://doi.org/10.3366/ijhac.2024.0325>).
- Sassolini *et al.* 2019 = Eva Sassolini, Anas Fahad Khan, Marco Biffi, Monica Monachini, Simonetta Montemagni, *Converting and structuring a digital historical dictionary of Italian: a case study*, in *Electronic lexicography in the 21st century: smart lexicography*. Proceedings of the eLex 2019 conference (1-3 October 2019, Sintra, Portugal), Brno, Lexical Computing CZ, pp. 603-621 (<https://elex.link/elex2019/proceedings-download/>).
- Sassolini-Biffi 2020 = Eva Sassolini e Marco Biffi, *Strategie e metodi per il recupero di dizionari storici*, in *Atti del IX Convegno Annuale AIUCD. La svolta inevitabile: sfide e prospettive per l'Informatica Umanistica*, a cura di Cristina Marras, Marco Passarotti, Greta Franzini, Eleonora Litta, Milano, Università Cattolica del Sacro Cuore, pp. 235-239, DOI: [10.6092/unibo/amsacta/6316](https://doi.org/10.6092/unibo/amsacta/6316).
- Sassolini *et al.* 2021 = Eva Sassolini, Marco Biffi, Francesca De Blasi, Elisa Guadagnini, Simonetta Montemagni, *La digitalizzazione del GDLI: un approccio linguistico per la corretta acquisizione del*

---

testo?, in *AIUCD 2021: DH per la società: e-guaglianza, partecipazione, diritti e valori nell'era digitale. Raccolta degli abstract estesi della 10ª conferenza nazionale*, a cura di Angelo Mario Del Grosso, Federico Boschetti, e Enrica Salvatori, pp. 159-166. Pisa, 2021, [DOI: 10.6092/unibo/amsacta/6712](https://doi.org/10.6092/unibo/amsacta/6712).

- Sassolini *et al.* 2024 = Eva Sassolini, Sebastiana Cuccurullo, Marco Biffi, *L'informatizzazione del GDLI: risultati, prospettive, sfide future*, in *Atti del XIII Convegno Annuale AIUCD. Me.Te. Digitali. Mediterraneo in rete tra testi e contesti*, a cura di Antonio Di Silvestro e Daria Spampinato, pp. 281-286. ISBN:978-88-942535-8-0, [DOI: 10.6092/unibo/amsacta/7927](https://doi.org/10.6092/unibo/amsacta/7927).
- Tasovac *et al.* 2018 = Toma Tasovac, Laurent Romary, Piotr Banski, Jack Bowers, Jesse de Does, Katrien Depuydt, Tomaž Erjavec, Alexander Geyken, Axel Herold, Vera Hildenbrandt, Mohamed Khemakhem, Boris Lehečka, Snežana Petrović, Ana Salgado and Andreas Witt, *TEI Lex-0: A baseline encoding for lexicographic data*. Version 0.9.4. DARIAH Working Group on Lexical Resources, <https://dariah-eric.github.io/lexicalresources/pages/TEILex0/TEILex0.html>.
- TEI P5 2025 = *Guidelines for Electronic Text Encoding and Interchange*, Text Encoding Initiative Consortium, gennaio 2025 (consultabile su <https://tei-c.org/release/doc/tei-p5-doc/en/Guidelines.pdf>).