FLAVIA SCIOLETTE[1] – EMILIANO GIOVANNETTI[1]

Un modello per domarli tutti: verso una rappresentazione del testo come esplicitazione di documento, lingua e contenuto[2]

Abstract

The aim of this research is to describe the first steps towards the theoretical elaboration of a holistic model to represent textual information. The focus of the model is the definition of "text", with its different dimensions, as a "diasystem". The set of elements, organized into distinct but strictly interconnected systems, wherein each element has an effect on the whole diasystem, is described in a model structured in the following components: graphic, linguistic, documental, discursive, and conceptual. In this work, the first attempts in the modeling of text will be shown through two case studies: the Babylonian Talmud and the DiTMAO (Dictionary of Old Occitan medico-botanical terminology).

1. *Il testo come fonte di informazione. Standard e Vocabolari*[3]

Il presente lavoro si propone di documentare lo stato di avanzamento di una serie di ricerche relative alla modellizzazione del testo, da un punto di vista computazionale, in corso presso l'Istituto di Linguistica Computazionale da parte del costituentesi Knowledge Laboratory (KLab).

"Il testo è tutto il nostro bene" scriveva Cesare Segre in "Ritorno alla critica" (Segre 2001: 99); se questa affermazione conferma un *habitus* proprio delle scienze umanistiche nel loro complesso, tanto più si adatta alle *Digital Humanities* (DH), intese come disciplina, approccio e soprattutto metodo. Sono infatti numerosi i *task*, ormai tradizionali, che vedono nel testo – o meglio, nelle diverse accezioni del testo – il presupposto fondante, tanto come bacino di dati, quanto come veicolo di informazione di varia natura (linguistica, contestuale, concettuale).

Si parla di diverse accezioni dell'oggetto "testo" proprio perché, a seconda dell'ambito, è possibile mutare la lente, la prospettiva di analisi; da mera stringa di

[1] Istituto di Linguistica Computazionale "A. Zampolli" – CNR.

[2] Il presente lavoro è frutto del lavoro congiunto di entrambi gli autori; in particolare la scrittura dei paragrafi 1 e 4 è di Emiliano Giovannetti, dei paragrafi 2 e 3 di Flavia Sciolette.

[3] Si utilizza il termine "vocabolario" nell'accezione comunemente nota nell'ambito del Semantic Web, dunque da intendersi come "vocabularies define the concepts and relationships (also referred to as "terms") used to describe and represent an area of concern. Vocabularies are used to classify the terms that can be used in a particular application, characterize possible relationships, and define possible constraints on using those terms". (https://www.w3.org/standards/semanticweb/ontology).

caratteri ai prodotti dell'attività linguistica dei parlanti fino ad arrivare alle edizioni digitali di documenti e opere di diversa complessità. Le definizioni di testo possono dunque essere molteplici, ma da qualunque punto di vista lo si osservi, occorre partire dall'assunto che il testo è un oggetto pluridimensionale; il primo problema pratico nella sua rappresentazione è la necessità, dunque, di inquadrare da un punto di vista teorico e metodologico questa intrinseca e multifattoriale complessità.

L'interesse per il web semantico e i *Linked Open Data* (LOD), paradigmi oramai ampiamente compresi nell'ambito della linguistica computazionale, in particolare per il sottoinsieme dei *Linguistic Linked Open Data* (LLOD, intesi come dati condivisi, aperti e interoperabili per gli scopi della linguistica e del Natural Language Processing, Cimiano et al. 2020), trova ormai da tempo spazio nel campo della codifica dei testi (già in Ciotti 2014: 1-10 e Ciotti & Tomasi 2016, Monella & Del Turco 2020: 148-155 per le DSE – Digital Scholarly Editions). Tale interesse non sembra essere orientato solo in un senso; data la crescente attività di conversione di risorse digitali, anche non native, in formato LOD-*compliant*, il modello *LExicon Model for ONtologies* (lemon), elaborato dal gruppo Lexica[4] (McCrae et al. 2017) ha di recente pubblicato Lexicog[5], un modulo rivolto alla conversione di dizionari tradizionali (non focalizzato dunque solo al dizionario-macchina, inteso come insieme di informazioni strutturate, di natura linguistica, elaborabili dalla macchina). Questa attenzione per il testo, inteso come attestazione, si registra inoltre anche grazie all'elaborazione di un ulteriore modulo dedicato a questi aspetti, attualmente in fase di lavorazione sempre nell'ambito del gruppo Lexica (Chiarcos et al. 2020). Nel modello lemon, tuttavia, questa riflessione sul testo appare funzionale – per non dire ancillare – alla rappresentazione del dato lessicale e lessicografico, in un'ottica tuttavia perfettamente coerente tanto con i presupposti del modello, quanto con la sua vocazione agnostica rispetto alla descrizione degli oggetti considerati nel *data-model*. La filosofia soggiacente a lemon, in linea con il paradigma LOD, demanda a vocabolari ontologici esterni la descrizione della lingua e conseguentemente anche del testo.

In questo senso, si segnala la distribuzione CRMtex (http://www.cidoc-crm.org/crmtex/home-8, pubblicata nel 2020), la più recente incarnazione di CIDOC-CRM (Doerr 2003), importante ontologia formale nel campo della conservazione del patrimonio culturale CRMtex si presenta come un'estensione deputata alla descrizione dei testi scritti antichi, intesi come entità comprendenti un insieme di glifi su un determinato supporto, dotati di scopo e in grado di veicolare un contenuto. Il modello introduce classi e proprietà indirizzati nello specifico agli esperti di discipline quali codicologia, epigrafia e paleografia. L'attenzione è pertanto rivolta al testo in quanto oggetto materiale e il vocabolario introdotto mira a fornire le etichette per una descrizione rigorosa, utile anche in termini di scambio di dati e interope-

[4] https://www.w3.org/community/ontolex/ Il gruppo fa parte della community W3C (World Wide Web Consortium), secondo il "W3C Community Final Specification Agreement (FSA). Di conseguenza il modello non è da considerarsi uno standard W3C.

[5] https://www.w3.org/2019/09/lexicog/.

rabilità su risorse diverse (si veda per esempio Bellotto 2020, per un'applicazione nell'ambito di manoscritti medievali, con analisi anche dei vantaggi dell'adozione del vocabolario unitamente a TEI per la risoluzione di alcune questioni teoriche e di metodo relative alla codifica).

Il panorama, sebbene fluido, mostra tuttavia approcci di modellizzazione del testo con un focus specifico, improntati al *task* piuttosto che al suo carattere n-dimensionale, ovvero della sua specifica complessità, in quanto insieme di documento, lingua e contenuto. Il presente lavoro tenta dunque di fornire una possibile risposta teorica in corso di elaborazione, fondata sia dal punto di vista semiotico che computazionale. Il punto di partenza sono i diversi casi d'uso forniti dall'attività svolta durante il progetto di traduzione del Talmud Babilonese in italiano e il progetto DiTMAO (Dictionary of Old Occitan medico-botanical terminology).

2. *Un modello per l'oggetto testo*

2.1 L'esplicitazione dell'informazione

La definizione rigorosa e formale di un oggetto e delle sue diverse componenti, delle relazioni che intercorrono tra di esse e delle proprietà che le caratterizzano, nonché dei diversi processi in cui possono essere coinvolte, è *conditio sine qua non* per l'applicazione di un approccio computazionale. Ogni operazione in cui è coinvolto un testo – sia essa la lettura, la scrittura o la traduzione – vede il coinvolgimento di un interprete e, conseguentemente, del suo punto di vista. In particolare, nel processo di traduzione, un interprete (ed esperto di dominio) utilizza l'informazione contenuta nel testo per la produzione di un nuovo oggetto (la traduzione). Questa informazione impiegata per strutturare il nuovo oggetto testo, tuttavia, non è rappresentata solo da quanto potrebbe essere veicolato esplicitamente dal testo – ovvero tutta l'informazione disponibile nel testo in quanto stringa di caratteri – ma comprende anche l'informazione implicita, ovvero quella che l'interprete umano è in grado di decodificare per mezzo della propria enciclopedia mentale (per esempio il significato di determinate espressioni idiomatiche).

Gli strumenti elaborati dal nostro gruppo di ricerca sono pensati per supportare l'attività dell'interprete umano grazie ai dati estratti: statistiche linguistiche, allineamenti di parola, occorrenze, frequenze, concordanze. Secondo un approccio votato alla collaboratività tra interpreti umani, sono stati elaborati strumenti volti alla traduzione assistita (come Traduco, sviluppato nell'ambito del progetto di traduzione del Talmud, Giovannetti et al. 2016) e alla creazione di risorse terminologiche e lessicali (come LexO, editor di lessici computazionali basato sul modello lemon, Bellandi 2021). Gli strumenti rappresentano uno dei quattro pilastri[6] su cui

[6] Gli altri tre pilastri sono: i) gli algoritmi dedicati al trattamento automatico della lingua, per esempio per l'estrazione della conoscenza e della terminologia (Giovannetti et al. 2020); ii) le risorse linguistiche (come nel caso del lessico computazionale dell'italiano Parole-Simple-Clips, il cui aggiornamento è in corso da parte del KLab al momento della redazione del presente contributo); iii) il modello volto

si basano le ricerche condotte dal KLab nell'ambito della modellazione, della rappresentazione e del trattamento del testo.

Secondo quanto appena esposto, i dati estraibili per mezzo degli strumenti non rappresentano tuttavia tutta l'informazione veicolata dal medium di trasmissione del testo. Ogni oggetto complesso, difatti, porta dell'informazione implicita, costituita da dati che non compaiono come segni espliciti – tali da poter essere individuati dagli strumenti e dunque estratti direttamente – ma che sono tuttavia presenti in ogni processo legato al testo (di cui la traduzione rappresenta probabilmente uno degli esempi più complessi). In altre parole, per poter sfruttare tutta l'informazione contenuta in un testo, in prima istanza, abbiamo bisogno di "esplicitarla". Necessitiamo quindi di poter modellare e rendere ogni elemento del testo un oggetto digitale esplicito, dotato di un proprio identificatore univoco (come *Uniform Resource Identifier*, o URI), delle proprietà caratterizzanti e delle relazioni con altri oggetti: chiamiamo questo processo "esplicitazione computazionale dell'informazione testuale". La seconda parte dell'articolo sarà dedicata all'esposizione di alcuni esempi in cui mostreremo questo processo nella pratica.

Una volta dati i presupposti per avere disponibile questa informazione, è necessario altresì disporre di un modello nel quale essa possa essere organizzata. Questo modello olistico deve dunque essere focalizzato sul testo come contenuto, sull'uso della lingua al suo interno (quindi non su una sua formalizzazione) e soprattutto deve essere sufficientemente espressivo da permettere l'utilizzo di standard esistenti, come quelli sopramenzionati. Di conseguenza, il nostro obiettivo non è, di per sé, la creazione di un nuovo *data-model* o un tentativo di sostituire standard esistenti, quanto piuttosto fornire: 1) una prospettiva di visione di insieme del dato testuale in ottica computazionale; 2) un vocabolario per le interazioni tra componenti che solitamente vengono considerate separatamente, utile per *task* che richiedono di elaborare informazione di diversa natura; 3) uno schema per agevolare la conversione e l'interoperabilità tra risorse basate su modelli diversi, aventi il testo come oggetto di interesse. Da qui il titolo del presente contributo, volutamente iperbolico, da leggere come obiettivo ideale per la realizzazione di un modello del testo in grado di fornire una descrizione dei suoi diversi aspetti, utile per specialisti di diversi ambiti.

In particolare, per il linguista computazionale e l'informatico della lingua questo modello intende agevolare la sperimentazione di elaborazioni della lingua (in analisi e in generazione) attraverso approcci in grado di integrare conoscenza linguistica e conoscenza del mondo e, inoltre, creare risorse linguistiche ricche e articolate (lessici, terminologie, corpora annotati, corpora allineati, ecc.), con strumenti e metodologie utili all'arricchimento delle risorse stesse, attraverso un modello rigoroso atto all'elaborazione di informazioni appartenenti a risorse di natura diversa. Per lo studioso del testo e della lingua, si intende fornire un modello (affiancato da un set di strumenti e algoritmi per utilizzarlo) per la creazione di risorse linguistiche e corpora annotati che offrano funzionalità di interrogazione avanzata, una formaliz-

alla rappresentazione dell'informazione testuale, oggetto di questo articolo e ultimo menzionato, ma in realtà fondamento di tutta l'architettura delineata.

zazione dell'informazione puntuale e rigorosa che sia nativamente condivisibile con la propria comunità di riferimento e, infine, la possibilità di arricchire collaborativamente le risorse già esistenti. Oltre all'estrazione di dati, il processo di esplicitazione dell'informazione si pone come ulteriore mezzo per lo studio del testo, attraverso possibilità di interrogazioni complesse per il corpus.

Preme precisare in questa sede che il modello non nasce per sostituire altri modelli già esistenti, focalizzati su specifici sistemi, né si propone come un'ontologia del testo. La proposta ambisce soprattutto a fornire una sovrastruttura per l'organizzazione generale di tutta l'informazione contenuta in un testo, a prescindere dalle tecniche utilizzate per estrarla in modo automatico o semiautomatico.

La necessità di un tale meta-modello nasce dagli esperimenti condotti su testi particolarmente complessi, spesso scritti in varietà con poca documentazione o non-standard o legate a specifici domini; queste tipologie testuali spesso risultano difficili da analizzare secondo approcci stocastici, in quanto possono essere scarsamente rappresentate nei corpora di addestramento. In linea generale l'obiettivo tuttavia non è, anche in questo caso, porsi in diretta antitesi con differenti approcci, quanto piuttosto affiancarsi a questi ultimi.

2.2 Il modello: diasistema, sistema e dimensioni

Il modello che qui si intende presentare, da un punto di vista teorico, si avvale delle innovazioni metodologiche proposte nei lavori di Tito Orlandi (formalizzate in Orlandi 2010) relativamente all'edizione critica digitale; sebbene Orlandi parta dai manoscritti, gli assunti di base che propone sono applicabili all'oggetto testuale nelle sue diverse accezioni. Secondo Orlandi, il testo è un "diasistema", inteso come prodotto, oggetto complesso risultante dall'interazione dialettica di più sistemi complessi (quindi sistema a sua volta), riprendendo una terminologia utilizzata già dal sopra menzionato Cesare Segre (ma la nozione di "sistema" e di "sistemi" in cui sono inseriti i testi ha una tradizione di lungo corso negli studi umanistici), in cui interagiscono diversi sottosistemi, tra cui quello linguistico (composto dalle parole grafiche, concepite come entità a sé stanti) e quello grafematico (di cui le unità sono i diversi grafemi aventi valore distintivo). Ognuno degli elementi considerato in questi sistemi interagisce con l'altro.

Parliamo di "sistema", infatti, in presenza di due o più elementi che soddisfino le seguenti condizioni: 1) ogni elemento del sistema ha un effetto sul comportamento del sistema nella sua interezza; 2) gli elementi sono interdipendenti tra di loro in quanto legati da relazioni; 3) gli elementi di un sottoinsieme hanno un effetto sul sistema nella sua interezza (Ackoff 1971; Skyttner 1996). La metafora del diasistema, inoltre, risulta particolarmente produttiva anche in un ambito di analisi computazionale del testo, in quanto ci consente di scomporre l'informazione testuale in moduli (i sistemi, per l'appunto) secondo elementi omogenei in relazione, senza concepirne un'obbligatoria gerarchizzazione (e quindi riuscendo a evitare il ricorso a nozioni come "livello" o "strato" del testo). In questo modo possiamo descrivere tutti gli elementi semanticamente e semioticamente fondanti di un testo.

Gli elementi del testo possono essere raggruppati tra loro in modo omogeneo in base alle loro caratteristiche e relazioni; per esempio, possiamo raggruppare tutti gli elementi di lingua; tutti i concetti espressi all'interno di un testo; tutte le caratteristiche materiali del veicolo di trasmissione, e costituire quindi un sistema (linguistico, concettuale o documentale). Allo stesso modo, possono essere messi in relazione elementi appartenenti a sistemi diversi, per esempio nel caso di termini da legare ai rispettivi concetti, nel corrispondente sistema.

In un modello del testo, secondo l'impostazione teorica orlandiana, è possibile distinguere diversi sistemi; per le esigenze delle nostre ricerche ne abbiamo, al momento, individuati cinque: 1) il sistema linguistico, ovvero il sistema riferito alla lingua, in quanto codice e mezzo di espressione dei contenuti del testo, nonché realizzazione di una varietà propria dell'agente che ha prodotto il testo; 2) il sistema grafico, ovvero il sistema riferito al *vehiculum*, inteso come l'insieme dei segni che compongono l'oggetto testo; 3) il sistema documentale, ovvero la rappresentazione del testo in quanto documento, comprendente le sue caratteristiche materiali e le informazioni codificabili come metadati; 4) il sistema discorsivo, riferito alle regole legate alla tradizione discorsiva del testo, alle regole di pragmatica e all'insieme delle caratteristiche non riconducibili a una grammatica esplicita; 5) il sistema concettuale, ovvero l'espressione della conoscenza del mondo espressa dal testo e dall'agente che lo ha prodotto.

I sistemi consentono di ordinare l'informazione testuale e di "spacchettarla" in diverse tipologie, che possono essere messe in relazione tra loro.

Ogni sistema, a sua volta, può essere diviso in dimensioni: definiamo "dimensione" un determinato aspetto del sistema entro il quale trovano spazio elementi omogenei, appartenenti a un insieme individuato secondo un determinato punto di vista, che chiameremo "prospettiva". Alcune delle dimensioni di un certo sistema devono essere considerate "raccomandate", se si tiene conto della natura del sistema a cui appartengono (per esempio la dimensione sintattica o quella morfologica nel sistema linguistico), tuttavia altre possono dipendere dalla teoria di riferimento o dal *task* prefissato.

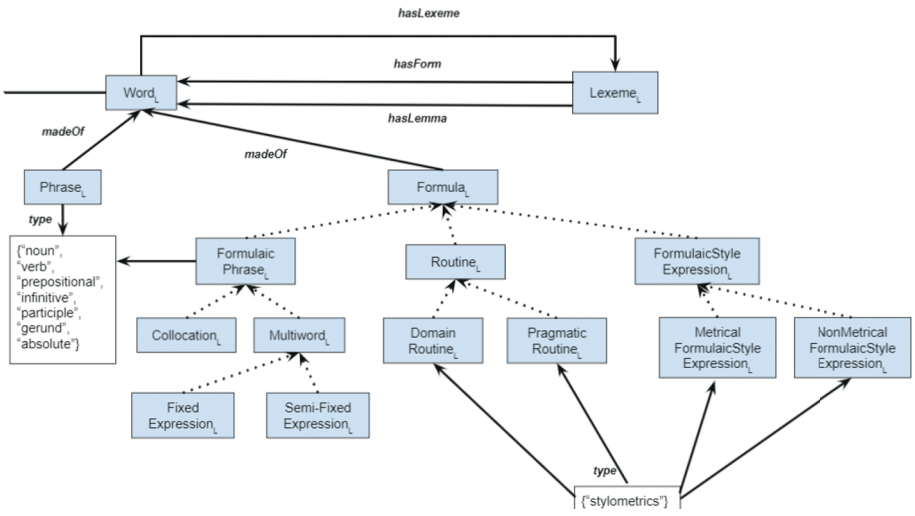
Gli elementi contenuti in una data dimensione costituiscono le unità di analisi e veicolano l'informazione da esplicitare, relativa a una specifica dimensione; la dimensione terminologica, ad esempio, si riferisce ai termini di un determinato dominio.

La nozione di dimensione, come sopra detto, non presuppone un ordinamento gerarchico, nonostante due o più dimensioni possano essere legate gerarchicamente da un punto di vista procedurale (per esempio, si consideri la dimensione sintattica di una lingua flessiva, che richiede comunque una relazione con la dimensione morfologica, perché altrimenti, da un punto di vista procedurale, non sarebbe possibile identificare l'organizzazione delle sue strutture sintattiche. L'organizzazione non strettamente gerarchica del modello, tuttavia, permette la descrizione anche di casi diametralmente opposti, come per le lingue isolanti). I legami possono sussistere tra elementi di dimensioni dello stesso sistema – come nel caso dei sintagmi, che

devono necessariamente considerare elementi di lessico, morfologici e sintattici – e tra dimensioni di diversi sistemi, nel momento in cui, per esempio, si formalizza la relazione tra un termine e un concetto (definendo così una relazione tra il sistema linguistico e il sistema concettuale) oppure la relazione tra una determinata struttura sintattica e il suo utilizzo in specifici generi testuali (relazione tra sistema linguistico e sistema discorsivo).

Di seguito si presenta un esempio di formalizzazione di alcuni elementi del modello. Abbiamo scelto di rappresentare il concetto di "Formula", inteso come componente del *formulaic language*, secondo l'accezione utilizzata in Wray (2013): "'formulaic language' refers to sequences of words that are in some regard not entirely predictable, whether on account of a meaning that is wildly or subtly different from the words they contain, a function that is only achieved with the whole expression, or features of structure such as morphology or word order that are non-canonical." La figura 1 mostra le relazioni tra gli elementi in tassonomia e le proprietà di ciascuna classe. Per spiegare l'esempio, si prenda come punto di partenza il concetto di Formula, inteso come classe delle formule in una Lingua (L). Le frecce tratteggiate descrivono le sottoclassi della classe Formula: i) Formulaic Phrase, ii) Routine, iii) Formulaic expression, a loro volta distinte in ulteriori sottoclassi. Le frecce non tratteggiate istituiscono le relazioni tra classi (come nel caso della relazione tra la classe delle formule e la classe delle parole) e tra classi e istanze (come nel caso della classe di Phrase, costituita dalle istanze con attributo *type* e i valori indicati in tabella). In questo caso la classe Formula nel modello è legata alla classe Word dalla relazione madeOf.

Figura 1 - *Rappresentazione della classe "Formula" nel sistema linguistico del modello. Le frecce tratteggiate indicano la relazione di "isA".*



Si tratta di un esempio di dimensioni contenute e di cui non sono ancora disponibili tutte le singole ramificazioni, utile tuttavia per chiarire la struttura che si intende dare all'informazione testuale.

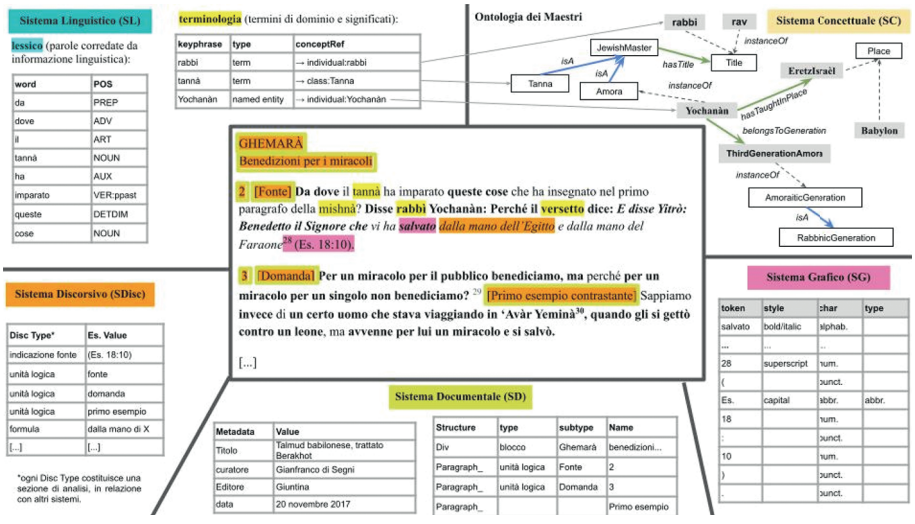
Nel paragrafo successivo mostreremo alcuni casi d'uso di applicazione del modello, utili per fornire un primo inquadramento dei possibili dati con cui è possibile popolare risorse basate su di esso. I casi d'uso illustrati sono stati presi da due progetti che hanno coinvolto il trattamento di risorse testuali e linguistiche e sulla base dei quali sono state inizialmente concepite alcune parti del modello stesso.

3. *Casi d'uso*

3.1 La traduzione del Talmud Babilonese

Il primo caso d'uso illustrato tratta l'esplicitazione e la strutturazione, nei cinque sistemi del modello, dell'informazione contenuta nel testo della traduzione italiana del Talmud babilonese. Ogni elemento da esplicitare appare annotato nel testo e, allo stesso tempo, strutturato nel sistema e nella dimensione opportuna. L'annotazione del testo, nel nostro paradigma di esplicitazione, non viene effettuata *inline* – ovvero come tipologia di *mark-up* in cui gli elementi sono disposti nella stringa di testo – ma ogni elemento viene annotato mediante riferimenti localizzati esternamente al testo (*stand-off markup*) che identificano, con opportuni indici, l'inizio e la fine dell'annotazione nel testo. Nell'illustrazione di figura 2, ogni elemento esplicitato viene collocato in una tabella posta in una sezione relativa a uno specifico sistema.

Figura 2 - *Esempio di esplicitazione e rappresentazione dell'informazione testuale nel caso d'uso della traduzione del Talmud*



Nel primo riquadro in alto a sinistra è rappresentato il sistema linguistico, strutturato attraverso una tabella che raccoglie gli elementi della dimensione lessicale (le

parole corredate da informazione linguistica) e della dimensione terminologica (i termini di domino[7], le entità nominate e i rispettivi sensi). Ricordiamo che il sistema linguistico può comprendere anche aspetti di morfologia, sintassi, ecc., da formalizzarsi mediante apposite dimensioni.

La tabella del lessico mostra, a sinistra, la colonna *word* dedicata alle forme di superficie delle parole, a destra, la relativa parte del discorso (PoS), mentre la terminologia è organizzata in *keyphrase* (la forma canonica di superficie), *type* (l'appartenenza a una determinata classe, come "term" o "named entity") e *conceptRef* (la relazione tra la parola e la relativa entità concettuale).

La parola *tannà*, per esempio, può essere dunque annotata ed esplicitata, secondo una differente prospettiva: come *word* (elemento del lessico) o *term* (termine di dominio); in questo secondo caso, il termine "tannà" viene associato, tramite conceptRef, alla classe ontologica dei Tannà (saggi rabbini del periodo mishnaico), che, nel sistema concettuale, è codificata come sottoclasse di *JewishMaster*. La parte ontologica, con le classi, gli individui e le relazioni che descrivono la natura del mondo riferito dal testo in esame, viene descritta e formalizzata nel sistema concettuale. I termini sono espressi dunque come lessicalizzazioni, collocate nel testo, di individui o classi, le cui singole proprietà concettuali sono descritte nel riquadro in alto a destra.

Nel riquadro in basso, sempre a destra, sono rappresentati gli elementi del sistema grafico; sebbene in un testo a stampa queste informazioni potrebbero sembrare più contenute o meno rilevanti, è comunque possibile catalogare e immagazzinare informazione, per esempio, relativa al font e allo stile del carattere. Difatti, nell'ambito del progetto Talmud, il grassetto (*Bold*, nel campo "Style" nel sistema grafico) ha un valore distintivo, in quanto indica le porzioni di testo tradotte letteralmente dal Talmud. Per poter sfruttare questo tipo di informazioni, a ogni token, come si può notare nella prima colonna disposta in alto a sinistra, vengono associate informazioni relative a carattere, stile e tipologia, in maniera tale da avere una rappresentazione formale anche di segni di interpunzione, cifre e abbreviazioni.

La sezione in basso al centro è dedicata al sistema documentale; comprende i metadati del testo in esame, organizzati nella tabella di sinistra, e le informazioni pertinenti alla struttura del testo, secondo una divisione per unità logiche. Ogni valore è definito attraverso l'associazione con *type, subtype* e *name*.

La sezione in basso a sinistra dell'immagine rappresenta il sistema discorsivo; all'interno di questa sezione si annotano i valori che, nel sistema documentale, corrispondono a delle unità di struttura del testo; queste ultime, tuttavia, possono essere associate anche alla classe *DiscType*, in quanto rappresentano una caratteristica del genere. La divisione in unità logiche del testo talmudico difatti non risponde solo a criteri editoriali di struttura materiale della pagina, ma le componenti costituiscono vere e proprie divisioni, strutturate secondo regole precise, legate all'andamento

[7] Il campione di 4166 termini di dominio utilizzato per i primi test di modello è stato estratto grazie al software T2K (Dell'Orletta et al. 2014); la procedura nel dettaglio è descritta in Giovannetti et al. 2020.

del discorso nell'ambito delle discussioni rabbiniche. Lo schema "Fonte-Domanda-Esempio" viene scomposto nei suoi elementi fondamentali, che costituiscono valori nella tabella del sistema discorsivo. Non riguarda dunque solo la struttura del testo come fisicamente disposto nell'impaginazione, ma risponde a regole precise del genere testuale, in quanto pattern proprio del testo talmudico. Le annotazioni sovrapposte al testo sottolineano la duplice natura informativa del segmento di attestazione scelto.

La selezione dei casi d'uso mira a fornire una panoramica delle possibilità di utilizzo del modello. Di fronte alla ricchezza dell'esempio fornito dal testo talmudico, potrebbe infatti essere lecito chiedersi se l'espressività del modello debba trovare sua giustificazione solo in un utilizzo completo di tutta l'architettura dei sistemi.

L'esempio che segue, pertanto, descrive una situazione che non richiede l'uso di tutti i sistemi, mostrando quanto il modello sia flessibile e adattabile a diversi contesti di studio.

3.2 Un'entrata nel DiTMAO

Nel caso del DiTMAO, il dizionario di antico occitano[8] dei termini medico-botanici, si considera come esempio il termine assafetida e la sua attestazione, contenuta nel manoscritto Vat. Ebr. 550, alla carta 115r. Il manoscritto contiene il testo *Tractatulus de Pestilentia*, che ha una sua edizione di riferimento, ma l'occorrenza di assafetida si registra nel successivo glossario non edito. Di conseguenza l'attestazione è data solo dallo spoglio manuale del manoscritto, disponibile in formato immagine[9] non ricercabile. Si è scelto di presentare questo esempio proprio per la natura del supporto e per mostrare come il modello si adatti a un approccio multilingue, utile nell'ottica di risorse lessicografiche.

A differenza di quanto è stato fatto nell'esempio del testo talmudico, l'annotazione del testo non può essere effettuata specificando un carattere di inizio e di fine all'interno di una sequenza testuale, ma deve implementarsi attraverso la definizione di una porzione di immagine relativa al manoscritto. In questo caso specifico, la carta 115r si presenta divisa in tre colonne: la prima colonna identifica il lemma; la seconda comprende la formula – struttura linguistica introdotta in 2.1 – che collega il lemma alla glossa, presente nella terza colonna. Questa struttura del folio "a colonne" si riflette nel sistema documentale, descritto nelle tabelle che seguono. Le tre colonne vengono descritte nei campi *structure*, *type* e *subtype* (tabella 1), i cui valori comprendono la struttura materiale del testo (in questo caso *column,* le colonne in cui è divisa la pagina) la tipologia testuale in cui sono inserite (List) e infine il subtype, che indica l'elemento presente ("word", "formula", "gloss"). La word descrive la forma di superficie dell'entrata descritta nel dizionario, mentre "formula" registra che l'occorrenza compare sempre preceduta da "esso è" (nel testo "אוה", tabella 2). La formula compare nella struttura del testo, in quanto colonna, e al tempo stesso è compresa

[8] Per la composizione del corpus: https://www.uni-goettingen.de/en/the+texts/487591.html.
[9] https://digi.vatlib.it/view/MSS_Vat.ebr.550/0233.

nei valori di DiscType in quanto costituisce una specifica caratteristica della tipologia testuale del glossario. Nel sistema grafico vengono aggiunti i dati relativi alla *scripta*, alla trascrizione in caratteri ebraici e alla traslitterazione del token.

Tabella 1 - *Intabellamento dei valori nel sistema documentale.*

Structure	*type*	*subtype*
Column_1	_list	word
Column_2	_list	formula
Column_3	_list	gloss

Tabella 2 - *Intabellamento dei valori per il DiscType nel sistema discorsivo*

DiscType	*value*
word_source	list
word_gloss	list
formula_gloss	"esso è"

La relazione tra termine e classe si esprime attraverso la *conceptRef*, in quanto istanza di una classe dell'ontologia "Substance"; si sceglie di rimanere nella dimensione terminologica, senza ulteriori collegamenti al sistema concettuale. In questo modo il modello consente la modellazione dell'entrata lessicale secondo il vocabolario lemon, ma consente anche di "spacchettare" l'informazione relativa a una delle attestazioni, dividerla in singole componenti che possono essere messe in relazione con l'entrata stessa e utilizzate in diversi processi. Esempi tipici sono l'optical character recognition (OCR) o il *pattern-disambiguation*, che possono in questo modo avvalersi di informazione relativa sia al contesto della parola e alla sua posizione in uno schema fondato – secondo la tipologia testuale – sia al significato disambiguato della parola precedente.

4. *A che punto siamo? Prospettive future*

In questo intervento abbiamo voluto rappresentare i primi passi volti alla realizzazione di un modello olistico per la rappresentazione computazionale del testo e la strutturazione di informazione testuale eterogenea.

Abbiamo passato in rassegna i principali vocabolari e standard attualmente esistenti per la rappresentazione del testo in formato digitale per differenti scopi; successivamente abbiamo introdotto la nozione di processo di esplicitazione computazionale dell'informazione testuale, da ordinare secondo un modello organizzato in sistemi e dimensioni. Infine abbiamo illustrato due casi d'uso provenienti da contesti molto differenti tra loro (la traduzione di un testo complesso e multilingua come il Talmud e la costruzione di una terminologia per l'occitano medievale) per dimostrare le differenti possibilità di utilizzo del modello.

Trattandosi di una ricerca tuttora in corso, alcuni aspetti progettuali e architetturali relativi alla definizione di specifiche componenti del modello e delle possibili implementazioni delle risorse su di esso basate devono ancora essere chiariti. Da un punto di vista più teorico, e coerentemente alla nostra volontà di i) riutilizzare, laddove possibile, modelli locali (lessicali, ontologici, ecc.) già disponibili e, allo stesso tempo, ii) aderire alle buone pratiche del paradigma dei Linked Data, intendiamo innanzitutto consolidare la definizione dei sistemi linguistico e concettuale, il primo da basarsi sul già citato lemon, e il secondo sul linguaggio OWL (Web Ontology Language), con interventi di adattamento qualora necessari. Inoltre, riteniamo prioritario definire formalmente le modalità di collegamento di ognuno degli elementi presenti nella risorsa alle relative porzioni di testo in cui essi sono stati esplicitati; ciò potrebbe essere possibile attraverso l'adozione del modulo di lemon dedicato alle attestazioni (non appena sarà reso disponibile) oppure, più in generale, collegando gli elementi esplicitati, tramite i rispettivi URI, alle annotazioni definite sul testo, mediante lo stand-off markup.

Le prime risorse che saranno realizzate sulla base del modello presentato (e pubblicate come LLOD) afferiranno al dominio religioso, e verranno allestite attingendo ai testi e ai dati trattati nell'ambito del citato Progetto Traduzione del Talmud Babilonese e del progetto PRIN 2017 "Representing religious diversity in Europe: past and present features". Per quanto riguarda le applicazioni pratiche del modello, si stanno conducendo primi esperimenti nell'ambito di task di full text search, condotto sul caso studio costituito dal testo del Talmud. Un modello fortemente orientato al testo e volto all'organizzazione di informazioni di varia natura, si rivela particolarmente vantaggioso su testi complessi, che necessitano di un sistema di interrogazione sofisticato e che combini diverse caratteristiche (semantiche, morfologiche, ontologiche, ecc.ecc.) al fine di recuperare informazioni a grana molto fine.

Bibliografia

Ackoff, Russel. 1971. Towards a system of systems concepts. *Management Sciences* 17(11). 661-671.

Bellandi, Andrea. 2021. LexO: An Open-source System for Managing OntoLex-Lemon Resources. *Language Resources & Evaluation*, 55, 1093-1126. https://doi.org/10.1007/s10579-021-09546-4.

Bellandi, Andrea & Giovannetti, Emiliano & Weingart, Anja. 2018. Multilingual and Multiword Phenomena in a lemon Old Occitan Medico-Botanical Lexicon. *Information* 9(3). doi:10.3390/info9030052.

Bellotto, Anna. 2020. Medieval manuscript descriptions and the Semantic Web: analysing the impact of CIDOC CRM on Italian codicological-paleographical data. *DHQ: Digital Humanities Quarterly* 14(1). digitalhumanities.org/dhq/vol/14/1/000449/000449.html.

Chiarcos, Christian & Ionov, Maxim & de Does, Jesse & Depuydt, Katrien & Khan, Anas Fahad & Stolk, Sander & Declerck, Thierry & McCrae, John P. 2020. Modelling Frequency and Attestations for OntoLex-Lemon. In *Proceedings of the Globalex Workshop on Linked*

Lexicography, Language Resources and Evaluation Conference (LREC 2020), Marseille, 11-16 May 2020. 1-9.

Cimiano, Philip & Chiarcos, Christian & McCrae, John P. & Gracia, Jorge. 2020. Linguistic linked open data cloud. In Cimiano, Philip (ed.) *Linguistic Linked Data*, 29-41. Cham: Springer.

Ciotti, Fabio. 2014. Tematologia e metodi digitali: dal markup alle ontologie. In Alfonzetti, Beatrice & Baldassarri, Guido & Tomasi, Franco (a cura di) *I cantieri dell'italianistica. Ricerca, didattica e organizzazione agli inizi del XXI secolo. Atti del XVII congresso dell'ADI – Associazione degli Italianisti (Roma Sapienza, 18-21 settembre 2013)*, 1-10. Roma: Adi editore.

Ciotti, Fabio & Tomasi, Francesca. 2016. Formal ontologies, Linked Data and TEI Semantics. *Journal of the Text Encoding Iniziative* 9. 1-23. https://doi.org/10.4000/jtei.1480.

Dell'Orletta Felice & Venturi Giulia & Cimino Andrea & Montemagni Simonetta. 2014. T2K²: a System for Automatically Extracting and Organizing Knowledge from Texts. *Proceedings of the 9th Edition of International Conference on Language Resources and Evaluation (LREC 2014)*, 26-31 May, Reykjavik, Iceland, 2062-2070.

Doerr, Martin. 2003. The CIDOC Conceptual Reference Module: An Ontological Approach to Semantic Interoperability of Metadata. *AI Magazine* 24(3). 75-92. doi:10.1609/aimag.v24i3.1720.

Giovannetti, Emiliano & Albanesi, Davide & Bellandi, Andrea & Benotto, Giulia. 2016. Traduco: A collaborative web-based CAT environment for the interpretation and translation of texts. *Digital Scholarship in the Humanities,* I47-i62. doi:10.1093/llc/fqw054.

Giovannetti, Emiliano & Bellandi, Andrea & Dattilo, David & Del Grosso, Angelo Maria & Marchi, Simone & Pecchioli, Alessandra & Piccini, Silvia. 2020. The Terminology of the Babylonian Talmud: Extraction, Representation and Use in the Context of Computational Linguistics. *Materia Giudaica* XXV. 61-74.

McCrae, John P. & Bosque-Gil, Julia & Gracia, Jorge & Buitelaar, Paul & Cimiano, Philipp. 2017. *The OntoLex-Lemon Model: development and applications*, http://john.mccr.ae/papers/mccrae2017ontolex.pdf.

Monella, Paolo & Rosselli Del Turco, Roberto. 2020. Extending the DSE: LOD support and TEI/IIIF integration in EVT. In Marras, Cristina & Passarotti, Marco & Franzini, Greta & Litta, Eleonora (a cura di) *Atti del IX Convegno annuale AIUCD. La svolta inevitabile: sfide e prospettive per l'Informatica Umanistica*. 148-155.

Orlandi, Tito. 2010. *Informatica testuale. teoria e prassi*. Roma: Laterza.

Segre, Cesare. 2001. *Ritorno alla critica*. Torino: Einaudi.

Skyttner, Lars. 2001. *General systems theory*, Singapore: World Scientific Publishing Co.

Wray, Alison. 2013. Formulaic Language. *Language Teaching* 46(3). 316-334. doi: 10.1017/S0261444813000013.