OVI OPERA DEL VOCABOLARIO ITALIANO
ISTITUTO DEL CONSIGLIO NAZIONALE DELLE RICERCHE

Criteri per l'aggiornamento
del *Corpus TLIO per il vocabolario* e del *Corpus OVI dell'italiano antico*

Con il progetto «CoVo. Il corpus del vocabolario italiano delle origini: aggiornamento filologico e interoperabilità» (PRIN 2015), coordinato da Lino Leonardi presso l'Università di Siena e attivo nel triennio dal febbraio 2017 al febbraio 2020, l'OVI ha riavviato l'aggiornamento sistematico del *Corpus TLIO* e del *Corpus OVI dell'italiano antico*. Grazie alla collaborazione in particolare di due Unità di Ricerca del progetto oltre a quella centrale, una presso l'OVI (responsabile Pär Larson) e l'altra presso l'Università per Stranieri di Siena (responsabile Giuseppe Marrani), tale opera di aggiornamento ha potuto essere impostata in modo sistematico, consentendo un ripensamento più generale del sistema di corpora che l'OVI mette a disposizione della comunità scientifica tramite il sistema GATTOweb.

Si fornisce qui una sintesi della nuova conformazione dei due corpora principali dell'OVI, e dei criteri adottati per il loro aggiornamento. A partire dal 2018, i testi via via aggiornati sono periodicamente segnalati sul sito web dell'OVI.

Corpus TLIO per il vocabolario.

Funzioni. Il *Corpus TLIO per il vocabolario* è stato fin dalla sua costituzione, e si conferma oggi, il corpus di riferimento per la redazione delle voci del «Tesoro della Lingua Italiana delle Origini» (TLIO), il vocabolario storico in allestimento a cura dell'OVI. Il corpus è lemmatizzato, non in forma esaustiva (non tutte le occorrenze sono associate a un lemma), ma in forma sistematica (tutte le forme sono associate a un lemma, o a più d'uno), ed è quindi interrogabile per lemmi, oltre che per forme grafiche.

Consistenza. Il *Corpus TLIO per il vocabolario* è stato costituito in modo da comprendere tutti i testi di tutte le varietà italo-romanze delle Origini, pubblicati in edizioni affidabili; non di rado i testi editi sono stati rivisti e corretti prima dell'inserimento nel corpus (lavoro documentato nelle schede filologiche consultabili in rete). Il termine ultimo era stato fissato simbolicamente alla morte di Boccaccio, 1375, ma si è poi allargato alla fine del sec. XIV.

Criteri per l'aggiornamento. Si provvede a due operazioni: la sostituzione di testi già presenti nel corpus, ma pubblicati in edizioni recenti più affidabili; l'aggiunta di testi finora esclusi dal corpus. Entrambe le operazioni comportano un lavoro notevole (verifica e sostituzione della lemmatizzazione, o lemmatizzazione ex novo; ricadute sulla redazione del TLIO), che rende impraticabile – stanti le risorse a disposizione – un aggiornamento esaustivo, sebbene l'onere sia talvolta facilitato dalla presenza di alcune delle nuove edizioni nei corpora settoriali allestiti nel frattempo dall'OVI (DiVo, LirIO). Si sono pertanto stabiliti i seguenti criteri:

> **Sostituzione**: Le edizioni più recenti di testi già presenti nel corpus sono sostituite a quelle precedenti solo dopo puntuale verifica sia della loro maggiore affidabilità in termini filologici, sia della misura in cui le modifiche rispetto all'edizione precedente ineriscono il lessico.
>
> **Aggiunta**: I nuovi inserimenti sono legati in generale alla rilevanza lessicografica. In primo luogo, si è riversata nel *Corpus TLIO per il vocabolario* la maggior parte dei testi contenuti nel *Corpus TLIO Aggiuntivo*, che era un corpus provvisorio i cui testi erano destinati a essere lemmatizzati e a entrare nel *Corpus TLIO*; esso è stato quindi disattivato. Per il resto, si sono seguiti i seguenti criteri di inclusione:
>
> - tutti i testi con datazione entro il XIII secolo;

- oltre questo limite cronologico, testi che rispondano ad almeno una di queste caratteristiche:
    - appartenenza ad aree linguistiche scarsamente documentate, escludendo cioè le seguenti: fior., pis., lucch., tosc.occ., pist., prat., sen., venez., perug. (oltre a quelle genericamente indicate come tosc., ven., sic.);
    - eccezionale rilevanza lessicale e/o culturale (per es. le *Vite dei Santi Padri* del Cavalca nella nuova ed. Delcorno, 1321-30 (tosc.occ.)); in particolare si terrà conto dei testi che contengono un lessico specifico, per es. trattati tecnici (medici, astronomici, artistici), statuti o documenti di arti o mestieri, ricettari.

L'indice di qualità TS (= "testi significativi per la documentazione della specifica varietà linguistica") è un fattore da considerare, ma non impone un'inclusione automatica (potranno non essere inclusi testi TS con localizzazione ampiamente documentata o datazione bassa: per es. Iacopo di Coluccino Bonavia, 1347-1416 (lucch.); viceversa si potranno fare eccezioni in presenza di buone ragioni, come nel caso delle scritture femminili: per es. le *Lettere* di Dora Guidalotti del Bene, 1381-92 (fior.)).

Corpus OVI dell'italiano antico.

Funzioni. Il *Corpus OVI dell'italiano antico* comprende il *Corpus TLIO per il vocabolario*, e lo estende fino a includere tendenzialmente tutti i testi pubblicati databili entro la fine del sec. XIV che non sono rientrati nei criteri selettivi per l'inclusione nel *Corpus TLIO*, in modo da consentire la ricerca su tutto il patrimonio testuale dell'italiano antico. I testi che esulano dal *Corpus TLIO* non sono lemmatizzati, per cui l'intero *Corpus OVI* è presentato senza lemmatizzazione; è tuttavia possibile utilizzare la funzione di ricerca per "lemmi muti".

Consistenza. Il *Corpus OVI* è stato fino al 2018 la somma di *Corpus TLIO* e di *Corpus TLIO Aggiuntivo* (vedi sopra): i testi dell'aggiuntivo erano destinati a essere lemmatizzati e a passare nel *Corpus TLIO*, ed erano quindi sempre in numero limitato. Nella nuova configurazione, il *Corpus OVI* intende includere un numero consistente di testi che non possono essere lemmatizzati a breve termine, ma che saranno comunque offerti alla ricerca.

Criteri per l'aggiornamento. Nel *Corpus OVI dell'italiano antico*, oltre a recepire tutti gli aggiornamenti disposti per il *Corpus TLIO*, si inseriscono i testi finora assenti che non rientrano nei criteri di inclusione nel *Corpus TLIO* (vedi sopra), quindi tendenzialmente tutti i testi editi databili entro il 1400. Motivi di ordine pratico impongono tuttavia di tener conto di alcune priorità, per cui si privilegiano le seguenti categorie: (a) i testi provenienti da aree scarsamente documentate (vedi sopra), limitandosi per le aree già ben documentate ai testi ante 1375; (b) i testi già compresi in altri corpora dell'OVI (Artesia, Datini, DiVo, LirIO); (c) i testi significativi per la loro rilevanza lessicale e/o culturale, soprattutto se scarsamente documentata nel *Corpus TLIO* (per es. la letteratura religiosa). Si eviteranno comunque i testi pubblicati in edizioni manifestamente inaffidabili.