DOI: 10.17469/O2106SLI000014

PAOLO D'ACHILLE, CLAUDIO IACOBINI

Il corpus MIDIA: concezione, realizzazione, impieghi

MIDIA è un corpus diacronico bilanciato della lingua italiana liberamente consultabile in rete che comprende testi che vanno dall'inizio del XIII alla prima metà del XX secolo per un totale di circa otto milioni di occorrenze. Si caratterizza rispetto ad altri corpora per la scansione temporale, la tipologia di testi, la lemmatizzazione. In questo intervento descriviamo la concezione e la realizzazione di MIDIA, i suoi possibili sviluppi, e illustriamo due esempi di impiego di MIDIA per ricerche diacroniche di tipo morfologico e sintattico che considerano anche la prospettiva dei generi testuali.

Parole chiave: corpus, lingua italiana, diacronia, morfologia, generi testuali.

1. *Introduzione*

Il nostro contributo è diviso in due parti: nella prima parte (§§ 2-4) presentiamo il corpus MIDIA illustrandone brevemente la concezione, la struttura attuale e le prospettive di sviluppo; nella seconda parte (§ 5) sono proposti due esempi di utilizzo dei dati ricavabili da MIDIA tratti da nostre ricerche in corso.

2. *Caratteristiche di MIDIA*

MIDIA (acronimo di Morfologia dell'Italiano in DIAcronia) è un corpus liberamente consultabile in rete a partire dal settembre 2014 all'indirizzo http://www.corpusmidia.unito.it in quanto ospitato nei server dell'Università di Torino, dapprima in un server fisico, poi, nel corso del 2021, in un servizio cloud che ha richiesto alcuni aggiornamenti al software. Abbiamo approfittato di questo passaggio per apportare anche alcune piccole modifiche, che consistono nella correzione di alcuni errori materiali, nella integrazione al corpus di alcuni testi al fine di garantire un ancora migliore bilanciamento, e qualche

integrazione alla documentazione a corredo del sito insieme a una interfaccia in lingua inglese.

In estrema sintesi, MIDIA è un corpus diacronico bilanciato della lingua italiana che comprende testi che vanno dall'inizio del XIII alla prima metà del XX secolo, per un totale di quasi otto milioni di occorrenze. È stato realizzato grazie a un finanziamento di un progetto PRIN 2009, coordinato da Paolo D'Achille, che ha visto la proficua collaborazione tra linguisti generali e storici della lingua italiana. Il corpus MIDIA presenta due peculiarità: la prima è quella di non suddividere i testi raccolti nel corpus secondo una estrinseca suddivisione in secoli (modalità che è di gran lunga quella privilegiata negli studi di italianistica, e adottata anche nel corpus LIZ poi BIZ), ma di adottare una suddivisione temporale in cinque periodi cronologici significativi per la storia della lingua italiana; la seconda, quella di operare una distinzione in generi testuali, prendendo in esame anche scritture estranee alla sfera letteraria.

I cinque periodi cronologici (1. Dall'inizio del Duecento al 1375; 2. dal 1376 al 1532; 3. dal 1533 al 1691; 4. dal 1692 al 1840; 5. dal 1841 al 1947) sono scanditi da fatti di storia linguistica, letteraria e culturale che possono essere considerati come punti di svolta nella storia della lingua italiana.

Il primo periodo (dall'inizio del Duecento al 1375) parte dallo sviluppo della letteratura (e in genere della scrittura in volgare) in area toscana fino all'anno della morte di Boccaccio e dell'inizio dell'attività cancelleresca da parte di Coluccio Salutati; la data finale è la stessa che delimita il corpus testuale dell'OVI – TLIO.

Il secondo periodo (dal 1375 al 1532) abbraccia l'esperienza dell'Umanesimo e del Rinascimento, e in particolare, per quanto riguarda la lingua, accoglie testi che si collocano tra lo sviluppo del fiorentino "argenteo" e la scelta in direzione classicista del fiorentino "aureo" teorizzata nelle *Prose della volgar lingua* di Pietro Bembo (1525). La data finale coincide con quella della terza edizione dell'*Orlando Furioso*, attuazione in poesia delle teorie bembiane.

Il terzo periodo (dal 1533 al 1691) comprende il tardo Rinascimento, il Manierismo e il Barocco. La data di chiusura coincide con la terza edizione del *Vocabolario degli Accademici della Crusca* (1691), all'indomani della fondazione dell'Arcadia (1690).

Il quarto periodo (dal 1692 al 1840) coincide con l'età dell'Arcadia, dell'Illuminismo e del Romanticismo: è questa, sostanzialmente, l'epoca in cui alcuni studiosi (Durante 1981; Tesi 2005), poco propensi a riconoscere una continuità tra italiano antico e italiano moderno, hanno collocato la nascita dell'italiano moderno. Il periodo termina con l'edizione definitiva dei *Promessi sposi*, basata, come è noto, sul fiorentino dell'uso vivo, e per tanti aspetti modello linguistico dell'italiano postunitario.

Il quinto periodo (dal 1841 al 1947) è quello in cui a partire dal Risorgimento, passando per l'Italia unita e le due guerre mondiali, si arriva alla nascita della Repubblica e alla promulgazione della Costituzione. Sull'importanza linguistica dell'unificazione hanno giustamente insistito vari studiosi, primo fra tutti Tullio De Mauro (1963), che più di recente ha valorizzato anche, sul piano linguistico, gli anni della Repubblica (De Mauro 2014). L'approdo quasi alla metà del Novecento – prima, dunque, di quest'ultimo periodo – fa sì che i testi selezionati, pur se non troppo lontani del tempo, possano fornire alcuni elementi di differenziazione in diacronia rispetto all'italiano contemporaneo, oggetto degli studi più recenti sulla formazione delle parole.

La distinzione in generi testuali è un'altra peculiarità di MIDIA. I testi del corpus sono suddivisi in sette generi: i. testi espositivi; ii. testi giuridico-amministrativi; iii. testi personali; iv. poesia; v. prosa letteraria; vi. testi scientifici; vii. teatro, oratoria, mimesi dialogica. Rileviamo anzitutto che la distinzione all'interno dei testi letterari di due generi, poesia e prosa letteraria, riprende una distinzione tradizionale (percepita come tale già nel Cinquecento ed effettivamente molto importante nella storia della lingua italiana, in cui l'istituto della poesia – come ha mostrato Serianni 2009 – ha garantito la sopravvivenza, a tutti i livelli di analisi linguistica, di tratti usciti da tempo dall'uso), mentre la letteratura teatrale (su cui negli ultimi decenni si sono intensificati gli studi, aperti da un magistrale saggio di Nencioni 1976), è stata inserita nella sezione teatro, oratoria, mimesi dialogica, che accoglie testi scritti in vista di una fruizione orale o derivati da essa (tra cui prediche, discorsi, registrazioni di verbali di processi) e altre simulazioni di dialogo (quali i manuali di conversazione), al fine di cogliere, per quanto possibile, fenomeni rappresentativi della modalità parlata. Le altre quattro sezioni rappresentano le maggiori novità del

corpus. Una è costituita da testi personali (lettere, autobiografie, diari, memorie, libri di conti) in genere non destinati alla pubblicazione e che, dato il loro carattere privato, possono aprire finestre su aspetti della lingua d'uso specialmente in ambito familiare che molte persone colte utilizzavano nei secoli passati accanto al dialetto. La sezione dei testi espositivi comprende trattati, saggi, descrizioni, biografie e altre opere non rientranti nella categoria della prosa d'arte e disponibili ad accogliere tecnicismi e voci di matrice locale. Ancora maggiore, ovviamente, è la quantità di tecnicismi che si possono trovare nella sezione dei testi scientifici, che comprende soprattutto opere che hanno per oggetto le cosiddette scienze dure: la matematica, la fisica, la biologia, la chimica, la medicina. Vale la pena di precisare che in questo caso per i periodi più recenti abbiamo raccolto anche testi di discipline quali la statistica e la psicologia, mentre specialmente per i primi due periodi temporali il corpus accoglie opere di alchimia, bestiari, volgarizzamenti di trattati scientifici classici e altre opere di simile contenuto. Infine, la sezione dei testi giuridico-amministrativi raccoglie leggi, statuti, regolamenti, atti amministrativi, che rappresentano, accanto ai trattati scientifici, i testi definiti "molto vincolanti" nella tipologia testuale di Francesco Sabatini (1990).

3. *L'annotazione del corpus*

Un corpus così differenziato per generi testuali ed esteso nel tempo ha richiesto l'impiego di un programma di annotazione automatica che fosse adatto a trattare un corpus con tali caratteristiche. I criteri di annotazione e di lemmatizzazione, così come il programma di interrogazione e l'interfaccia web, sono stati concepiti e realizzati dall'unità di ricerca dell'Università di Salerno coordinata da Claudio Iacobini, che si è avvalsa del prezioso contributo di diversi collaboratori, tra i quali in particolare Giovanna Schirato per la parte linguistica e Aurelio De Rosa, attualmente Software Engineering Manager presso Facebook a Londra, per la parte informatica.

L'interfaccia di interrogazione è articolata e flessibile: permette infatti la combinazione di diversi criteri, tra cui la parte del discorso, la forma (o parti di essa), il lemma, il contesto, il periodo temporale, il genere testuale, l'autore del testo. I risultati, oltre che essere visualiz-

zati, possono essere scaricati in formato .csv; è possibile fare ricerche anche tramite espressioni regolari.

Il ricco polimorfismo e la stratificazione lessicale della lingua italiana nelle sue diverse fasi cronologiche hanno rappresentato uno dei principali problemi da affrontare al momento di scegliere quale programma di annotazione automatica utilizzare e come adattarlo al meglio alle nostre esigenze. Infatti, i programmi di annotazione automatica disponibili mirano nella quasi totalità all'analisi dell'attuale stato sincronico di lingua (o al più a un determinato stato sincronico), oppure a una determinata modalità (ad esempio lingua parlata, invece che scritta) oppure a uno specifico genere testuale. Una delle caratteristiche distintive di MIDIA, cioè quella di rendere disponibili e confrontabili in un corpus bilanciato testi di diverse epoche e generi, è stata dunque anche la causa di maggiore difficoltà nella etichettatura e lemmatizzazione del corpus.

Per poter fornire un'analisi automatica quanto più possibile adeguata alle variazioni dipendenti dai diversi generi linguistici e dall'estensione temporale dei testi presenti nel corpus, la soluzione più efficiente si è dimostrata essere l'utilizzo del programma TreeTagger combinato con un formario che modifica e integra quello realizzato da Marco Baroni (consultabile all'indirizzo https://docs.sslmit.unibo.it/doku.php?id=resources:morph-it) per la lingua italiana contemporanea[1]. Il nostro formario, che attualmente conta circa 550.000 forme associate ciascuna a un lemma o a una o più parti del discorso, è stato arricchito di forme tratte da testi appartenenti ai diversi generi compresi nel corpus in tutta la sua estensione temporale. L'arricchimento del formario si è rivelato una strategia particolarmente efficace per migliorare significativamente gli errori di assegnazione di parte del discorso o di lemmatizzazione inevitabili per un programma tarato su testi dell'italiano contemporaneo.

Il formario da noi raccolto si può considerare un prodotto secondario del lavoro di realizzazione di MIDIA e, oltre che indispensabile per l'analisi automatica del nostro corpus, può essere anche riutilizzabile per altri lavori di etichettatura di corpora che riguardino periodi dell'italiano precedenti a quello contemporaneo. Speriamo di riuscire

[1] Ringraziamo ancora una volta Marco Baroni per il prezioso e amichevole aiuto fornitoci nelle fasi iniziali del nostro lavoro.

ad accrescerlo ulteriormente e di raffinarlo grazie alla partecipazione a nuovi progetti ispirati o collegati a MIDIA.

L'impegno profuso nel miglioramento dello strumento di annotazione automatica e gli interventi semi-automatici successivi miranti a correggere errori sistematici risultati dal processo di annotazione automatica precedenti alla messa on-line di MIDIA non hanno potuto ovviamente impedire che il corpus oggi consultabile presenti errori di annotazione, alcuni dei quali sono stati del resto da noi stessi individuati nel corso delle nostre ricerche. Al momento non ci è stato possibile correggere i singoli errori di annotazione individuati, perché ciò comporterebbe onerosi interventi manuali che sono al di là delle nostre possibilità. Chi utilizza MIDIA deve dunque in ogni caso effettuare un controllo dei dati ottenuti.

4. *Finalità, limiti e possibili sviluppi*

Il fine per cui è stato realizzato MIDIA è quello di costituire la base per un progetto che speriamo prima o poi di portare a compimento: la pubblicazione di un testo di riferimento per lo studio della formazione delle parole dell'italiano in prospettiva diacronica che completi e si integri con quello curato da Grossmann & Rainer (2004) per la formazione delle parole dell'italiano in prospettiva sincronica.

Gli studi che finora hanno utilizzato come base di dati MIDIA, in effetti, hanno riguardato principalmente la morfologia derivazionale (tra i contributi al riguardo segnaliamo D'Achille & Grossmann 2016 su *-(t)ore* e *-trice*), ma diversi hanno affrontato anche altri ambiti, quali lo studio della morfologia flessiva, del lessico e delle costruzioni. Del resto, già in occasione del convegno di chiusura del PRIN (2014), come documentano gli atti editi a cura di D'Achille & Grossmann (2017), alcune delle diverse possibilità di uso del corpus erano state indicate; da allora, ovviamente, il quadro si è molto arricchito e non è irrilevante il numero dei lavori, apparsi sia in Italia sia soprattutto all'estero, di cui abbiamo avuto notizia, che a MIDIA hanno attinto.

Siamo consapevoli del limite quantitativo rappresentato dalle dimensioni del corpus MIDIA e proprio per questo, anche in vista dell'obiettivo iniziale di un ampio studio diacronico della formazione delle parole in italiano, avevamo presentato nel 2016 un nuovo progetto PRIN, che avrebbe consentito di allargare il corpus, cosa più

facile rispetto agli anni iniziali del progetto, dato che negli ultimi anni la disponibilità di testi in rete è diventata incomparabilmente maggiore di allora, anche per epoche e generi testuali che erano risultati allora abbastanza sguarniti e il cui riempimento ha costituito un notevole sforzo al fine della acquisizione in formato digitale di testi pubblicati solo a stampa. Il nuovo progetto prevedeva anche l'inserimento di testi dialettali, in modo da ampliare la riflessione sulla formazione delle parole all'intero dominio italo-romanzo, sia in sincronia sia in diacronia. Purtroppo, però, il nuovo progetto – pur essendo stato valutato positivamente – non riuscì a entrare nella rosa dei progetti finanziati.

Nonostante questa delusione, possiamo dichiararci piuttosto soddisfatti dell'accoglienza che l'impianto del progetto e i risultati ricavabili da MIDIA hanno avuto presso la comunità scientifica, in particolare all'estero. Possiamo anticipare che siamo stati coinvolti in due iniziative in corso che hanno MIDIA come riferimento importante, l'annotazione del corpus CODIT (*Corpus diacronico dell'italiano*) e il progetto CoRaLHis (*Comparison of Romance Languages through History*). Nella breve presentazione di questi due progetti sarà evidente la loro stretta correlazione con MIDIA e il ruolo che MIDIA ha svolto nella loro ideazione e realizzazione.

Il corpus CODIT[2], ideato e realizzato da Maria Silvia Micheli, è un corpus diacronico bilanciato di italiano scritto di circa 33 milioni di token (che evidentemente costituisce un significativo incremento quantitativo rispetto a MIDIA). La periodizzazione di CODIT riprende quella del corpus MIDIA e anche la tipologia dei testi è largamente coincidente. Il corpus CODIT è attualmente consultabile attraverso il portale dell'Istituto del corpus nazionale ceco http://www.korpus.cz collegato alla Università Carolina di Praga. Al momento si possono interrogare solo testi grezzi (non annotati) a partire dal seguente link: https://www.korpus.cz/kontext/query?corpname=codit. Il lavoro appena intrapreso di lemmatizzazione e annotazione in parti del discorso (promosso da Maria Silvia Micheli insieme a Jan Radimský) utilizza il formario raccolto per MIDIA. Si tratta evidentemente di una interazione fruttuosa, in primo luogo per l'etichettatura di CODIT, ma anche per l'ulteriore futuro arricchimento del formario MIDIA, che potrà incorporare le forme ricavate dal corpus CODIT.

[2] Su cui si possono avere maggiori informazioni dal sito https://wiki.korpus.cz/doku.php/en:cnk:codit.

CoRaLHis (acronimo di *Comparison of Romance Languages through History*) è un progetto già approvato e finanziato, ma non ancora partito, che ha come sede principale l'Università Sorbona di Parigi ed è coordinato da Anne Carlier e da Elisabeth Stark dell'Università di Zurigo. Il progetto ha lo scopo di realizzare una risorsa digitale multilingue ad accesso aperto che permetta ricerche empiriche, comparative e diacroniche basate su di un corpus di testi dal XIII al XVIII secolo comparabili per epoca e per generi testuali per le tre principali sotto-aree della Romania: area iberica, gallica e italica. Anche per questo progetto, le scelte adottate per MIDIA hanno fornito un riferimento utile, che va oltre lo studio della lingua italiana e l'ambito della formazione delle parole. L'integrazione del corpus MIDIA in CoRaLHis offrirà possibilità comparative tra le varietà romanze che arricchiranno notevolmente le possibilità di ricerca già offerte da MIDIA. Per rendere possibile la comparazione sarà necessaria una piccola integrazione ai metadati di MIDIA. Le indicazioni temporali dei testi dovranno infatti essere fornite non solo in riferimento alla periodizzazione basata su eventi rilevanti per la storia della lingua italiana, ma sarà necessario indicare il secolo o partizioni temporali più precise a cui ricondurre i testi. La modularità di MIDIA rende agevolmente possibile questa integrazione, che può essere facilmente ricavata dalle date di redazione o pubblicazione di ciascun testo che sono indicate nelle schede che lo accompagnano.

Crediamo quindi che MIDIA abbia dimostrato di poter contribuire con la propria architettura allo sviluppo di progetti importanti non solo per la storia della lingua italiana, ma anche per una visione storico-comparativa delle lingue romanze.

5. *Due esempi di ricerche basate su MIDIA*

Dopo aver delineato le caratteristiche essenziali di MIDIA, intendiamo qui di seguito contribuire a mettere in luce le possibilità offerte da MIDIA illustrando i risultati di due ricerche condotte specificamente per quest'occasione (oltre a quelle documentate nei contributi raccolti in D'Achille & Grossmann 2017 e in altri lavori posteriori): una di carattere propriamente morfologico e una di tipo sintattico, i cui dati quantitativi fanno riferimento alla versione di MIDIA non ancora aggiornata.

5.1 Il prefisso negativo *anti-*

Sia pur nei limiti della sua estensione quantitativa, il corpus MIDIA permette di ricavare utili indicazioni sulla formazione delle parole dell'italiano in prospettiva diacronica, quali ad esempio la diffusione nell'uso del prefisso negativo *anti-*. Da un impiego inizialmente ristretto a termini di ambito religioso nei soli testi di tipo espositivo (a parte una sporadica attestazione in un testo poetico), il prefisso ha avuto una lenta ma progressiva espansione nel numero di formazioni e ambiti semantici nei periodi successivi, fino ad arrivare nei testi dell'ultimo periodo a coprire tutti i generi testuali grazie a un numero di derivati significativamente maggiore rispetto a tutte le epoche precedenti.

Le parole derivate con *anti-* non sono presenti nei testi del primo periodo (dall'inizio del Duecento al 1375) con la sola eccezione di *anticristo*, parola di origine tardolatina, derivata a sua volta dal greco, attestata in un testo poetico, le *Rime* di Cecco Angiolieri. Nel secondo periodo (dal 1376 al 1532) le parole attestate sono ancora molto poche (*anticristiano*, *antipapa*) e solo in testi di tipo espositivo consistenti in cronache, storie e commenti di ambito religioso (la *Storia di fra' Michele Minorita*, le *Croniche* di Giovanni Sercambi, *Le Sposizioni di Vangeli* di Franco Sacchetti). Nel terzo periodo (dal 1533 al 1691), oltre a testi di tipo espositivo (si veda il termine di origine aristotelica *antiparistasi*), il prefisso è presente anche in parole usate in testi scientifici (*anticopernicano*). Nel quarto periodo (dal 1692 al 1840) i derivati con *anti-*, oltre che in testi scientifici ed espositivi, si trovano anche in testi giuridici (*antisociale*) e letterari (*anti-geometrico*). Solo nel periodo più recente (dal 1841 al 1947) si assiste a una decisa affermazione del prefisso, i cui derivati sono presenti in tutti i generi testuali del nostro corpus, e raggiungono un numero di formazioni decisamente più alto rispetto a quelle di tutti i secoli precedenti, oltre che una maggiore varietà di ambiti semantici (*anti-astensionista*, *anti-individualistico*, *anti-progressivo*, *antiborghese*, *anticonservatore*, *anticostituzionale*, *antiguerresco*, *antiintellettualismo*, *antimilitarista*, *antimodernismo*, *antipatriotta*, *antiré*, *antiscientifico*). Grazie ai dati ricavabili da MIDIA, viene confermato e documentato l'iter cronologico della diffusione dei prefissi nominali e aggettivali delineato in Iacobini (2019), ed è anche possibile fare un raffronto con i dati relativi all'impiego del prefisso *anti-* in francese e spagnolo pubblica-

ti da Martín García (1996), Fradin (1997), Montero Curiel (1998), Huertas Martínez (2015).

5.2 La perifrasi verbale *stare* + gerundio

Per quello che riguarda la sintassi, abbiamo preso invece in considerazione le occorrenze della perifrasi verbale *stare* + gerundio, che, come è noto, è diffusissima nell'italiano contemporaneo e che si è progressivamente estesa anche a verbi che già sul piano semantico indicano un'azione durativa. I dati di MIDIA, che, come si è detto, sono relativi a fasi precedenti agli sviluppi più recenti dell'italiano, sono riportati nella Tabella 1, che distribuisce le occorrenze in periodi (I-V) e generi testuali (così abbreviati: Po(esia), Pr(osa letteraria), T(eatro, ecc.), (Testi) E(spositivi), (Testi) S(cientifici), (Testi) G(iuridici), (Testi) Pe(rsonali):

Tabella 1 - *Le occorrenze di* stare + *gerundio in MIDIA*

	Po	*Pr*	*T*	*E*	*S*	*G*	*Pe*	TOT.
I	2	1	1	-	-	1	1	**5**
II	7	-	3	-	-	-	1	**11**
III	2	5	6	1	1	1	27	**43**
IV	11	12	5	2	2	-	29	**61**
V	5	18	31	9	7	1	17	**88**
TOT.	**27**	**36**	**46**	**12**	**10**	**3**	**74**	**208**

Dalla tabella risulta chiaramente come, sebbene la struttura sia documentata *ab antiquo*, le attestazioni nei primi periodi siano rare, specie in certi generi testuali, in cui mancano del tutto. La diffusione si ha a partire dal Periodo III, soprattutto nei testi personali, e poi più decisamente nel Periodo V, in cui la perifrasi è documentata in tutti i generi testuali, anche nella prosa letteraria e soprattutto nel teatro, più aperto all'oralità. Crediamo che i dati, pur nella forma bruta con cui sono stati presentati, possano apportare un contributo al tema, molto dibattuto soprattutto negli ultimi anni, della continuità tra italiano del passato e italiano contemporaneo: indicano infatti l'importanza della prospettiva di analisi quantitativa, che guarda alla frequenza di forme e strutture, accanto a quella qualitativa, finora privilegiata, che polarizza l'analisi in base al criterio della presenza/assenza.

A rafforzare l'ipotesi di un mutamento quantitativo avvenuto di recente in italiano per quanto riguarda la struttura in esame, aiuta

il confronto con altre perifrasi, anzitutto con *stare a* + infinito, che, come è noto, oggi o assume un carattere regionale, caratterizzando (con l'infinito apocopato) le varietà romana, laziale e abruzzese, oppure è standard, ma si usa solo in contesti in cui è impossibile il costrutto alternativo (in frase negativa, con verbi di percezione, ecc.; cfr. D'Achille & Giovanardi 1998, a cui si rimanda per una prima informazione bibliografica anche sulle altre perifrasi segnalate).

I dati forniti da MIDIA sono quelli riportati nella Tabella 2.

Tabella 2 - *Le occorrenze di* stare a + *infinito in MIDIA*

	Po	*Pr*	*T*	*E*	*S*	*G*	*Pe*	*TOT.*
I	7	3	-	-	5	8	-	**23**
II	8	9	7	11	2	6	2	**45**
III	4	10	11	-	3	2	10	**40**
IV	7	6	8	-	1	-	2	**24**
V	1	15	34	3	-	-	7	**60**
TOT.	**27**	**43**	**60**	**14**	**11**	**16**	**21**	**192**

Se il numero complessivo delle occorrenze è di poco inferiore a quello di *stare* + gerundio, la distribuzione sul piano cronologico e dei generi testuali è diversa: c'è un maggior numero di assenze, ma nel complesso si vede come la struttura, che prevaleva sulla concorrente nei primi due periodi, abbia ceduto il campo a partire dal Periodo III e più consistentemente nel Periodo IV e nel V, in cui tuttavia è ancora ben attestata, anche in questo caso soprattutto nella prosa letteraria e nel teatro, limitatamente ai contesti sintattici sopra richiamati.

Per completezza, a titolo di confronto, riportiamo nella Tabella 3 i dati relativi alla perifrasi "imminenziale" (che, normalmente, non è in concorrenza con le due precedenti) *stare per* + infinito.

Tabella 3 - *Le occorrenze di* stare per + *infinito in MIDIA*

	Po	*Pr*	*T*	*E*	*S*	*G*	*Pe*	*TOT.*
I	-	2	-	-	-	8	-	**10**
II	-	2	-	1	-	6	2	**11**
III	2	2	1	1	-	2	10	**18**
IV	2	7	-	2	-	-	2	**13**
V	1	15	8	5	4	-	7	**40**
TOT.	**5**	**28**	**9**	**9**	**4**	**16**	**21**	**92**

Anche questa perifrasi ha una continuità di presenze (se pure con un numero di assenze ancora maggiore, in particolare nei testi scientifici) fin dal Periodo I, ma il numero delle occorrenze risulta sempre inferiore a 10, tranne che nei testi personali del Periodo III e nella prosa letteraria del Periodo V, in cui dunque tutte e tre le perifrasi sono ben rappresentate, perché certamente funzionali al genere narrativo (largamente prevalente, in questo periodo, nei testi del corpus).

Tornando ai dati di *stare* + gerundio, è interessante anche il confronto con due perifrasi strutturalmente analoghe – anche se aspettualmente diverse (Squartini 1990; Brianti 1992; Dessì Schmid 2021) – che in un certo senso si possono considerare "alternative", *venire* + gerundio e *andare* + gerundio. Nel primo caso, in verità, come risulta dalla Tabella 4, i dati non risultano particolarmente significativi.

Tabella 4 - *Le occorrenze di* venire + *gerundio in MIDIA*

	Po	*Pr*	*T*	*E*	*S*	*G*	*Pe*	TOT.
I	10	8	-	3	6	1	-	**28**
II	10	9	5	1	4	-	2	**31**
III	7	2	6	2	7	1	4	**29**
IV	7	12	4	7	1	-	2	**33**
V	11	2	1	25	13	1	2	**55**
TOT.	**45**	**33**	**16**	**38**	**31**	**3**	**10**	**176**

Il numero delle occorrenze è più o meno equivalente a quello di *stare* + gerundio, ma risulta distribuito più equamente nei vari periodi e generi testuali, con scarse presenze tra i testi personali e con un incremento, nel Periodo V, più nei testi espositivi e scientifici che non nella prosa letteraria.

Quanto ad *andare* + gerundio, come risulta dalla Tabella 5, ha un numero di occorrenze significativamente più alto rispetto a tutte le altre perifrasi finora esaminate.

Tabella 5 - *Le occorrenze di* andare + *gerundio in MIDIA*

	Po	*Pr*	*T*	*E*	*S*	*G*	*Pe*	*TOT.*
I	58	33	6	22	5	15	-	**139**
II	78	40	24	57	23	8	23	**253**
III	40	58	61	60	152	19	81	**471**
IV	31	24	32	34	23	2	56	**202**
V	19	28	21	17	25	-	24	**134**
TOT.	**226**	**183**	**144**	**190**	**228**	**44**	**184**	**1199**

La struttura risulta ben diffusa anche nel Periodo I (dove manca solo nei testi personali) e ha una particolare espansione nel Periodo III, quello di costituzione della norma, con un numero esorbitante di presenze (che merita un esame specifico) nella prosa scientifica. Anche nel Periodo V il numero delle occorrenze, pur se un po' in calo, resta notevole e anzi, sia sul piano complessivo, sia nei singoli generi – a parte quelli giuridici, in cui, stranamente, manca, e il teatro, dove però le occorrenze non sono affatto trascurabili –, supera quello di *stare* + gerundio.

In questo caso, dunque, il "mutamento" quantitativo dell'italiano, che vede oggi *andare* + gerundio in declino probabilmente a causa dell'avanzata di *stare* + gerundio, è avvenuto oltre i limiti cronologici del nostro corpus, che tuttavia rivela già questa tendenza.

6. *Conclusioni*

In definitiva, pensiamo che MIDIA possa offrirsi come uno strumento utile, da usare soprattutto in combinazione con altri (corpora più ampi, che non offrono però le stesse possibilità di ricerche, e opere lessicografiche in rete), per indagini a diversi livelli di analisi linguistica, che possono fornire sia dati precisi su fatti specifici, sia motivi di riflessione su questioni di carattere più generale. Per questo abbiamo voluto che fosse presente in questo convegno dedicato ai corpora.

Riferimenti bibliografici

BIZ = *Biblioteca italiana Zanichelli*. Bologna: Zanichelli, 2010, DVD.

Brianti, Giovanna. 1992. *Périphrases aspectuelles de l'italien. Le cas de* andare*,* venire *et* stare + *gérondif*. Bern: Peter Lang.

D'Achille, Paolo & Giovanardi, Claudio. 1998. Conservazione e innovazione nella sintassi verbale dal romanesco del Belli al romanaccio contemporaneo. *Dal Belli ar Cipolla. Conservazione e innovazione nel romanesco contemporaneo*, 43–65. Roma: Carocci, 2001.

D'Achille, Paolo & Grossmann, Maria. 2016. I suffissati in *-(t)ore* e *-trice* nell'italiano del periodo 1841-1947. In Ruffino, Giovanni & Castiglione, Marina (a cura di), *La lingua variabile nei testi letterari, artistici e funzionali contemporanei. Analisi, interpretazione, traduzione. Atti del XIII Congresso della SILFI (Palermo, 22-24 settembre 2014)*, 787–805. Firenze: Franco Cesati.

D'Achille, Paolo & Grossmann, Maria (a cura di). 2017. *Per la storia della formazione delle parole in italiano: un nuovo corpus in rete (MIDIA) e nuove prospettive di studio*. Firenze: Franco Cesati.

De Mauro, Tullio. 1963. *Storia linguistica dell'Italia unita*. Bari: Laterza.

De Mauro, Tullio. 2014. *Storia linguistica dell'Italia repubblicana dal 1946 ai nostri giorni*. Roma-Bari: Laterza.

Dessì Schmid, Sarah. 2021. Zur Beziehung von progressiven Verbalperiphrasen und *states*. Ein erster Bericht aus Studien zu romanischen Sprachen. *Romanistisches Jahrbuch* 72(1). 31–62.

Durante, Marcello. 1981. *Dal latino all'italiano moderno. Saggio di storia linguistica e culturale*. Bologna: Zanichelli.

Fradin, Bernard. 1997. Une préfixation complexe: le cas de *anti-*. *Neuphilologische Mitteilungen* 98(4). 333–349.

Grossmann, Maria & Rainer, Franz (a cura di). 2004. *La formazione delle parole in italiano*. Tübingen: Niemeyer.

Huertas Martínez, Sheila. 2015. Aspectos de la formación de palabras en *anti-* en el español del siglo XIX. *Études Romanes de Brno* 36(1). 41–60.

Iacobini, Claudio. 2019. "Rapiéçages faits avec sa propre étoffe": Discontinuity and convergence in Romance prefixation. *Word Structure* 12(2). 176–207.

LIZ = LIZ 4.0. *Letteratura italiana Zanichelli*. CD-ROM dei testi della letteratura italiana. 4ª ed. per Windows. Bologna: Zanichelli, 2001.

Martín García, Josefa. 1996. Los valores semánticos y conceptuales de los prefijos *anti-* y *contra-* en español. *Cuadernos de Lingüística* 4. 133–150.

Montero Curiel, María Luisa. 1998. La evolución del prefijo *anti-*. In García Turza, Claudio & González Bachiller, Fabián & Mangado Martínez, José Javier (a cura di), *Actas del IV Congreso Internacional de Historia de la Lengua Española*, vol. 3, 321–328. La Rioja: Universidad de La Rioja.

Nencioni, Giovanni. 1976. Parlato-parlato, parlato-scritto, parlato-recitato. *Di scritto e di parlato. Discorsi linguistici*, 126–179. Bologna: Zanichelli, 1981.

OVI = Istituto Opera del Vocabolario Italiano, *Corpus OVI dell'Italiano antico*, http://gattoweb.ovi.cnr.it/.

Sabatini, Francesco. 1990. Analisi del linguaggio giuridico. Il testo normativo in una tipologia generale dei testi. *L'italiano nel mondo moderno. Saggi scelti dal 1968 al 2009*, vol. 2, 273–320. Napoli: Liguori, 2011.

Serianni, Luca. 2009. *La lingua poetica italiana. Grammatica e testi*. Roma: Carocci.

Squartini, Mario. 1990. Contributo per la caratterizzazione aspettuale delle perifrasi italiane *andare* + gerundio, *stare* + gerundio, *venire* + gerundio. *Studi e saggi linguistici* 30. 117–212.

Tesi, Riccardo. 2005. *Storia dell'italiano. La lingua moderna e contemporanea*. Bologna: Zanichelli.

TLIO = *Tesoro della Lingua Italiana delle Origini*, http://tlio.ovi.cnr.it/TLIO/.