DOI: 10.17469/O2106SLI000020

ANGELA FERRARI, LETIZIA LALA, FILIPPO PECORARI

La punteggiatura italiana attraverso i corpora. Teoria, sincronia e diacronia

Tra il 2015 e il 2020 sono stati attivi all'Università di Basilea due progetti di ricerca dedicati allo studio della punteggiatura italiana in prospettiva sincronica e diacronica. Le ricerche si sono avvalse di una metodologia *corpus-based*, fondata sull'analisi di dati estratti da corpora elettronici e raccolte di testi non annotate. L'articolo riassume dapprima i risultati principali dei due progetti, mettendo in evidenza il ruolo fondamentale svolto dai corpora nell'elaborazione di una teoria della punteggiatura e nella descrizione dell'evoluzione del sistema interpuntivo nel tempo. In un secondo momento, è proposta una riflessione metodologica su alcuni limiti che i corpora elettronici pongono all'analisi della punteggiatura: limiti connessi principalmente al trattamento di segni specifici (lineetta, punto a capo) e all'impossibilità di accedere a testi interi per l'analisi di fenomeni che coinvolgono ampie porzioni testuali.

Parole chiave: punteggiatura, punteggiatura comunicativa, linguistica dei corpora, linguistica del testo.

1. *Introduzione*[1]

Tra il 2015 e il 2020, abbiamo sviluppato due progetti[2], finanziati dal Fondo Nazionale Svizzero per la Ricerca Scientifica, incentrati sulla punteggiatura italiana, che abbiamo studiato dapprima in prospettiva sincronica[3] e poi in prospettiva diacronica[4]. Per queste ricerche ci

[1] La stesura del contributo è dovuta a Letizia Lala per i §§ 1-2 e a Filippo Pecorari per il § 3. Angela Ferrari ha diretto entrambi i progetti menzionati.

[2] Ai progetti hanno partecipato: Angela Ferrari (direttrice), Letizia Lala, Fiammetta Longo, Filippo Pecorari, Benedetta Rosi, Roska Stojmenova.

[3] *PUNT-IT – Le funzioni informativo-testuali della punteggiatura nell'italiano contemporaneo, tra sintassi e prosodia* (100012_156119, febb. 2015 – genn. 2018).

[4] *PUNT-IT2 – La punteggiatura italiana in prospettiva diacronica: dallo standard al neo-standard, e dal Cinquecento al Novecento* (100012_175741, febb. 2018 – genn. 2020).

siamo avvalsi di una metodologia *corpus-based*, analizzando raccolte di grandi quantità di dati, consultabili e confrontabili. Per il nostro tipo di analisi, incentrata sugli usi interpuntivi, abbiamo potuto utilizzare corpora elettronici annotati, disponibili alla consultazione online (e.g. CORIS, DiaCORIS, CONTRAST-IT), e raccolte di testi non annotate (e.g. PUNT-IT, Web2Corpus_it, Narrativa anni 2000, Corpus La Stampa).

L'applicazione della *corpus analysis* ci ha consentito di individuare fenomeni specifici ed è stata fondamentale per arrivare alle generalizzazioni necessarie per elaborare una teoria e tratteggiare le evoluzioni del sistema su larga scala. Ha però mostrato anche alcune criticità che ci proponiamo di commentare (§ 3) una volta illustrati i risultati delle nostre ricerche (§ 2).

2. *Risultati delle ricerche*

2.1 La prima fase: lo studio teorico e sincronico

Nella prima fase della ricerca, dedicata all'osservazione della punteggiatura nell'italiano contemporaneo, abbiamo passato in rassegna l'intero sistema interpuntivo. Scostandoci dalla vulgata grammaticale e saggistica, abbiamo mostrato come le tradizionali interpretazioni della punteggiatura approntate in termini sintattici e/o prosodici siano fuorvianti dal punto di vista teorico e in ogni caso concretamente inapplicabili, e come l'uso contemporaneo dell'interpunzione in italiano non possa che definirsi in termini comunicativo-testuali[5]. Più precisamente, la punteggiatura svolge nella scrittura contemporanea due funzioni, che possono anche intersecarsi:

a. segmenta il testo nelle sue unità comunicative costitutive e (eventualmente) le gerarchizza (è il caso del punto, della virgola, del punto e virgola, dei due punti, della lineetta, delle parentesi);
b. introduce nel testo valori interattivi: inferenze (come per i puntini di sospensione), atteggiamenti illocutivi (come per il punto interrogativo e il punto esclamativo), piani polifonici (come per le virgolette).

[5] La definizione del valore comunicativo-testuale di fondo dei segni interpuntivi è stata proposta nel volume collettivo Ferrari *et al.* (2018).

Le nostre ricerche hanno mostrato che la concezione diffusa che la punteggiatura segnali snodi sintattici o che indichi le curve intonative e prosodiche di una realizzazione orale è inadeguata non solo in base a questioni teoriche (in particolare una concezione strutturale *vs*. funzionale della lingua), ma anche in quanto solo un'interpretazione comunicativo-testuale riesce davvero a rendere conto degli usi interpuntivi nell'italiano contemporaneo. In effetti, ogni tentativo di trattare la punteggiatura in chiave sintattica ha mostrato grossi limiti, finendo per scomporsi in distinzioni e sotto-distinzioni, e contraddicendosi con una massa di eccezioni ed eccezioni alle eccezioni (Ferrari & Lala 2011 e 2013). Anche gli usi che la tradizione più stabilmente riconduce alla sintassi mostrano in realtà regolarità di carattere testuale. Si pensi al punto, a cui tradizionalmente si attribuisce la funzione di chiudere una frase sintattica, ma che in realtà segnala la chiusura di un'unità testuale (l'Enunciato) autonoma da un punto di vista illocutivo-testuale, che non è affatto obbligatorio che abbia natura sintattica frasale, come mostrano impieghi come i seguenti[6]:

(1) Ho conosciuto tua moglie. **Affascinante.**
Oggi ho guidato io. **Fino a Firenze!**

Anche la virgola ha una funzione di natura testuale, avendo il ruolo di scandire l'Enunciato al suo interno in sotto-unità (Unità Informative) la cui natura è, ancora una volta, testuale e non sintattica:

(2) Si avvicinò**, lentamente,** e sorrise.
Mi è parso triste, **e arrabbiato**.

Abbiamo studiato e chiarito anche la relazione tra punteggiatura e prosodia di lettura. Ne è emerso che, se esiste in effetti una relazione tra presenza di un segno e intonazione di lettura, essa è però indiretta, sotto-specificata e parziale. Indiretta, perché mediata dai valori comunicativi ad esso associati; sotto-specificata, in quanto l'unità delimitata da un segno può ricevere profili intonativi diversi in base alla funzione informativo-illocutiva che è chiamata a svolgere; e parziale, in quanto è determinata non solo dalla presenza del segno, ma dalla combinazione con le indicazioni date dal lessico, dalla morfosintassi e dal cotesto.

[6] Per un approfondimento del modello teorico sul quale si basano queste analisi cfr. Ferrari *et al*. (2008).

Studiando gli impieghi sui corpora analizzati siamo risaliti al valore specifico di ogni segno di punteggiatura, arrivando a definizioni, che, allontanandosi dalla vulgata, sono tutte riconducibili all'una e/o all'altra delle due funzioni generali appena proposte (cfr. *supra*) e realmente in grado di rendere conto degli impieghi nei testi.

Si prenda ad esempio il punto interrogativo, al quale viene tradizionalmente riservato un trattamento sintattico (chiuderebbe nello scritto la frase interrogativa) e/o prosodico (attribuirebbe alla sequenza che chiude un'intonazione 'interrogativa'). In realtà lo studio sui corpora ha mostrato che: (i) in moltissimi casi il punto interrogativo chiude unità che non hanno affatto una natura frasale (*Perché mai? E quindi? Giornata pesante?*); (ii) la restituzione nell'orale di unità chiuse da questo segno non indirizza verso un unico profilo intonativo, ma verso un paradigma di realizzazioni anche piuttosto diverse tra loro: *Che cosa fai?* (domanda aperta), *Che cosa fai?* (richiesta di conferma), *Che cosa fai?!* (domanda accompagnata da tono enfatico, perentorio). Scartata dunque la lettura tradizionale, il nostro studio su corpora ci ha permesso di stabilire che il vero valore del punto interrogativo è di tipo (comunicativo) interattivo: esso introduce una richiesta di reazione, linguistica o non-linguistica, rivolta all'interlocutore. In contesto monologico, l'interazione sollecitata dal segno si produce tra scrittore e lettore; in contesto dialogico, tra i partecipanti allo scambio (Lala 2018).

2.2 La seconda fase: lo studio diacronico

L'aver compreso a fondo la punteggiatura italiana contemporanea ci ha portati a interrogarci su due questioni teoriche importanti di orientamento diacronico: (i) la punteggiatura italiana è sempre stata quello che è oggi? (ii) la punteggiatura italiana negli ultimi decenni è cambiata?

Adottando un'attenta metodologia *corpus-based* abbiamo intrapreso uno studio della punteggiatura in ottica diacronica che si è posto l'obiettivo di investire in queste due direzioni: così, da una parte abbiamo studiato la storia della punteggiatura dal Cinquecento a oggi (diacronia lunga), e dall'altra le sue evoluzioni più recenti, relative all'epoca del passaggio dell'italiano dallo standard al neostandard (diacronia breve).

2.2.1 Diacronia lunga

Per il primo ambito, di diacronia lunga, si è trattato di delineare e spiegare la storia della punteggiatura italiana, della sua concezione e dei suoi impieghi, dal Cinquecento fino al Novecento.

La ricerca è stata svolta basandosi sulle più importanti grammatiche e su un ampio corpus di scritture rappresentative. Abbiamo studiato il sistema interpuntivo in generale e ogni singolo segno di punteggiatura. I risultati ottenuti sono stati significativi: è emerso come l'italiano sia passato da un uso che combinava il criterio prosodico con quello morfosintattico (Cinquecento-primo Seicento), a un uso più rigorosamente morfosintattico (Seicento-secondo Settecento), infine a un uso comunicativo (stabilizzato nel secondo Ottocento, per raffinarsi sempre di più nel corso del Novecento).

Si è dunque potuto appurare come l'evoluzione nella storia della punteggiatura italiana equivalga in buona parte al passaggio da una *ratio* morfosintattica a una *ratio* comunicativo-testuale; passaggio registrato nel secondo Ottocento dalle grammatiche, anche su spinta delle scelte manzoniane.

Questo orientamento è particolarmente visibile se si osservano i mutamenti negli usi della virgola. Come si sa, la *ratio* morfosintattica prevede che la virgola compaia ogni volta che emerge un confine tra reggente e subordinata, qualunque sia la natura di quest'ultima; e, per la coordinazione, che la virgola accompagni sempre e comunque il collegamento copulativo. La *ratio* comunicativa chiede invece la virgola solo nei casi in cui la proposizione subordinata sia autonoma dal punto di vista informativo (il che la esclude nei casi in cui la subordinata è compattata semanticamente con la principale, come nelle completive post-reggente e nelle relative restrittive); mentre per la coordinazione, la regolarità comunicativa conduce a ometterla a meno che non emergano particolari necessità di ordine comunicativo, quali la disambiguazione o la focalizzazione. Così usi come quelli che vediamo negli esempi (3) e (4), improntati a regolarità di natura morfo-sintattica, perfettamente adeguati nel XVIII secolo[7]:

(3) Di questo n'ebbi una chiara pruova, **quando** mi fu concesso l'onore dalla Maestà Vostra d'ammirare la diligenza, **che**

[7] Gli esempi sono riprodotti fedelmente, senza normalizzare eventuali devianze o peculiarità grafiche.

usate nell'esaminare i minimi oggetti col Microscopio. (Della Torre 1755, in Ferrari 2020)

(4) A me resta di supplicare in fine l'A. V., **che** le piaccia nella presente offerta risguardare con occhio clemente l'estraordinaria volontà, **che** porto di corrispondere [...] all'obbligo che tengo alla Serenissima sua Casa, **ed** al contento che sento d'esserle nato, **e** di doverle morire devotissimo suddito, **e** servitore, **e** qui con ogni umiltà me le inchino. (Molza 1750, in Ferrari 2020)

hanno ceduto il passo a impieghi legati a una concezione comunicativa della punteggiatura, che è andata stabilizzandosi nella prima metà del XIX secolo, e che è all'origine dell'alternanza di presenza *vs.* assenza di virgola nelle copulative in (5):

(5) Sia dedicato a voi questo libro, dove io cercava, come si cerca spesso colla poesia, di consacrare il mio dolore, **e** col quale al presente (nè posso già dirlo senza lacrime) prendo comiato dalle lettere **e** dagli studi. Sperai / che questi cari studi avrebbero sostentata la mia vecchiezza, **e** credetti colla perdita di tutti gli altri piaceri, di tutti gli altri beni della fanciullezza **e** della gioventù, avere acquistato un bene che da nessuna forza, da nessuna sventura mi fosse tolto. Ma io non aveva appena vent'anni, quando da quella infermità di nervi **e** di viscere, che privandomi della mia vita, non mi dà speranza della morte, quel mio solo bene mi fu ridotto a meno che a mezzo; poi, due anni prima dei trenta, mi è stato tolto del tutto, **e** credo oramai per sempre. Ben sapete che queste medesime carte io non ho potute leggere, **e** per emendarle m'è convenuto servirmi degli occhi **e** della mano d'altri. (Leopardi 1831, in Ferrari 2020)

2.2.2 Diacronia breve

Per l'analisi in diacronia a breve gittata, l'obiettivo è stato quello di verificare se negli ultimi decenni, accanto alla punteggiatura standard, se ne stesse disegnando una neo-standard, nello stesso modo in cui hanno preso forma un neo-standard morfologico, sintattico e lessicale.

Sono emersi in effetti alcuni usi significativi che vanno in questa direzione: (i) il diffondersi di uno *style coupé* creato dal susseguirsi di punti che spezzano la linearità del dettato (6); (ii) la cosiddetta *virgola passepartout*, che sta espandendo il proprio ambito d'azione andando ad occupare spazi/funzioni tradizionalmente riservati a segni più forti (7); (iii) il diffondersi della lineetta singola di origine inglese (8); (iv)

il calo significativo d'impiego dei segni semanticamente ricchi, in particolare del punto esclamativo (9) e dei due punti (10); (v) il declino del punto e virgola, segno i cui impieghi rimangono frequenti solo in alcuni generi testuali conservativi (in particolare, nel linguaggio giuridico-amministrativo, con il ruolo di scandire enumerazioni) (11):

(6) Si era messa la gonna e la giacca grigia che usava quando faceva le cose importanti. Il golf girocollo. Le perle. E le scarpe blu con i tacchi alti. (Ammaniti, in Ferrari & Lala 2021: 22)

(7) Ha scelto la batteria. Dave Grohl, ex batterista dei Nirvana, approverebbe. («La Stampa», 6 luglio 2019, in Demartini 2019)

(8) L'ultima funzione, "Poke", permetteva infine di mandare a un utente solo un segnale di interesse, del tutto privo di contenuto – il destinatario riceveva un avviso che spiegava solamente che il mittente l'aveva "toccato, stuzzicato" («poked»). (Tavosanis, in Longo 2018a: 145)

(9) Franceschino. Che bello sentire la sua voce, ora. [*vs.* !] Che sollievo pensare che lui c'è davvero, non è una questione di crederci o non crederci. [*vs.* !] Che bello sentirlo partire sparato come al solito [...]. [*vs.* !] (Ferrante, in Lala 2019: 336)

(10) Il problema era mia madre, [*vs.* :] con lei le cose non andavano mai per il verso giusto. [...] Di sicuro non era felice, [*vs.* :] le fatiche di casa la logoravano e i soldi non bastavano mai. (Ferrante, in Lala 2019: 335) [*vs.* :]

(11) La Corte: riunisce i ricorsi; dichiara ammissibile ed accoglie il ricorso principale e rigetta il ricorso incidentale; dichiara la giurisdizione del giudice italiano; cassa la sentenza impugnata e rinvia la causa al Tribunale di Genova, che deciderà anche sulle spese del giudizio di legittimità. (Cass., sez. un. civ., 10-03-1998, n. 2642, in Dell'Anna 2017: 139)

La spinta verso il cambiamento e l'affermarsi di queste tendenze proviene in buona parte da scritture marcate in diafasia, come quelle letterarie degli ultimi cinquant'anni, o marcate in diamesia, come quelle che rientrano nella *Computer-Mediated Communication*, senza dimenticare l'influenza di impieghi tipici in altre lingue.

3. *Alcuni limiti delle ricerche interpuntive su corpora*

Come si è detto in §§ 1-2, tutte le ricerche qui menzionate si sono avvalse di una metodologia *corpus-based*, senza la quale non avremmo potuto elaborare una teoria della punteggiatura né tratteggiare le evoluzioni del sistema interpuntivo. L'uso dei corpora ha però fatto emergere anche alcuni limiti che questi strumenti pongono all'analisi della punteggiatura. Va precisato che i limiti osservati intaccano solo in maniera marginale l'utilità dei corpora come strumento per un'analisi della punteggiatura nei testi: come si vedrà, si tratta di limiti che riguardano una casistica limitata di segni e di fenomeni interpuntivi. Riteniamo tuttavia che sia opportuno stimolare una riflessione di carattere metodologico, anche perché a nostra conoscenza mancano studi specifici sull'uso dei corpora come strumento per indagini sulla punteggiatura: forse, dunque, può essere utile proporre qualche osservazione cercando di far interagire il punto di vista teorico sulla punteggiatura – privilegiato nelle nostre ricerche – con quello applicato alla costruzione di corpora.

Per l'analisi abbiamo ragionato retrospettivamente su un campione di corpora dell'italiano che ci sono stati utili in diverse fasi delle nostre ricerche: i corpora bolognesi CORIS (Rossini Favretti 2000) e DiaCORIS (Onelli *et al*. 2006), il PEC-Perugia Corpus (Spina 2014), il Primo Tesoro della Lingua Letteraria Italiana del Novecento (De Mauro 2007) e i *web corpora* ospitati da Sketch Engine (Jakubíček *et al*. 2013), con particolare riferimento a itTenTen.

3.1 La non-tokenizzazione dei segni di punteggiatura

Un primo limite che abbiamo riscontrato consiste nell'assenza della tokenizzazione dei segni di punteggiatura[8]. Si tratta evidentemente di un limite pressoché insormontabile per l'analisi interpuntiva: se il corpus non riconosce i segni di punteggiatura, e non può essere interrogato su di essi, la ricerca risulta impossibile.

Nel nostro campione di riferimento, l'unico corpus che presenta questo limite è il Primo Tesoro della Lingua Letteraria Italiana del Novecento, che raccoglie i testi di 100 romanzi vincitori o partecipanti al Premio Strega tra il 1947 e il 2006. A fronte della grande raf-

[8] Cfr. Lenci *et al*. 2005 (105-107) per una presa di posizione, dalla prospettiva computazionale, a favore della considerazione dei segni interpuntivi come token indipendenti.

finatezza che il corpus consente nelle analisi lessicali (ad es. sulle marche d'uso dei lessemi), esso non ammette ricerche sulla punteggiatura. Ciò risulta piuttosto sorprendente, perché tra le categorie oggetto di annotazione morfosintattica ci sono anche quelle di "ideogramma" e di "simbolo", che comprendono segni sicuramente meno interessanti per l'analisi linguistica rispetto ai segni di punteggiatura: segni come, ad esempio, <&>, <+>, <$> per la prima categoria e <[ ]>, <|> per la seconda categoria, che annovera peraltro anche due segni interpuntivi propri di lingue diverse dall'italiano, ovvero il punto esclamativo <¡> e interrogativo <¿> capovolti usati in spagnolo.

3.2 Il trattamento della lineetta

Al netto del problema appena menzionato, rilevante ma anche molto raro, gli altri limiti che abbiamo riscontrato riguardano principalmente il modo in cui sono considerati alcuni segni di punteggiatura nella tokenizzazione del corpus.

Un primo problema riguarda il trattamento della lineetta, un segno interpuntivo «alloglotto» (Longo 2020: 232) che l'italiano ha acquisito dall'inglese verso la fine del Settecento. Il problema principale che si incontra lavorando sui corpora è l'assimilazione tra i token relativi a due segni molto diversi dal punto di vista funzionale, ovvero la lineetta <–> e il trattino <->. La lineetta, secondo il punto di vista teorico, può essere considerata un segno interpuntivo a tutti gli effetti, perché contribuisce alla costruzione del messaggio testuale fornendo indicazioni sulla segmentazione del testo; il trattino, invece, è un segno paragrafematico non interpuntivo (Longo 2020: 234), che agisce a livello lessicale segnalando una relazione tra due lessemi o lavorando all'interno del lessema, come nelle parole composte. È evidentemente molto diversa la funzione che ha la lineetta in (12), una funzione simile a quella delle parentesi, rispetto alla funzione che ha il trattino in (13):

(12) A quelle parole – nette e assertive – i mercati hanno cambiato direzione. («Corriere della Sera», in Longo 2018b: 131)

(13) Salerno-Reggio Calabria, calcio-mercato, tecnico-scientifico, eco-incentivi (in Tonani 2011)

Il trattamento della lineetta e del trattino, nella maggior parte dei corpora del campione, non prevede una distinzione dei due segni su basi funzionali. Le opzioni che si riscontrano sono due:

i. Ci sono corpora, come il CORIS e il PEC, che assimilano integralmente i due simboli grafici: nel CORIS la ricerca della lineetta non dà risultati, e tutto ciò che si trova – tanto usi come in (12) quanto usi come in (13) – compare tokenizzato come trattino e richiede di essere ricercato come tale; nel PEC la ricerca del trattino e la ricerca della lineetta danno esattamente gli stessi risultati.
ii. Ci sono corpora, come il DiaCORIS e itTenTen, che distinguono sì i due simboli grafici, ma su basi esclusivamente formali e non funzionali. Questo trattamento può porre problemi alla ricerca, perché i testi non sono sempre uniformi nella realizzazione della lineetta e del trattino con due simboli distinti, e in particolare la lineetta è spesso realizzata graficamente nella forma breve tipica del trattino. Di fatto, chi volesse esaminare le occorrenze del segno interpuntivo "lineetta" sarebbe obbligato a discriminare gli esempi manualmente, in maniera non diversa da quanto impone il caso (i).

3.3 Il trattamento dell'a capo

Un altro limite relativo al trattamento di forme specifiche è quello che riguarda gli a capo. Nelle ricerche teoriche che abbiamo condotto sulla punteggiatura (cfr. Ferrari 2018), il punto a capo è classificato come un segno a sé stante rispetto al punto fermo, perché ha la funzione di delimitare un'unità testuale specifica, ovvero il Capoverso, che può contenere al suo interno più Enunciati delimitati da punti fermi o da altri segni. Più in generale, per un'analisi *corpus-based* della punteggiatura in prospettiva testuale sarebbe importante poter accedere alla scansione in capoversi del testo, non solo in relazione agli usi del punto a capo, ma anche a quelli di altri segni accompagnati dall'a capo: ad esempio, dei due punti che introducono un elenco, oppure delle lineette usate come segnale grafico dei punti di una lista.

Nei corpora abbiamo riscontrato tre opzioni relativamente al trattamento dell'a capo:

i. Ci sono corpora, come il CORIS e il DiaCORIS, in cui la scansione del testo in capoversi è assente. In questo caso, si perde la possibilità di distinguere il punto a capo dal punto fermo, o di riflettere sull'associazione tra l'a capo e altri segni interpuntivi.

ii. In altri corpora, come il PEC, si adotta una soluzione in qualche misura opposta a quella in (i): nelle finestre di contesto mostrate dall'interfaccia tutte le frasi sono isolate da un a capo, a prescindere dalla presenza effettiva di un a capo nel testo originario. Anche in questo caso, la scansione originaria in capoversi è evidentemente non ricostruibile.
iii. Un'opzione più utile all'analisi interpuntiva è quella presente in itTenTen, che marca i confini di frase e di capoverso con due tag XML specifici: <s> per la frase (*sentence*) e <p> per il capoverso (*paragraph*). Questa scelta di annotazione consente di discriminare i punti a capo dai punti fermi non solo *a posteriori*, alla lettura delle concordanze, ma anche al momento della ricerca automatica, attraverso l'impiego di stringhe apposite[9].

3.4 L'accessibilità ai testi del corpus

A conclusione di questa breve rassegna, proponiamo un'ultima osservazione che ci porta oltre il dominio della punteggiatura, a toccare più in generale i rapporti tra corpora e linguistica del testo. Alcuni tra i fenomeni interpuntivi considerati nelle nostre ricerche coinvolgono non singoli enunciati o brevi sequenze testuali, ma blocchi di testo più estesi, quando non addirittura testi interi. È per esempio il caso della virgola *passepartout* menzionata in § 2.2.2., che può essere usata in lunghe sequenze di testo. Si tratta di un espediente tipico del testo letterario, che serve principalmente a mimare una tirata di parlato senza interruzioni da parte di un personaggio, come nell'esempio seguente:

(14) Che brutta gente – attaccò a dire senza fermarsi più fino alla metropolitana di Piazza Amedeo –, hai visto la vecchia come t'ha trattata, s'è voluta vendicare, non può sopportare che Nadia, educata apposta per essere la meglio di tutte, Nadia che doveva darle tante soddisfazioni, non combina niente di buono, s'è messa col muratore e le fa la puttana sotto gli occhi: sì, non può sopportarlo, ma tu fai male a dispiacerti, fottitene, non glielo dovevi lasciare il tuo libro, non dovevi chiedere se voleva la dedica, soprattutto non gliela dovevi fare, questa è gente che bisogna trattare a calci in culo, il tuo difetto è che sei troppo buona, abbocchi a tutto ciò che dicono quelli che

[9] Per cercare i soli punti a capo, ed escludere i punti fermi, si può usare la stringa in *Corpus Query Language* [lemma=""] <p>.

> hanno studiato come se la testa ce l'avessero soltanto loro, e invece non è così, rilassati, va', sposati, fa' il viaggio di nozze, ti sei preoccupata troppo per me, scrivi un altro romanzo, lo sai che m'aspetto da te cose bellissime, ti voglio bene. (Ferrante, in Ferrari 2017: 148)

Un altro fenomeno che coinvolge ampie porzioni testuali è l'uso dei puntini di sospensione come segno esclusivo, che si sostituisce a tutti o quasi gli altri segni segmentanti. Questo uso è piuttosto comune in rete, nei testi scritti dai non professionisti della scrittura come il seguente:

(15) Mi rivolgo a tutte le Persone che non stanno facendo altro che insultare gigio e tutta la nostra famiglia.. Gigio sin da piccolo e tifoso del Milan.. per lui giocare con la maglia del Milan e' un sogno..ha sempre onorato e dato L anima per questi colori.. ha pianto per ogni sconfitta.. fino a ieri eravate tutti con gigio.. ora senza sapere nulla state insultando tutta la famiglia, scrivendo frasi che la nostra famiglia non augura nemmeno al peggior nemico.. la nostra famiglia ha gioito e pianto con tutti voi tifosi.. il Milan ha una storia incredibile.. e nessuno può metterlo in dubbio..Per le persone che hanno scritto messaggi a favore di gigio ci tengo a dire che voi avete capito davvero che persona e' gigio.. qualunque gesto che ha fatto e qualunque frase ha detto o scritto.. L ha fatto davvero per amore del Milan.. gigio e' soprattutto un tifoso del Milan.. Come voi..e chi lo insulta non è tifoso del Milan..ora potete anche riempire di insulti questa foto.. ma la famiglia lasciatela stare.. loro ci hanno sempre insegnato i veri valori della vita... per quelli che invece continuano a dire che io devo ringraziare a gigio perché mi da i soldi.. vi dico che a me mai nessuno mi ha regalato qualcosa.. ogni anno lotto per guadagnare quello che mi merito.. grazie.. (instagram.com/antodonnarumma90, 16.06.2017, in Pecorari 2019: 161-162)

Per l'analisi di fenomeni come quelli qui esemplificati, i corpora pongono al testualista il problema – ormai annoso – dell'accessibilità ai testi contenuti nel corpus. In linea di massima, la maggior parte dei corpora non consente l'accesso ai testi nella loro interezza, ma solo a porzioni limitate. Il problema, come è noto, è essenzialmente legale, e dipende dai vincoli posti dalle norme sul copyright (cfr. Allora & Barbera 2007): i dati del corpus possono essere distribuiti al pubblico

soltanto nei limiti di una finestra di contesto di poche decine o centinaia di caratteri.

La principale soluzione pratica che abbiamo adottato nelle nostre ricerche per ovviare a questo problema è stata quella di costruire raccolte di testi *ad hoc*, progettate per essere rappresentative di un tipo testuale o di una varietà linguistica, a partire da testi disponibili pubblicamente in rete: articoli pubblicati negli archivi online dei quotidiani, saggi e articoli scientifici ad accesso libero, testi normativi e amministrativi ecc. È per esempio questo il caso del corpus PUNT-IT, che è stato il nostro principale terreno di indagine della punteggiatura negli ultimi anni[10], e anche del corpus It-Ist_CH, che ci sta sostenendo da qualche tempo nello studio dell'italiano istituzionale svizzero in prospettiva testuale[11].

I corpora così costruiti hanno tendenzialmente dimensioni più ridotte dei corpora elettronici di ultima generazione (PUNT-IT, per esempio, conta circa 500.000 parole, mentre It-Ist_CH ne comprende 2.500.000), e peraltro, secondo alcune definizioni più restrittive di "corpus", non potrebbero nemmeno essere chiamati corpora in senso stretto, dal momento che i testi non sono tokenizzati e non sono addizionati di markup[12]. Tuttavia, la possibilità che queste risorse offrono di accedere agilmente e senza alcuna limitazione ai testi interi costituisce un vantaggio pratico di enorme rilevanza per l'analisi testuale, specialmente in casi particolari come gli usi interpuntivi esemplificati in (14) e (15).

[10] Il corpus PUNT-IT è stato costruito nell'ambito del progetto omonimo e contiene al suo interno testi giornalistici, saggistici e giuridico-amministrativi scritti negli ultimi trent'anni (1985-2015), rappresentativi della scrittura funzionale contemporanea di registro medio-alto.

[11] Il corpus It-Ist_CH è stato costruito nell'ambito del progetto FNS "L'italiano istituzionale svizzero: analisi, valutazioni, prospettive" attualmente in corso ed è disponibile ad accesso libero sul sito del progetto: cfr. https://sites.google.com/view/progettoitistch/corpus e la descrizione del corpus in Ferrari *et al*. (2022).

[12] Secondo Barbera *et al*. (2007), il corpus è una «[r]accolta di testi in formato elettronico uniformemente trattati (ossia *almeno tokenizzati ed addizionati di un markup adeguato*) in modo da essere gestibili ed interrogabili informaticamente» (p. 26) [corsivo nostro].

Riferimenti bibliografici

Allora, Adriano & Barbera, Manuel. 2007. Il problema legale dei corpora. Prime approssimazioni. In Barbera, Manuel & Corino, Elisa & Onesti, Cristina (a cura di), *Corpora e linguistica in rete*, 109–118. Perugia: Guerra.

Barbera, Manuel & Corino, Elisa & Onesti, Cristina. 2007. Cosa è un corpus? Per una definizione più rigorosa di corpus, token, markup. In Barbera, Manuel & Corino, Elisa & Onesti, Cristina (a cura di), *Corpora e linguistica in rete*, 25–88. Perugia: Guerra.

Dell'Anna, Maria Vittoria. 2017. Veniamo al punto. Interpunzione e dintorni nei testi giudiziari italiani. In Ferrari, Angela & Lala, Letizia & Pecorari, Filippo (a cura di), *L'interpunzione oggi (e ieri). L'italiano e altre lingue europee*, 131–146. Firenze: Cesati.

Demartini, Silvia. 2019. I punti della situazione. Viaggio nella punteggiatura dell'italiano di oggi. 3. La virgola splice. (https://www.treccani.it/magazine/lingua_italiana/articoli/scritto_e_parlato/punteggiatura3.html) (Consultato il 02.02.2022.)

De Mauro, Tullio. 2007. *Primo Tesoro della Lingua Letteraria Italiana del Novecento*. Torino: UTET/Fondazione Bellonci.

Ferrari, Angela. 2017. Usi "estesi" del punto e della virgola nella scrittura italiana contemporanea. *La lingua italiana. Storia, strutture, testi* XIII. 137–153.

Ferrari, Angela. 2018. Il punto a capo. In Ferrari, Angela & Lala, Letizia & Longo, Fiammetta & Pecorari, Filippo & Rosi, Benedetta & Stojmenova, Roska, *La punteggiatura italiana contemporanea. Un'analisi comunicativo-testuale*, 95–107. Roma: Carocci.

Ferrari, Angela. 2020. Note sull'uso della virgola ai margini della scrittura letteraria e saggistica tra Sette e Ottocento. *Margini. Giornale della dedica e altro* 14. (https://www.margini.unibas.ch/web/rivista/numero_14/saggi/articolo1/ferrari.html) (Consultato il 02.02.2022.)

Ferrari, Angela & Cignetti, Luca & De Cesare, Anna-Maria & Lala, Letizia & Mandelli, Magda & Ricci, Claudia & Roggia, Carlo Enrico. 2008. *L'interfaccia lingua-testo. Natura e funzioni dell'articolazione informativa dell'enunciato*. Alessandria: Edizioni dell'Orso.

Ferrari, Angela & De Cesare, Anna-Maria & Evangelista, Daria & Lala, Letizia & Marengo, Terry & Pecorari, Filippo & Piantanida, Giovanni & Rosi, Benedetta. 2022. Il corpus It-Ist_CH: un corpus rappresentativo dell'italiano istituzionale svizzero. In Baranzini, Laura & Casoni, Matteo & Christopher, Sabine (a cura di), *Linguisti in contatto 3. Ricerche di*

linguistica italiana in Svizzera e sulla Svizzera, 57–69. Bellinzona: Osservatorio linguistico della Svizzera italiana.

Ferrari, Angela & Lala, Letizia. 2011. Les emplois de la virgule en italien contemporain. De la perspective phono-syntaxique à la perspective textuelle. In Favriaud, Michel (a cura di), *Ponctuation(s) et architecturation du discours à l'écrit*, 53–88. Paris: Larousse/Armand Colin.

Ferrari, Angela & Lala, Letizia. 2013. La virgola nell'italiano contemporaneo. Per un approccio testuale (più) radicale. *Studi di Grammatica Italiana* XXIX-XXX. 479–501.

Ferrari, Angela & Lala, Letizia. 2021. *Interpunzioni creative. Esempi letterari degli anni Duemila*. Firenze: Cesati.

Ferrari, Angela & Lala, Letizia & Longo, Fiammetta & Pecorari, Filippo & Rosi, Benedetta & Stojmenova, Roska. 2018. *La punteggiatura italiana contemporanea. Un'analisi comunicativo-testuale*. Roma: Carocci.

Jakubíček, Miloš & Kilgarriff, Adam & Kovář, Vojtěch & Rychlý, Pavel & Suchomel, Vít. 2013. The TenTen corpus family. In *Proceeding of the 7th International Corpus Linguistics Conference CL*, 125–127.

Lala, Letizia. 2018. Il punto interrogativo. In Ferrari, Angela & Lala, Letizia & Longo, Fiammetta & Pecorari, Filippo & Rosi, Benedetta & Stojmenova, Roska, *La punteggiatura italiana contemporanea. Un'analisi comunicativo-testuale*, 183–199. Roma: Carocci.

Lala, Letizia. 2019. Sulle tendenze interpuntive nella narrativa italiana contemporanea. In Moretti, Bruno & Kunz, Aline & Natale, Silvia & Krakenberger, Etna (a cura di), *Le tendenze dell'italiano contemporaneo rivisitate. Atti del LII Congresso Internazionale di Studi della Società di Linguistica Italiana (Berna, 6-8 settembre 2018)*, 323–341. Milano: Officinaventuno.

Lenci, Alessandro & Montemagni, Simonetta & Pirrelli, Vito. 2005. *Testo e computer. Elementi di linguistica computazionale*. Roma: Carocci.

Longo, Fiammetta. 2018a. La lineetta singola. In Ferrari, Angela & Lala, Letizia & Longo, Fiammetta & Pecorari, Filippo & Rosi, Benedetta & Stojmenova, Roska, *La punteggiatura italiana contemporanea. Un'analisi comunicativo-testuale*, 141–153. Roma: Carocci.

Longo, Fiammetta. 2018b. Le lineette doppie. In Ferrari, Angela & Lala, Letizia & Longo, Fiammetta & Pecorari, Filippo & Rosi, Benedetta & Stojmenova, Roska, *La punteggiatura italiana contemporanea. Un'analisi comunicativo-testuale*, 127–140. Roma: Carocci.

Longo, Fiammetta. 2020. La lineetta nelle grammatiche dell'Ottocento. In Ferrari, Angela & Lala, Letizia & Pecorari, Filippo & Stojmenova Weber,

Roska (a cura di), *Capitoli di storia della punteggiatura italiana*, 231–246. Alessandria: Edizioni dell'Orso.

Onelli, Corinna & Proietti, Domenico & Seidenari, Corrado & Tamburini, Fabio. 2006. The DiaCORIS project: a diachronic corpus of written Italian. In *Proceedings of the 5th International Conference on Language Resources and Evaluation – LREC 2006*, Genova, 1212–1215.

Pecorari, Filippo. 2019. Punteggiatura in rete: i puntini di sospensione nella comunicazione mediata dal computer. *Linguistica e filologia* 39. 129–175.

Rossini Favretti, Rema. 2000. Progettazione e costruzione di un corpus di italiano scritto: CORIS/CODIS. In Rossini Favretti, Rema (a cura di), *Linguistica e informatica. Multimedialità, corpora e percorsi di apprendimento*, 39–56. Roma: Bulzoni.

Spina, Stefania. 2014. Il Perugia Corpus: una risorsa di riferimento per l'italiano. Composizione, annotazione e valutazione. In Basili, Roberto & Lenci, Alessandro & Magnini, Bernardo (a cura di), *The First Italian Conference on Computational Linguistics. Proceedings*, 354–359. Pisa: Pisa University Press.

Tonani, Elisa. 2011. Trattino. In Simone, Raffaele (a cura di), *Enciclopedia dell'italiano Treccani*, 1520–1522. Roma: Istituto della Enciclopedia Italiana. (https://www.treccani.it/enciclopedia/trattino_%28Enciclopedia-dell%27Italiano%29/) (Consultato il 02.02.2022.)