DOI: 10.17469/O2106SLI000022

ANNA-MARIA DE CESARE

La concezione delle congiunzioni e degli avverbi negli schemi di annotazione dei corpora d'italiano scritto: breve ricognizione e alcune proposte

Il presente studio vuole fare il punto sulle PoS usate nell'annotazione dei più recenti corpora d'italiano scritto. Per motivi di spazio, ci soffermiamo su due PoS: quelle legate alle congiunzioni e agli avverbi. Gli obiettivi perseguiti sono i seguenti: ricostruire la concezione teorico-descrittiva delle congiunzioni e degli avverbi nella linguistica dei corpora legata all'italiano; valutare quanto questa concezione sia vicina a quella della grammatica tradizionale e fare emergere gli elementi innovativi; infine, alla luce dei risultati ottenuti, formulare una serie di desiderata e proposte per rivedere le PoS associate alle congiunzioni e agli avverbi, tenendo conto delle recenti messe a punto della ricerca teorica e computazionale.

Parole chiave: schemi di annotazione, PoS-tag, congiunzioni, avverbi, corpora d'italiano scritto.

1. *Introduzione*

Tra i molteplici livelli ai quali si può annotare un corpus, il più basico, e da decenni ormai ritenuto standard (cfr. le linee guida EAGLES discusse in Monachini 1995), è senza dubbio quello relativo al markup delle parti del discorso (nome, verbo, aggettivo, pronome, articolo, avverbio, preposizione, congiunzione e interiezione) – nella linguistica dei corpora più comunemente chiamate *parts of speech* (PoS). Si tratta di un livello di annotazione linguistica del testo che interessa la morfosintassi, e la cui unità di riferimento è la *parola*.[1]

[1] Come è noto, il termine 'parola' non ha nessuna concretezza a livello teorico: una sua definizione più puntuale si basa su criteri di natura grafica, morfologica e seman-

Tenuto conto del loro carattere ontologico, il presente studio si propone di fare il punto sulle PoS usate nell'annotazione dei corpora d'italiano scritto (contemporaneo). Per motivi di spazio, ci concentreremo qui solo sulle PoS legate alle parti invariabili del discorso, in particolare sulle congiunzioni e sugli avverbi. Gli obiettivi perseguiti in questa sede sono tre:[2]

1. Ricostruire la concezione teorico-descrittiva di congiunzioni e avverbi nella linguistica dei corpora dell'italiano scritto (§ 2).
2. Valutare quanto questa concezione sia vicina a quella della grammatica tradizionale e fare emergere gli elementi innovativi, in sintonia con proposte teoriche recenti (§ 3).
3. Chiarire la necessità di rivedere le PoS associate alle congiunzioni e agli avverbi; sarebbe a nostro avviso utile tenere conto in modo più sistematico delle recenti messe a punto della ricerca – sia teorica sia computazionale – sulle PoS (§ 4).

2. *La concezione teorico-descrittiva delle congiunzioni e degli avverbi nella linguistica dei corpora dell'italiano scritto: una ricognizione*

La nostra ricognizione si basa su due aspetti: la natura delle etichette (E. *tags*) relative alle congiunzioni e agli avverbi nei *tagset* di dieci recenti corpora d'italiano scritto; e la natura dei lemmi etichettati come congiunzione e avverbio nelle liste di tre corpora a confronto. Il primo aspetto permette soprattutto di capire le proprietà definitorie delle due PoS studiate in questa sede; il secondo, invece, fa luce sull'estensione di queste due classi di parole.

2.1 Analisi delle etichette nei tagset di dieci corpora d'italiano scritto

Gli schemi di annotazione relativi al livello morfosintattico, in particolare alle parti del discorso, sono stati analizzati in modo attento per l'italiano da ultimo nel rapporto tecnico di Venturi 2009 (cfr. anche Barbera 2007b). Il rapporto si sofferma su 12 sistemi diversi per an-

tica. Per una riflessione teorica in merito, rimandiamo alle pagine di Graffi (1994: 36-37) e Chiari (2007: 49).

[2] Tralasceremo qui tutte le questioni relative all'applicazione automatica di PoS-*tagger*, ampiamente discusse nell'ambito della linguistica computazionale (cfr., tra altri, Bernardi *et al*. 2006 e Tamburini 2016 sull'italiano; Schmid 1994 per altre lingue).

notare le PoS, di cui otto sono concepiti per annotare testi scritti (tra questi, uno è pensato per i testi antichi: cfr. Barbera 2007a), tre per annotare discorsi parlati e un ultimo per annotare testi appartenenti a entrambe le tipologie (cfr. Venturi 2009: 5).

Nel presente studio ci soffermiamo su tre (degli otto) schemi di annotazione dell'italiano scritto considerati da Venturi 2009; si tratta di schemi applicati a dieci recenti corpora (cfr. i dati della Tabella 1 in Appendice):

a. il PoS-tagset concepito seguendo le linee guida EAGLES (Monachini 1995), applicato al CORIS-CODIS (cfr. Tamburini 2000);
b. lo schema progettato da Marco Baroni, che si configura oggi come quello più fortunato (è applicato a ben 8 dei 10 corpora da noi considerati);
c. lo schema di Achim Stein, applicato al corpus Araneum Italicum Maius (per una descrizione più puntuale, cfr. § 2.2).

L'ordine in cui abbiamo menzionato i tre schemi di annotazione riflette la loro ideazione cronologica: il PoS-tagset descritto al primo punto è stato sviluppato prima degli altri due; inoltre, come già indicato, lo schema di Stein si basa su quello di Baroni. Non per questo, e lo vedremo nei prossimi paragrafi, le PoS relative alle congiunzioni e agli avverbi sono più articolate nell'ultimo schema.

2.2 Valutazione delle *tag* relative alle congiunzioni e agli avverbi

Le etichette concepite per le congiunzioni e gli avverbi nei tre schemi di annotazione considerati sono molto semplici, soprattutto se paragonate a quelle di PoS variabili, *in primis* del verbo e del nome. Nei tre schemi considerati, le congiunzioni e gli avverbi sono associati a una o (al massimo) due *tag* (cfr. Tabella 1 in Appendice).

Nel PoS-tagset sviluppato in base agli standard EAGLES (applicato al CORIS-CODIS), c'è una sola *tag* associata agli avverbi (ADV), mentre sono due le *tag* applicate alle congiunzioni: CONJ_C (che sta per Congiunzioni Coordinanti) e CONJ_S (per congiunzioni subordinanti). Nello schema di annotazione previsto da Baroni si trova invece una sola *tag* per le congiunzioni (CON), mentre sono due le *tag* concepite per gli avverbi: ADV_mente (per gli avverbi in *-mente*) e ADV (per tutti gli altri avverbi). Infine, nello schema di Stein tro-

viamo una sola *tag* per gli avverbi (ADV) e una sola *tag* anche per le congiunzioni (CON).

Un confronto tra i tre schemi di annotazione permette di rilevare alcune differenze nella scelta delle *tag* associate alle congiunzioni e agli avverbi. Nel caso delle congiunzioni, o si usa una sola etichetta generale, chiamata 'Conjunction' (Baroni, Stein), oppure si usano due etichette più specifiche, chiamate 'Coordinating conjunction' e 'Subordinating conjunction', senza che vi sia una *tag* generale per la classe (EAGLES). Per quanto riguarda gli avverbi, o si usa l'etichetta generale 'Adverb' (EAGLES, Stein), o si opta per due etichette (Baroni): 'Adverb' (che va in questo caso intesa alla luce della seconda: sono tutti gli avverbi non derivati con il suffisso *-mente*) e 'Adverb in -mente'. Nei tre schemi di annotazione considerati, l'interpretazione dell'etichetta ADV non è dunque uniforme.

A quanto detto si aggiunge una differenza relativa ai criteri in base ai quali le due PoS si articolano al loro interno: le congiunzioni sono suddivise in base a un criterio sintattico (coordinazione vs. subordinazione), mentre gli avverbi sono distinti in base a un criterio morfologico (± derivazione con il suffisso *-mente*).

2.3 Analisi dei lemmi etichettati CON e ADV: dati quantitativi

Passiamo ora all'analisi dei lemmi etichettati come 'congiunzione' (CON) e 'avverbio' (ADV) in tre corpora d'italiano presi a campione tra i dieci considerati:[3]

1. Timestamped JSI web corpus Italian 2014-2021 (TIt)
2. itTenTen16 (itTT)
3. Araneum Italicum Maius (AItM)

I tre corpora sono stati selezionati in base al fatto che presentano importanti punti in comune: la loro modalità di creazione (tramite web-crawling) e la tipologia alla quale appartengono. Si tratta di 'corpora di ambito generale' (così secondo SkE), con un'importante area di sovrapposizione relativa alla prosa giornalistica. Allo stesso tempo i tre corpora si differenziano per almeno due aspetti: la loro dimensione e il loro schema di annotazione (cfr. Tabella 2).

[3] I tre corpora sono disponibili sulla piattaforma Sketch Engine (SkE), cfr. Kilgarriff *et al*. 2014.

Tabella 2 - *Proprietà di tre corpora a confronto*

Corpus	*Dimensione (n. di parole)*	*File di parametri*
TIt	ca. 8,7 miliardi	TT Baroni
itTT	ca. 5 miliardi	TT Baroni
AItM	ca. 900 milioni	TT Stein

La frequenza assoluta dei lemmi etichettati CON e ADV nei tre corpora è proposta nella Tabella 3.[4]

Tabella 3 - *Frequenza assoluta dei lemmi CON e ADV*

Corpus	*CON*	*ADV*
TIt	50	16.154
itTT	47	38.124
AItM	783	1.912

Da un punto di vista quantitativo generale va dapprima rilevato che la categoria degli avverbi (che include anche quelli in *-mente*) ha un'estensione molto maggiore di quella delle congiunzioni. Tale differenza riflette proprietà semantiche e categoriali ben note delle due classi (cfr. De Cesare 2019: 23-25): gli avverbi formano una classe lessicale aperta, i cui elementi costitutivi hanno un significato perlopiù denotativo; le congiunzioni entrano invece in una classe grammaticale chiusa, i cui membri veicolano un significato di tipo istruzionale.

Se osserviamo poi i risultati ottenuti per le due PoS in modo separato, spicca soprattutto la differenza tra i dati dei tre corpora. Il dato relativo agli avverbi è senza dubbio più difficile da spiegare: non vi è infatti nessuna correlazione tra la dimensione dei tre corpora e la quantità di avverbi presenti in ognuno di essi. Il numero di ADV presenti in TIt e itTT (due corpora annotati con lo stesso schema), è inversamente proporzionale alla dimensione delle due risorse: sono due volte meno numerosi nel primo corpus (16.154 vs. 38.124), che è però

[4] Pur essendo disponibili sulla stessa piattaforma (SkE), i tre corpora non sono sempre interrogabili allo stesso modo perché presentano interfacce in parte diverse. Per ottenere informazioni sui lemmi associati alle due PoS che ci interessano bisogna seguire due cammini distinti: nel caso di TIt/itTT si possono cercare i lemmi taggati CON e ADV direttamente nella "lista di frequenza". Nel caso di AItM, invece, si deve per forza partire da una ricerca relativa a tutte le *tag*, per poi selezionare i lemmi relativi alle CON e agli ADV.

due volte più ampio del secondo (8,7 vs 5 miliardi di parole). Vi è poi il fatto che il numero di avverbi in AItM è sorprendentemente basso (1912). La differenza tra i corpora (almeno tra i due primi) sembra da ricondurre alla natura dei testi di cui si compongono. Per capire meglio i risultati ottenuti, bisognerebbe proporre un'analisi più approfondita, che in questa sede non siamo tuttavia in grado di offrire.

I dati relativi alle congiunzioni sono in parte più facili da interpretare: i due primi corpora (TIt e itTT) presentano praticamente lo stesso numero di lemmi etichettati CON (rispettivamente 50 e 47). Questa cifra riflette bene le proprietà intrinseche della classe, che non è solo chiusa, ma anche composta da un numero relativamente ridotto di membri. In altre parole, diversamente dagli avverbi, il numero di lemmi taggati CON non dipende dalla dimensione del corpus. Alla luce di queste considerazioni, sorprende dunque ancora una volta il dato ottenuto per il corpus AItM, che è però questa volta inaspettatamente elevato (783). Una verifica più puntuale dei lemmi associati alla *tag* CON in questo corpus permette di fare luce sull'ampio numero di forme riportate: AItM presenta un'etichettatura 'difettosa' almeno a partire dai lemmi che occupano il rango d'uso 100. Se scorriamo i lemmi etichettati CON dopo il rango 100, troviamo per esempio forme come 'O_o', 'intr-o', 'e-o', 'Ukiyo-e', 'WEB-MA', 'saxdax-e', che non hanno chiaramente nulla a che fare con le CON, e non sono neanche veri e propri lemmi. Si tratta in molti casi semplicemente di forme che includono una congiunzione prototipica (*e, ma, o*).

2.4 Analisi dei lemmi etichettati CON: valutazione qualitativa

Soffermiamoci ora sulle forme di CON che entrano a far parte delle tre liste di frequenza (nel caso del corpus AItM limitiamo l'analisi ai primi 50 lemmi).

Nei tre corpora le teste di lista sono assolutamente identiche: troviamo (nello stesso ordine) *e, ma, o, ed* e *se*. Sono poi presenti, ma non più allo stesso rango d'uso, la variante grafica *od* e le varianti grafico-semantiche *ovvero, oppure* e *ossia*. Inoltre, sempre nei tre corpora, sono inclusi *né* e la variante *nè* (la prima occupa sempre un rango d'uso più elevato), *sebbene* e *seppure*. Altre CON presenti in tutte e tre le liste sono *mentre* e *sia* (che occupano ranghi elevati), *nonostante* e *bensì* (che occupano ranghi intermedi).

Comune alle tre liste è poi la presenza di *che*. Il lemma non occupa però gli stessi ranghi d'uso, né è associato alla stessa frequenza. In TIt e itTT, *che* ha una frequenza d'impiego molto bassa (esso compare, rispettivamente, 1.388 e 1.732 volte) e occupa uno degli ultimi ranghi. In AItM, invece, *che* compare 110269 volte e occupa il rango 10. La differenza tra TIt e itTT da una parte e AItM dall'altra si spiega facilmente alla luce del fatto che i primi due corpora prevedono una *tag* ad hoc per la parola *che*, ovvero CHE (torneremo a parlare di questa scelta nel § 3.2).

Tutte e tre le liste di CON includono anche un numero piuttosto consistente di lemmi composti da *che*, come per esempio (citando solo le cinque voci più frequenti) *nonché, affinché, poiché, finché, anziché* (TIt); *nonché, poiché, affinché, finché, anziché* (itTT); *perché, perchè, nonché, poiché, affinché* (AItM). I lemmi che contengono la forma *che* sono numerosi soprattutto nei due primi corpora, dove formano circa la metà delle due liste. In AItM, i lemmi formati con *che* sono invece meno numerosi (ci sono ca. 10 entrate in meno). Il dato si spiega con il fatto che nel terzo corpus altri lemmi taggati CON hanno una frequenza d'uso più elevata e occupano i primi 50 ranghi (declassando a ranghi più bassi forme marginali di CON basate su *che*: è il caso di *sicché, giacché, allorché*). Si tratta in particolare di *come, quando, visto, oltre, qualora, così, salvo, ebbene.* Queste forme sono CON solo in AItM. In TIt e itTT sono quasi tutte etichettate WH. Lo stesso vale del resto anche per *perché*.

Va infine rilevato che nelle tre liste di CON compaiono forme come *dopo, tra / fra* ecc., che possiamo descrivere in modo unitario: sono parole polifunzionali, principalmente taggate PRE. Per queste forme si osservano però differenze importanti tra i tre corpora. Prima di tutto in merito al loro numero: ve ne sono due in itTT (*senza* e *dopo*), quattro in TIt (*dopo, senza, tra* e *fra*) e sei in AItM (*fino, prima, per, senza, dopo, a*). Un'altra differenza riguarda la loro frequenza d'uso: *dopo*/CON è per esempio molto più frequente nei due primi corpora (903.400 occ. in TIt; 298.125 occ. in itTT e 728 occ. in AItM). In questo caso, di nuovo, le differenze non si spiegano unicamente in base alla diversa dimensione dei tre corpora (vi sono tre volte più occorrenze in TIt che in itTT).

3. *Valutazione delle PoS associate alle congiunzioni e agli avverbi: tra concezione tradizionale e proposte innovative*

3.1 Cenni sulla concezione tradizionale delle congiunzioni e degli avverbi

Nella lezione tradizionale sulle parti del discorso (nell'ambito delle grammatiche italiane rappresentata per esempio da Serianni 2000), le parole sono raggruppate in classi (o paradigmi) in base alla loro flessione: a un primo livello di analisi, si distinguono le parole variabili da quelle invariabili. Le parti variabili possono essere ulteriormente distinte (in nome, verbo, aggettivo, articolo, pronome) osservando più da vicino i tratti flessivi di cui si compongono (per es. il genere, numero, la persona, il tempo ecc.). Questo *modus operandi* non è invece possibile per le parole invariabili che – per definizione – non possono essere flesse. La distinzione tra le quattro classi invariabili del discorso (avverbio, congiunzione, preposizione e interiezione) verte dunque su criteri di natura semantica e/o sintattica (De Cesare 2019: 26-27).

In parte a seguito della natura dei criteri in base ai quali sono suddivise, le categorie invariabili del discorso non hanno confini molto netti. Basta considerare l'esempio paradigmatico della parola invariabile *dopo*, che secondo la grammatica tradizionale può fungere da avverbio, congiunzione e preposizione. La sua specifica classe di appartenenza dipende dalle proprietà sintattiche che intrattiene con il resto della frase. Quando la parola *dopo* non entra in relazione con nessun altro costituente seguente, cioè non regge un complemento, la si assegna alla categoria degli avverbi (cfr. 1); quando, invece, *dopo* regge una frase (anche di modo finito, come *aver sostenuto l'esame*), abbiamo una congiunzione (2); e quando regge un sintagma nominale (come *l'esame*), una preposizione (3):

(1) Ti sentirai meglio dopo. (avverbio)
(2) Ti sentirai meglio dopo *aver sostenuto l'esame*. (congiunzione)
(3) Ti sentirai meglio dopo *l'esame*. (preposizione)

Nella grammatica tradizionale ognuna delle quattro categorie invariabili è ulteriormente suddivisa in base a criteri morfosintattici. Nel caso delle due classi che ci interessano, possiamo osservare la suddivisione riportata in Tabella 4: le congiunzioni sono suddivise in tre gruppi (De Cesare 2019: 41); gli avverbi, invece, in quattro gruppi (De Cesare 2019: 52).

Tabella 4 - *Suddivisione delle congiunzioni e degli avverbi (ed esempi)*

Congiunzioni	*Avverbi*
semplici (*e, né, ma, quindi, anche*)	semplici (*ora*)
composte (*oppure*)	composti (*soprattutto*)
locuzioni (*dal momento che, visto che*)	locuzioni (*d'ora in poi*)
	derivati (*ovviamente, bocconi*)

3.2 CON e ADV: tra tradizione e innovazione

Complessivamente, la nostra indagine – che andrebbe naturalmente allargata e approfondita – permette di osservare che la concezione teorico-descrittiva delle congiunzioni e degli avverbi associata ai corpora d'italiano scritto più recenti è relativamente tradizionale.

Questo è vero prima di tutto se osserviamo le *tag* previste nei tre schemi usati per annotare dieci recenti corpora d'italiano scritto: in questi schemi troviamo infatti due etichette di primo livello – CON per le congiunzioni e ADV per gli avverbi. Inoltre, alla stregua delle scelte della grammatica tradizionale, le articolazioni interne alle due *tag* si basano su criteri sintattici (CONJ_C e CONJ_S) e morfologici (cfr. ADV_mente vs. ADV, che copre tutte le altre forme): le *tag* più specifiche sono dunque concepite in base a criteri di natura eterogenea. A questo punto bisogna però notare una cosa importante: le scelte adottate per etichettare i corpora analizzati sono più semplici di quelle della grammatica tradizionale. In effetti, come abbiamo visto (cfr. Tabella 4) la grammatica tradizionale individua ben quattro categorie morfologiche diverse dell'avverbio (cfr. De Cesare 2019: 52): gli avverbi semplici, composti, derivati e le locuzioni avverbiali.

Un confronto tra i tre schemi di annotazione permette anche di osservare delle differenze tra i *tagset*. Lo schema relativo agli standard EAGLES è quello più tradizionale. Diversamente dagli altri due (quello di Baroni e di Stein), esso non prevede una serie di *tag* di primo livello legate alle congiunzioni e agli avverbi. Pensiamo in particolare alle etichette CHE e WH-WORD, ispirate alle proposte della linguistica teorica: la prima *tag* si rifà ad una nuova concezione della congiunzione *che*, concepita come 'complementatore'; la seconda *tag* accoglie invece la proposta di raggruppare parole come *chi, dove, perché* ecc. – nella grammatica tradizionale concepite come trasversali agli avverbi, pronomi e aggettivi (per una discussione, cfr. Salvi 2013: 37-39) – in un'unica nuova categoria (le *wh-word* della grammatica generativa).

Se spostiamo la nostra attenzione sulle forme che entrano a far parte del lemmario delle CON, possiamo osservare che nelle liste di frequenza dei tre corpora analizzati (annotati con i due schemi più recenti, di Baroni e Stein) sono presenti molte parole tradizionalmente concepite come congiunzioni, a cominciare dalle forme prototipiche di congiunzioni coordinanti (*e, ma, o*) – con le rispettive varianti grafiche (*ed, od*) e grafico-semantiche (*oppure, ossia, ovvero*) – e subordinanti (*se, che*).

Detto questo, è doveroso mette in luce che le tre liste analizzate presentano almeno un aspetto innovativo importante. Forme come *quindi, inoltre* e *anche*, che la grammatica tradizionale annovera tra le congiunzioni, non sono etichettate CON, ma ADV – si tratta dunque di avverbi (tali forme non compaiono nella lista delle ca. 700 CON del corpus AItM).

Per il resto, sia a livello terminologico che a quello concettuale, i corpora esaminati tengono poco conto delle proposte mosse negli ultimi anni nell'ambito della linguistica generale (cfr. Prandi 2007; Salvi 2013, 2014; De Cesare 2019) e nel campo della linguistica computazionale (Bernardi *et al*. 2006; D'Errico *et al*. 2016). Una di queste riguarda la categoria delle congiunzioni. Entrambi gli approcci (quello teorico della linguistica generale e quello a-teorico di quella computazionale) argomentano a favore dell'eliminazione di questa parte invariabile del discorso e propongono di ridistribuire le forme tradizionalmente concepite come congiunzioni nella classe degli avverbi (*inoltre, anche*), delle preposizioni (*dopo, fra, perché*) e degli operatori detti 'logici' o 'sintattici' (*e, o, ma; che, se*). Accogliere questa proposta nei file di parametri da applicare ai corpora d'italiano significherebbe anche elaborare maggiormente le *tag* associate alla categoria degli avverbi, proponendo un'articolazione più fine delle etichette di secondo livello.

4. *Conclusioni: tra desiderata e proposte*

I corpora rappresentativi dell'italiano scritto dell'ultima generazione, di cui abbiamo analizzato vari esempi paradigmatici (TIt, itTT e AiTM), si distinguono per la loro modalità di creazione e dimensione: sono il risultato di sistemi di webcrawling e contengono miliardi di parole. A fronte della rapida evoluzione registrata negli ultimi anni

nell'ambito della creazione di nuovi corpora, i progressi fatti nel campo della loro annotazione grammaticale sono piuttosto contenuti.[5]

La nostra analisi delle etichette concepite per le congiunzioni e gli avverbi negli schemi di annotazione impiegati in dieci recenti corpora d'italiano scritto (contemporaneo) permette di affermare che il metalinguaggio della linguistica dei corpora italiana non ha ancora recepito e integrato le proposte della linguistica teorica e computazionale: per l'annotazione delle due PoS impiega un tagging basato su una concezione tradizionale delle congiunzioni e degli avverbi e le *tag* sono relativamente fisse (non si registrano grandi cambiamenti dallo schema basato sugli standard EAGLE agli schemi successivi).

Alla luce dei risultati ottenuti in questo studio ci sembra dunque doveroso esprimere un desiderio: quello di tornare a riflettere sulle *tag* relative alle congiunzioni e agli avverbi negli schemi di annotazione dei testi d'italiano scritto (ma non solo). Converrebbe a nostro parere concepire nuove *tag* (di primo, ma anche di secondo livello), basate non più tanto sulle proprietà morfologiche delle parole quanto su quelle sintattiche (una riflessione importante in questo senso è in Schütze 1995 e Tamburini 2016).

Un buon punto di partenza per la revisione delle PoS associate agli avverbi è lo schema di annotazione morfosintattico del Turin University Treebank/TUT (Lesmo *et al*. 2002). In sintonia con la ricerca teorica sulle parti del discorso, in TUT parole come *sì* e *no* non sono più etichettate ADV (seguendo la concezione della grammatica tradizionale) ma PHRAS (che sta per *phrasals* 'frasali'), una nuova *tag* di primo livello concepita per la classe delle profrasi (sulle proprietà definitorie della categoria, cfr. Bernini 1995; per una discussione relativa al riassestamento del quadro relativo alle parti invariabili del discorso, cfr. De Cesare 2019: 59-62).

Lo schema TUT prevede poi varie *tag* di secondo livello, che sono tuttavia ancora largamente basate su criteri semantici tradizionali (basta considerare le classi degli avverbi di maniera, di quelli locativi, temporali, negativi e quantificativi). Da ciò consegue che nello schema TUT mancano alcune sottoclassi di avverbi ormai ben consolidate a livello teorico-descrittivo, tra cui quelle degli avverbi connettivi (in TUT, *inoltre* è per esempio taggato congiunzione coordinativa), frasa-

[5] La riflessione teorico-descrittiva sulle PoS non manca però in rapporto ai corpora d'italiano antico (cfr. Barbera 2007a; Iacobini *et al*. 2017).

li (sempre in TUT, *forse* è avverbio di dubbio) e focalizzanti. Questi ultimi includono parole come *anche, persino/perfino* (in TUT concepiti come 'avverbi di intensità'), *solo-soltanto* (etichettati 'avverbi di limitazione') e *nemmeno-neanche* (concepiti come 'avverbi di negazione').[6]

Una revisione delle PoS associate alle congiunzioni e agli avverbi permetterebbe di migliorare l'etichettatura automatica dei corpora. I risultati ottenuti per la parola *dopo* bastano per mostrare che una concezione tradizionale delle PoS è molto problematica. In itTT, per esempio, *dopo* è spesso etichettata CON laddove – almeno così secondo il punto di vista della grammatica tradizionale – si tratta chiaramente di un avverbio: cfr. i sintagmi *due mesi dopo* e *otto giorni dopo* nella schermata dell'Appendice 2. La ricerca teorica recente risolve questo problema raggruppando parole come *dopo* in un'unica classe: quella della preposizione (per dettagli, cfr. Salvi 2013: 119-121).

Problemi di etichettatura come quello appena illustrato andrebbero risolti al più presto, vuoi perché molte congiunzioni e molti avverbi (con significato non denotativo) sono associati a una frequenza d'uso elevata, vuoi perché hanno ricadute importanti sui risultati di altre ricerche, come la creazione di Word Sketches. Una revisione delle etichette relative alle congiunzioni e agli avverbi aumenterebbe il grado di accuratezza del tagging automatico e, di conseguenza, la descrizione dei dati linguistici.

Riferimenti bibliografici

Barbera, Manuel. 2007a. Un tagset per il corpus Taurinense. Italiano antico e linguistica dei corpora. In Barbera, Manuel & Corino, Elisa & Onesti, Cristina (a cura di), *Corpora e linguistica in rete*, 135-168. Perugia: Guerra Edizioni.

Barbera, Manuel. 2007b. Mapping dei tagset in bmanuel.org / corpora.unito.it. Tra *guidelines* e prolegomeni. In Barbera, Manuel & Corino, Elisa & Onesti, Cristina (a cura di), *Corpora e linguistica in rete*, 135-168. Perugia: Guerra Edizioni.

Baroni, Marco & Bernardini, Silvia & Comastri, Federico & Piccioni, Lorenzo & Volpi, Alessandra & Aston, Guy & Mazzoleni, Marco. 2004. Introducing the *la Repubblica* Corpus: A large, annotated, TEI(XML)-

[6] Per dettagli sugli avverbi focalizzanti, cfr. De Cesare (2019: 90-95).

compliant corpus of newspaper Italian. *Proceedings of LREC 2004*, 1771-1774. Lisbon: ELDA.

Bernardi, Raffaella & Bolognesi, Andrea & Seidenari, Corrado & Tamburini, Fabio. 2006. POS tagset design for Italian. In *Proceedings of the 5th International Conference on Language Resources and Evaluation*. Genova.

Bernini, Giuliano. 1995. Le profrasi. In Renzi, Lorenzo & Salvi, Giampaolo & Cardinaletti, Anna (a cura di), *Grande grammatica italiana di consultazione* 3, 175-222. Bologna: Il Mulino.

Chiari, Isabella. 2007. *Introduzione alla linguistica computazionale*. Bari: Laterza.

De Cesare, Anna-Maria. 2019. *Le parti invariabili del discorso*. Roma: Carocci.

D'Errico, Marianna & Grandi, Nicola & Paternesi Meloni, Serena & Tamburini, Fabio. 2016. Induzione di categorie grammaticali e lessicali. In Dedè, Francesco (a cura di), *Categorie grammaticali e classi di parole. Statuto e riflessi metalinguistici*, 115-137. Roma: Il Calamo.

Graffi, Giorgio. 1994. *Sintassi*. Bologna: Il Mulino.

Iacobini, Claudio & De Rosa, Aurelio & Schirato, Giovanna. 2017. Criteri e strategie di classificazione morfo-sintattica dei testi del corpus MIDIA. In D'Achille, Paolo & Grossmann, Maria (a cura di), *Per la storia della formazione delle parole in italiano. Un nuovo corpus in rete (MIDIA) e nuove prospettive di studio*, 33-51. Firenze: Cesati.

Kilgarriff, Adam & Baisa, Vít & Bušta, Jan & Jakubíček, Miloš & Kovář, Vojtěch & Michelfeit, Jan & Rychlý, Pavel & Vít Suchomel. 2014. The Sketch Engine: ten years on. *Lexicography* 1(1). 7-36.

Lesmo, Leonardo & Lombardo, Vincenzo & Bosco, Cristina. 2002. Treeback development. The TUT approach. In *Proceedings of ICON02*. Mumbay, India.

Monachini, Monica. 1995. ELM-IT: An Italian Incarnation of the EAGLES-TS. Definition of Lexicon Specification and Classification Guidelines. In *Technical report*. Pisa.

Prandi, Michele. 2007. Avverbi di collegamento e congiunzioni. In San Vicente, Félix (a cura di), *Partículas. Particelle. Estudios de lingüística contrastiva español e italiano*, 89-103. Bologna: CLUEB.

Salvi, Giampaolo. 2013. *Le parti del discorso*. Roma: Carocci.

Salvi, Giampaolo. 2014. La classificazione delle parti del discorso. *Italogramma* 8. 55-74.

Schmid, Helmut. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of International Conference on New Methods in Language Processing*. Manchester, UK.

Schütze, Hinrich. 1995. Distributional Part-of-speech Tagging. In *Proceedings of 7th EACL*, 141-148. Dublin, Ireland.

Serianni, Luca. 2000. *Italiano*, con la collaborazione di A. Castelvecchi. Torino: Garzanti.

Tamburini, Fabio. 2000. Annotazione grammaticale e lemmatizzazione di corpora in italiano. In Rossini Favretti, Rema (a cura di), *Linguistica e informatica: multimedialità, corpora e percorsi di apprendimento*, 57-73. Roma: Bulzoni.

Tamburini, Fabio. 2016. A BiLSTM-CRF PoS-tagger for Italian tweets using morphological information. In *CLiC-it/EVALITA 2016*. Napoli.

Venturi, Giulia. 2009. Rassegna comparativa degli schemi di annotazione morfosintattica per la lingua italiana. In *Rapporto Tecnico TRIPLE*.

Appendice

Tabella 1 - *Corpora e schemi di annotazione a confronto*

	Nome corpus	*Annotazione*	*Tag/Code*	*Parte del discorso*
1.	CORIS-CODIS	PoS-tagset concepito seguendo le linee guida EAGLES (Monachini 1995)	• ADV • CONJ_C • CONJ_S	• Avverbi • Cong. Coord. • Cong. Subord.
2.	la Repubblica	Tree Tagger basato sul file di parametri di Marco Baroni (Baroni *et al*. 2004)	• ADV	• Avverbio
3.	Timestamped JSI web corpus Italian		• ADV:mente	• Avverbio in *-mente*
4.	COMPARE-IT		• CON	• Congiunzione
5.	Paisà			
6.	OPUS2 Italian			
7.	itTenTen16			
8.	itWAC (reduced)			
9.	MIDIA			
10.	Araneum Italicum Maius (2014)	Tree Tagger basato sul file di parametri di Achim Stein	• ADV • CON	• Avverbio • Congiunzione

Appendice 2 - *La parola* dopo/*CON in itTT*

1	lanazione.it	e mi sbatte a terra . </s><s> Duecento metri	dopo	si ferma per prestarmi soccorso . </s><s> Arriva u
		in CON CLI VER:fin PRE NOUN SENT DET:num NOUN	CON	CLI VER:fin PRE VER:infi:cli NOUN SENT VER:fin AI
2	virgilio.it	roblemi sono molto differenti . </s><s> Il maggiore	dopo	vari cambiamenti di scuola (diversi sistemi) è
		NOUN VER:fin ADV ADJ SENT ART ADJ	CON	ADJ NOUN PRE NOUN PUN ADJ NOUN PUN AUX:s
3	innomedimaria.i...	ili di nuovo al cielo </s><s> - otto giorni	dopo	ai discepoli con Tommaso </s><s> - alcuni gi
		t:fin PRE NOUN ARTPRE NOUN NOCAT DET:num NOUN	CON	ARTPRE NOUN PRE NPR NOCAT DET:indef N
4	innomedimaria.i...	i discepoli con Tommaso </s><s> - alcuni giorni	dopo	sul Lago di Tiberiade , miracolo della pesca (
		PRE NOUN PRE NPR NOCAT DET:indef NOUN	CON	ARTPRE NPR PRE NPR PUN NOUN ARTPRE NOUN P
5	amazingcomics.i...	arrello . </s><s> Nato in provincia di Venezia ,	dopo	essermi diplomato all ' Istituto Statale d' Arte di
		NOUN SENT NPR PRE NOUN PRE NPR PUN	CON	AUX:infi:cli VER:ppast NOUN PUN NPR NPR PRE NPR ART
6	chiesacattolica...	rato in attesa di processo , rilasciato sei mesi	dopo	per non aver commesso il fatto ; e quella
		past PRE NOUN PRE NOUN PUN VER:ppast DET:num NOUN	CON	PRE NEG AUX:infi VER:ppast ART NOUN PUN CON PRO:demo
7	chiesacattolica...	he non ha avuto nessun ' altra imputazione	dopo	di allora e non sta godendo di alcun beneficio
		CHE NEG AUX:fin VER:ppast DET:indef PUN DET:indef NOUN	CON	PRE ADV CON NEG VER:fin VER:geru PRE DET:indef NOUN
8	quipo.it	altro la scatola sul n ° 2 , quadrato	dopo	quadrato , fino ad arrivare al dieci . </s>
		PRO:indef ART NOUN ARTPRE ADJ NOCAT NUM PUN ADJ	CON	ADJ PUN PRE PRE VER:infi ARTPRE DET:num SENT
9	indire.it	e i suoi fratelli si recano al tempio e ,	dopo	averlo purificato , demoliscono l' altare dei sa
		ON ART DET:poss NOUN CLI VER:fin ARTPRE NOUN CON PUN	CON	AUX:infi:cli VER:ppast PUN VER:fin ART NOUN ARTPRE N
10	innomedimaria.i...	ifiuto di Nazaret viene narrato da Marco molto tempo	dopo	che l' attività pubblica di Gesù è iniziata .
		NOUN PRE NPR AUX:fin VER:ppast PRE NPR ADV NOUN	CON	CHE ART NOUN ADJ PRE NPR AUX:fin VER:ppast SENT
11	repubblica.it	mo paese che ha riconosciuto l' Ucraina il giorno	dopo	che questa s ' è proclamata indipendente d
		DJ NOUN CHE AUX:fin VER:ppast ART NPR ART NOUN	CON	CHE DET:demo ADJ PUN AUX:fin VER:ppast ADJ ART
12	dnet.it	no del Mali . </s><s> Ciononostante due anni	dopo	rprenderanno gli scontri . </s><s> Ad aggravare l
		L ARTPRE NPR SENT NPR DET:num NOUN	CON	VER:fin ART NOUN SENT PRE VER:infi AI
13	amyresource.it	a che non dipenderà da un Pc o altro computer	dopo	che abbiamo visto AInc ritrattare costantemente le su
		CHE NEG VER:fin PRE ART NPR CON DET:indef NOUN	CON	CHE VER:fin NOUN NPR VER:infi ADV:mente ART DET
14	sangiorgiohotel...	stando sul lato sinistro della strada . </s><s>	Dopo	circa 800mt si scorgerà la statua intitolata a Garibal
		VER:geru ARTPRE NOUN ADJ ARTPRE NOUN SENT	CON	ADV NPR CLI VER:fin ART NOUN VER:ppast PRE NPR
15	lacasadellezucc...	Giganti " Is Concias " svoltare a destra :	dopo	circa 100m al bivio prendere la via Monte Nied
		E NPR PUN NOUN NPR PUN VER:infi PRE NOUN PUN	CON	ADV ADJ ARTPRE NOUN VER:infi ART NOUN NPR NPR
16	niederwieser.it	er la macinazione della carne . </s><s> Ora ,	dopo	otre trent ' anni di attività nel settore , Tippe
		RE ART NOUN ARTPRE NOUN SENT ADV PUN	CON	ADV ADJ PUN NOUN PRE NOUN ARTPRE NOUN PUN NP