Criteri di lemmatizzazione

La lemmatizzazione, ovvero l'operazione di ricondurre ogni parola di un testo alla forma base (o entrata di dizionario) è un'operazione lunga e complessa, sia perché attualmente ancora non si dispone di lemmatizzatori automatici totalmente efficienti, sia perchè i criteri di lemmatizzazione devono fare i conti con la complessità dei fenomeni linguistici.

I criteri che vengono scelti sono lo specchio della grammatica a cui si fa riferimento. Le norme da noi scelte riflettono le convinzioni (o i ragionevoli compromessi) a cui è giunto, anche dopo approfondite discussioni, il gruppo di persone che ha lavorato al progetto del corpus. Nel dichiararle esplicitamente pensiamo di poter mettere chiunque in condizione di usufruire del nostro lavoro.

Il corpus è stato lemmatizzato automaticamente, in prima battuta, con un lemmatizzatore gentilmente messo a disposizione gratuitamente dalla IBM Italia, attraverso la persona dell'Ing. Federico Mancini, poiché all' epoca non si disponeva di un analizzatore di testi. Si tratta dello stesso lemmatizzatore che è stato utilizzato per la lemmatizzazione del corpus di parlato LIP (De Mauro, Mancini, Vedovelli, Voghera, Etas 1994), a cui spesso si farà riferimento. Il lemmatizzatore (descritto da Federico Mancini nel capitolo 4 del LIP) ha operato seguendo le norme di lemmatizzazione esposte da Miriam Voghera nel capitolo 5.2 (pp. 89-96) del LIP.

Nel nostro gruppo di ricerca, non sempre ci siamo trovati concordi sulle scelte di lemmatizzazione del LIP, ma non abbiamo potuto modificare le opzioni, perché l'IBM ci ha fornito solo gli eseguibili e non la sorgente dei programmi.

E' stato però possibile apportare alcuni cambiamenti con procedure che hanno agito automaticamente o manualmente sull'output del lemmatizzatore. I principali cambiamenti operati sono stati tre:

1) Unificare in un' unica categoria, chiamata Nomi Propri, tutte le forme che il lemmatizzatore classifica con i codici D, E, F, L. Questa scelta è stata operata per considerazioni teoriche e pratiche; in particolare, in queste categorie il lemmatizzatore dava luogo a frequenti errori di classificazione, per esempio categorizzando i nomi propri di persona come cognomi e viceversa; il raggruppamento di tutti i nomi propri sotto un unico codice ha permesso di ridurre gli interventi di correzione manuale necessari.

2) Sostituire i codici A e O (abbreviazioni e esotismi) con il codice S (sostantivi). Essere un'abbreviazione o un esotismo è una proprietà di natura diversa da quella di appartenere a una determinata categoria grammaticale; inoltre, le parole classificate come abbreviazioni ed esotismi sono quasi tutte funzionalmente dei sostantivi. Anche questa operazione ha permesso di ridurre la quantità di interventi di correzione manuale.

3) Trattare le parole sintagmatiche al cui interno compaiono altri elementi.

A ciascuna parola (o ricorrenza) è stato associato un codice che la assegna alla classe grammaticale di appartenenza.

ARTICOLI Art.

Gli articoli sono stati lemmatizzati seguendo i criteri adottati dal LIP, in cui si lemmatizzano tutti gli articoli nella forma maschile singolare: "Sono registrate sotto gli stessi lemmi le ricorrenze degli articoli nelle preposizioni articolate. I partitivi sono lemmatizzati sotto l'articolo indefinito UNO" .

L'articolo partitivo nelle sue varie forme, veniva spesso confuso dal lemmmatizzatore con le forme della preposizione di articolata. Questi errori sono stati corretti manualmente. Analogamente, in frasi come di pane ne ho molto e di amici ne ho molti i due di sono stati trattati come casi di articolo partitivo.

Un caso particolare si è presentato con i partitivi che seguono il quantificatore un po'. Quest'ultimo è stato trattato come sintagmatica di tipo avverbiale solo in usi assoluti, quali fa un po' caldo, hai fame? Un po'. Nei casi in cui si è trovato il sintagma un po' di, si sono lemmatizzati isolatamente i singoli membri e di è stato trattato come un partitivo. Si è presentato anche il caso dell' articolo partitivo francese, nella frase rien que des mensonges: in questo caso il des è stato ricondotto all'articolo partitivo francese DU.

Le preposizioni articolate francesi au, du e des, omografa dell'articolo partitivo, sono state ricondotte alle preposizioni semplici francesi A e DE.

AGGETTIVI Agg.

Sono stati adottati gli stessi criteri del LIP, ovvero: “tutti gli aggettivi al maschile singolare e nella loro forma positiva. Vengono invece portate a lemma tutte le forme sintetiche di comparativo e superlativo: per esempio migliore, ottimo, ecc.. Gli aggettivi tronchi (bel, gran, ecc.) sono lemmatizzati come forme del lemma rappresentato dalla forma di citazione non tronca (bello, grande, ecc.). Sono registrati come aggettivi gli usi aggettivali dei participi passati in frasi come *il bar all'angolo é aperto".*

Gli aggettivi sostantivati sono stati categorizzati come sostantivi, per esempio in alloggi di presunti mafiosi e pregiudicati, la forma mafiosi è stata categorizzata come sostantivo; in prima del solito, prima del previsto, solito e previsto sono stati classificati come sostantivi e non aggettivi.

Gli aggettivi alterati (ad esempio, bellino) sono stati ricondotti al lemma di grado positivo corrispondente e contrassegnati con il codice M.

La distinzione tra uso aggettivale e uso participiale dei participi passati è stata frequentemente problematica. Si è scelto di categorizzare come participi i participi in uso assoluto nelle subordinate temporali (ad esempio nella frase: *un tempo usate come cucchiaini da tè*, *usate* è stato lemmatizzato come verbo, in quanto commutabile con che un tempo erano usate...) e quelli in funzione predicativa in frasi nominali, soprattutto titoli (ad esempio, *I segreti dei Marlborough rivelati in un' intervista; Sedotti dalle bambole d'epoca*). Sono stati invece classificati come aggettivi i participi usati come predicati nominali del verbo essere, anche se reggevano dei complementi (per es., è *scrupolosamente vestita con panni...; è corredata di accessori...*), in quanto si tratta di strutture parallele ad altre contenenti aggettivi, quali è *piena di..., è ricca di...* e simili.

Nei casi dubbi, per avere un prodotto omogeneo si è deciso di eseguire alcuni test.

Nella Grande grammatica italiana di consultazione, vol. 2, nel capitolo dedicato al sintagma aggettivale (curato da M.T. Guasti) vengono offerti tre test per la discriminazione tra Aggettivi e Participi:

1) i participi non possono essere modificati con -issimo e con molto, assai ecc., gli aggettivi sì.
2) solo il participio passato usato come attributo di un nome può essere sede di un clitico: *la notizia*

comunicatagli vs. *la ragazza simpaticagli*, *la notizia sconosciutagli*. Se è sede di un clitico, il participio è verbo e non aggettivo.
3) i participi sono compatibili con gli ausiliari essere e venire, gli aggettivi solo con essere.

Questi criteri sono stati applicati ad alcuni casi problematici incontrati:

a) *Un tempo usate come cucchiaini da tè*

criterio 1): *un tempo usatissime come cucchiaini da tè, molto/assai usate come cucchiaini da tè*; le frasi sono corrette per cui *usate* è Aggettivo;
criterio 2): il test del clitico sembra non pertinente in questo contesto;
criterio 3): gli ausiliari possono essere usati entrambi, ma solo a patto di ampliare il contesto: *che un tempo erano/venivano usate come cucchiaini ...*

Il solo test che risponde positivamente senza manipolazioni del contesto è quello che qualifica usate come Aggettivo.

b*) I segreti dei Marlborough rivelati in un'intervista*

criterio 1):*rivelatissimi, molto rivelati*
criterio 2):*rivelatici*
criterio 3): *vengono rivelati*

I test classificano rivelati come un Participio.

c) *Sedotti dalle bambole d'epoca*

criterio 1):*se dottissimi*
criterio 2): il clitico non si può applicare
criterio 3):veniamo sedotti

C'è una lieve prevalenza del valore participiale.

d) *E' scrupolosamente vestita con panni... è corredata di accessori*

criterio 1): *vestitissima, molto vestita...; corredatissima* sembra invece andare bene
criterio 2): clitico non applicabile;
criterio 3): *viene vestita* ha valore diverso da è vestita (che è stativo, mentre *viene vestita* è dinamico; lo stesso vale per *viene corredata* rispetto a *è corredata*).

Corredata risponde a tutti i test per l'aggettivo; vestita risponde meno positivamente ai test per l'aggettivo, ma risponde negativamente a quelli per il participio, e in ragione del parallelismo con corredata è stato classificato come aggettivo.

Nello stesso capitolo della Guasti viene considerato normale che un aggettivo regga complementi (ad esempio, *fedele a Maria*), quindi la presenza di complementi non è di per sé indice di verbalità. Piuttosto, potrebbe essere indice di verbalità la presenza (o la possibilità) di un complemento d' agente: ad esempi, *sedotti dalle bambole d'epoca* sarebbe più verbale di *vestita con panni, corredata di accessori* e simili.

Forniamo di seguito alcuni esempi di lemmatizzazione :

- *c'erano tutti Ugo escluso*/incluso/compreso/eccettuato: *escluso* è stato considerato una forma del verbo ESCLUDERE;
-*ormai mi sono abituato*: *abituato* è stato considerato una forma del verbo ABITUARE;
-*sono abituato alle sue manie*: *abituato* è stato considerato una forma dell'aggettivo ABITUATO;
-*è destinato alla sconfitta*: *destinato* è stato considerato una forma del verbo DESTINARE;
-*non è detto che...* : *detto* è stato considerato una forma del verbo DIRE;
-*un libro dedicato a Mario*: *dedicato* è stato considerato una forma del verbo DEDICARE;
-*sarà cosa fatta entro oggi* : è stato considerato una forma dell'aggettivo FATTO;
-*a sirene spiegate*: *spiegate* è stato considerato una forma dell'aggettivo SPIEGATO;
-*22 cervi morti ammazzati: morti ammazzati* sono stati considerati forme dei verbi MORIRE e AMMAZZARE;
-*sono 22 i morti ammazzati: morti* è stato considerato forma del sostantivo MORTO e ammazzati forma dell' aggettivo AMMAZZATO

Problemi analoghi si sono presentati con i **participi presenti**. Quando si aveva un deciso valore verbale (es. *Il treno proveniente da Roma*) sono stati lemmatizzati come verbi.

In casi di sintagmi **nome-nome** come:

I ricami oro
Una giacca avorio
Una gonna grigio ferro
Una gonna grigio perla

le parole *oro, avorio, grigio* sono state classificate come Aggettivi.

I **numerali ordinali** sono stati classificati come Aggettivi. Quindi: primo, secondo, terzo... quindicesimo ecc sono aggettivi, anche se scritti in forme alfanumeriche, come in 32esimo parallelo.

I **prefissi scritti staccati** dalla parola cui si riferiscono quali ex e mini , per es.: *la mia ex moglie; il mio mini appartamento* sono stati classificati come Aggettivi.

I **sostantivi prefissati con anti-**, laddove erano usati in funzione di modificatore, sono stati classificati come Aggettivi, per es., *provvedimenti antiemergenza*.

Gli elementi posti prima del nome, come in: *Un nuovo sentimento filo asiatico e anti europeo, Il vice Presidente, filo, anti e vice* sono stati classificati come aggettivi.

Gli **usi anaforici** di primo, secondo ecc. oppure di precedente, successivo, seguente e simili (si tratta di alcuni tra i vocaboli più usati nelle riprese anaforiche), usati nella combinazione ART. + AGG., sono stati trattati come aggettivi, supponendo che nella struttura soggiacente ci fosse il sostantivo corrispondente. Per esempio in : *Ho due macchine, prendo la nuova; abbiamo visto 3 film: il primo parla di X, il secondo di Y, il terzo di Z, nuova, primo, secondo, terzo*,sono stati classificati come aggettivi .

Sono stati inoltre classificati come aggettivi, nei contesti adeguati :

allora in *l'allora presidente*;
bene in *la Brescia bene*;
bis in *Berlusconi bis*;
cubo in *un metro cubo*;
doc in *leghisti doc*;

fu in *il fu Mattia Pascal*;
gay in *amore gay;*
già in *già presidente*;
in in *spiagge in*;
mezzo in *due anni e mezzo*;
nord in *nord coreano*;
PDS in *il senatore PDS;*
primo in *sono arrivato primo*;
rock in *musica rock*;
spray in *vernici spray*;
sud in *latitudine sud* (Sud è invece sostantivo in Sud America, cfr Sostantivi);
tv in *pubblicità tv*;
USA in *il dollaro USA*;
vip in *spiagge vip*.

Gli **aggettivi nelle parole sintagmatiche** (specialmente avverbiali) sono stati classificati come aggettivi, anche se non era presente nella sintagmatica un sostantivo esplicito: *per le spicce, spicce* aggettivo.

I **secondi membri di composti scritti separatamente** quali *parola chiave, famiglia tipo, madre bambina* sono stati considerati Sostantivi e non Aggettivi.

Nel caso di **etnici e altre parole** (in genere riferite ad umani) che possono essere usati sia come aggettivi sia come sostantivi, si è classificato nel modo che di volta in volta è apparso più adeguato al contesto, dal momento che non c'è una regola generale.

Gli **aggettivi a quattro uscite** che seguono o precedono un verbo, laddove esisteva un accordo di genere col soggetto della frase, sono stati classificati come Aggettivi; es. *camminano stanche*.

Aggettivi sintagmatici: sono stati classificati e trattati come sintagmatici una serie di aggettivi.

AVVERBI Avv

Il LIP si attiene alla classificazione corrente degli avverbi, pur non trovandola sempre soddisfacente in quanto non tiene conto delle diverse funzioni di parole considerate comunemente avverbi. In CoLFIS si è cercato di individuare la funzione di ogni occorrenza e di assegnarla quindi alla corretta categoria.

Quando gli avverbi ricorrono con forme alterate, sono stati riportati alla forma positiva: *lucidissimamente* è lemmatizzato come LUCIDAMENTE.

Qual quando svolge funzione di avverbio, come in *fugge qual cervo*, è stato classificato come Avverbio e lemmatizzato sotto QUALE. (Nello stesso tipo di contesto viene trattato come avverbio anche come).

Anche in *rappresenta una donna quale oggetto*, quale è classificato come Avverbio. Lo stesso vale per *un pentito della mafia salentina aveva fatto il nome di un politico quale ispiratore dell'attentato*... e per *rafforzare l'Europa quale luogo di produzione*.

In strutture come *un X quale quello di... quale* è stato classificato come Avverbio. Si osserva però che in questo tipo di strutture quale potrebbe ricorrere anche al plurale: la norma prevede quindi una controintuitiva classificazione come avverbio di alcune forme plurali. Tuttavia è apparso preferibile

conservare tale norma, in quanto l'uso di *quale* in questi contesti è del tutto parallelo a quello di *come*, che è stato classificato come avverbio.

Tutto in casi come *La porta era tutta sporca*: l'aggettivo ha chiaramente funzione avverbiale (cioè significa *La porta era completamente sporca* anziché *Tutta la porta era sporca*) ed è quindi stato classificato come Avverbio; anche in questo caso, quindi, si può avere una forma avverbiale flessa.
Dopo e **Prima** sono stati classificati come Preposizioni se reggono sostantivi (es. *Dopo cena...*) o verbi di modo non finito (infinito, participio, gerundio), quindi in proposizioni implicite (es. *Dopo mangiato, dopo aver mangiato*); sono stati classificati come Avverbi altrimenti (es. *Ci vediamo dopo; potevi pensarci prima o sette minuti dopo*).
Quanto è stato classificato come Avverbio quando introduce delle interrogative dirette (es. *Quanto fuma?*), o quando introduce il secondo termine di paragone (es. *E' tanto buona quanto è bella*); è stato considerato Aggettivo in frasi tipo *Quanti soldi hai?; Vorrei sapere quanta stoffa ci vuole per fare questo vestito*; è stato classificato Pronome in frasi tipo *Devo comprare della verdura, ma non so quanta me ne occorre; Quanti di voi parteciperanno alla gita? Non so quanto mi convenga.*
Tanto...Quanto correlativi sono stati classificati come Avverbi.
Eccetera è stato talora classificato come Avverbio e talora come Congiunzione e le forme *etc., ecc.* sono state ricondotte a ECCETERA

Avverbi sintagmatici: sono stati classificati e trattati come sintagmatici una serie di avverbi.

CONGIUNZIONI Cong.

Il LIP le lemmatizza uguali a se stesse (tranne ed e od che vengono lemmatizzate rispettivamente sotto E e O) e si attiene in genere ad una classificazione tradizionale. In Colfis si è cercato di individuare le diverse funzioni.

Distinzione tra congiunzioni e altre categorie verbali

Allora è stato classificato come Congiunzione quando è coordinato con se e quando è introduttore di frase (es. *Allora che facciamo?*); come Avverbio quando ha valore temporale (es. *Anche allora gli obiettivi erano gli stipendi*). Nell'espressione *E allora?* è stato classificato come congiunzione.
Perciò e **Pertanto** sono stati classificati come Congiunzioni.
Dunque è stato classificato come Congiunzione.
Ebbene è stato classificato come Congiunzione.
Quindi è stato classificato come Avverbio quando ha valore temporale, nel senso di poi (es. *Andai a Parigi, quindi a Bonn*), come Congiunzione quando ha valore causale (es. *Non so come sono andati i fatti, quindi non posso esprimere un giudizio*) o conclusivo (es. *Non mi piacciono i gialli, quindi non li compro*).
Insomma è stato classificato come Congiunzione quando introduce una proposizione (es. *Insomma, andiamo?*), come Avverbio quando ha il senso di praticamente, in poche parole (es. Non *occorre, insomma, che mi dilunghi su quest'argomento*) come Interiezione in uso assoluto (*es. Insomma!*).
Comunque è stato classificato come Congiunzione quando regge il congiuntivo (es. *Comunque vadano le cose, a Luglio partiremo*), e quando ha valore avversativo nel senso di tuttavia (es. E' *stata una cena improvvisata, comunque potevi avvisarmi lo stesso*), come Avverbio quando ha il senso di in ogni modo (es. *Vorrei che venissi a trovarmi comunque, Comunque io stasera non vengo*).
Quando è stato classificato come Avverbio quando introduce interrogative dirette (es. *Quando tornerai?*), Congiunzione quando introduce delle subordinate (es. *Dimmi quando tornerai; Verrò quando avrò finito questo lavoro, Quando* (nel senso di tutte le volte che, ogni volta che) *penso agli anni del liceo non posso che provare nostalgia; E' stato quella volta quando* (nel senso di nella quale, in cui) *ci siamo incontrati;*

Quando (nel senso di se, qualora) *c'è la salute, c'è tutto*; ecc.).
Da quando non è stato considerato parola sintagmatica, per cui in casi come: *Da quando lo conosci?*, *quando* è stato classificato come Avverbio mentre in *Lo conosco da quando è nato*, *quando* è stato classificato come Congiunzione.
Dove è stato classificato come Avverbio solo quando introduce interrogative dirette (es. *Dove andate?*); è stato classificato come Congiunzione quando introduce subordinate (es. *Dimmi dove hai intenzione di trascorrere le vacanze*). E' stato inoltre classificato come Congiunzione in molti casi in cui alcune grammatiche lo classificherebbero pronome relativo.
Allo stesso modo di **dove**, sono stati trattati anche **ove, donde** e **onde**.
Come è stato classificato come Avverbio nel secondo termine di paragone (es. *E' alto come Luigi*), quando ha il significato di *da, in qualità di, per esempio, in che modo* (es. *Come insegnante ti dico...*), quando introduce delle interrogative dirette (es. *Come stai?*) e in casi in cui ha il valore di coordinazione, ad esempio: *In un caso come nell'altro* e *Del loro come degli altri*; è stato invece considerato Congiunzione quando introduce delle subordinate (es. Non *so come comportarmi; Fai come fossi a casa tua; nell'inciso , come dire,; sordo come è...; Così come è stato scritto, questo libro è illegibile;*)
Appena è stato classificato come Avverbio quando ha il senso di *a fatica, a stento* (es. *Ci si vedeva appena; Facemmo appena in tempo*), di *soltanto, non di più* (es. *E' appena un ragazzo! Sono appena le otto!*), di *da poco* (es. *Sono appena arrivata*); è stato considerato Congiunzione quando introduce delle subordinate (es. *Appena arrivai mi corse incontro*).
In casi tipo *ogni volta che, tutte le volte che* e simili, il *che* (dove che ha valore temporale) è stato qualificato come Congiunzione.
Eccetera è stato classificato talora come Congiunzione e talora come Avverbio e le forme *etc., ecc.* sono state ricondotte a ECCETERA.
Nondimeno, **nonostante** e **ciononostante** sono state classificate come Congiunzioni.
Ciononostante nella forma **ciò nonostante** è stato lemmatizzato separatamente, ovvero ciò come Pronome e *nonostante* come Avverbio.

Congiunzioni sintagmatiche: sono state classificate e trattate come sintagmatiche una serie di congiunzioni.

INTERIEZIONI Inter.

Questa categoria comprende le interiezioni, i fonosimboli e le onomatopee. Nell'ambito di queste categorie sono state fatte delle scelte in parte diverse da quelle del LIP.

1.Fonosimboli
Il LIP individua una lista chiusa di fonosimboli lemmatizzati uguali a se stessi (cfr. LIP, p. 93 e Lista D, p. 531). In CoLFIS le eventuali ricorrenze di queste forme sono classificate con il codice Inter.

2.Interiezioni
IL LIP considera interiezioni quelle che solitamente vengono chiamate interiezioni secondarie: *caspita, mannaggia, accidenti*, ecc., mentre le interiezioni primarie sono classificate come fonosimboli (vedi sopra). Anche le formule di saluto quali *buongiorno, buonasera, ciao, salve, arrivederci* sono classificate nel LIP come interiezioni. Invece parole varie usate interiettivamente, quali *dai!* (verbo), *bene!* (avverbio), *guai!* (sostantivo) nel LIP sono riportate alle rispettive categorie.

In CoLFIS si è preferito categorizzare come interiezioni anche questi ultimi casi.

3. Onomatopee
Il LIP considera onomatopee le sequenze che riproducono e evocano un suono (*cucù, cri cri, bang, flap*) e non le distingue in base alla loro funzione.

In COLFIS sono state lemmatizzate uguali a se stesse e categorizzate come Interiezione. Quando hanno funzione di sostantivi sono state classificate in base alla loro funzione : ad esempio, in contesti come *il cri cri del grillo, cri cri* è stato classificato come Sostantivo (distribuzionalmente si comporta come verso in *il verso del grillo*), mentre in contesti come *il grillo fa cri cri, cri cri* è stato riconosciuto come onomatopea e quindi classificato come interiezione.

E' stata anche individuata una lista di interiezioni classificate come sintagmatiche.

NOMI PROPRI Nome

Nell'ambito di questa categoria ci siamo discostati abbastanza radicalmente dalle scelte del LIP. Il lemmatizzatore distingueva:

E = nome proprio di persona, prenome (Es. *Daniela, Alessandro*)
F = cognome
L = Nome proprio geografico (città, nazioni...)
D = Nome proprio di ditta

Nel LIP inoltre esiste una categoria N = Nome proprio non meglio specificato, che comprende nomi propri vari, esclusi antroponimi e toponimi. Anche solo osservando per campioni ciò che è classificato N nel LIP, appare evidente che le parole classificabili come nomi propri vanno al di là di quelle tradizionalmente assegnabili ai due settori dell'antroponomastica e della toponomastica.

Si è pertanto deciso di creare una categoria **NOME PROPRIO** (che accogliesse tutti i tipi di nomi propri incontrati nei testi, senza ulteriormente distinguere tra Nomi, Cognomi, Toponimi, Ditte e altro). Quindi, con programmi opportunamente scritti, abbiamo trasformato nel primo output del lemmatizzatore tutti i diversi tipi di nome proprio riconosciuti dal lemmatizzatore come Nomi personali, Cognomi, Toponimi o Ditte in **Nomi**, ovvero nome proprio. Questo ha permesso di aggirare spinosi problemi, come quello di classificare nomi di persona appartenenti ad altre tradizioni onomastiche; ad esempio, *Marco Tullio Cicerone, Giulio Cesare, Mao Tse-tung* (dove *Mao* è il cognome!), ecc.

Anche i soprannomi sono stati inseriti nell'unica categoria Nome proprio.

Nelle **INIZIALI DI NOMI PROPRI** puntate, per es. G.C., si è ritenuto che il *punto* facesse parte dell'abbreviazione e quindi non andasse lemmatizzato come punteggiatura (come fa il lemmatizzatore).

Si elencano di seguito altri casi classificati come Nomi propri (E).

1. Nomi geografici
Rientrano in questa categoria luoghi in senso lato:

- edifici e monumenti (es. il Pantheon, la Scala)
- strade (es. Nomentana, Tiburtina)
- quartieri (es. Tiburtino)
- città e paesi
- nazioni
- continenti

- pianeti
- stelle
- mari, oceani
- laghi
- fiumi
- montagne

Sono stati trattati come Nomi propri anche nomi del tipo *abitare nel bergamasco, trevigiano, vicentino, modenese* ecc.; Non sono stati considerati nomi propri *meridione* e *mezzogiorno* nel senso de *il sud*; il caso *Unione Europea* (e la sua sigla UE) è stato considerato nome proprio sintagmatico, mentre sono stati sono considerati Sostantivi i casi *unione monetaria* e *unione monetaria europea*; è stato considerato nome proprio anche *Belpaese* per *Italia*.

2. Nomi commerciali:

- nomi di ditte: es. Palmolive
- nomi di prodotti: es. Ajax
- nomi di negozi: es: Upim, Il cavallo, La boutique del formaggio

Casi come *La boutique del formaggio* sono stati considerati nomi sintagmatici.

3. Altri casi di nomi propri abbastanza diffusi:

-nomi di serie di automobili: es. Uno, Tipo, Fiesta
- nomi di squadre di calcio
- nomi di partiti politici
- nomi di enti e associazioni
- nomi di testate e titoli di opere dell'ingegno e di trasmissioni radiotelevisive*: il Messaggero, i Promessi Sposi, la Traviata, Volare, Il sabato del villaggio, Forum, Samarcanda, TG3...*
- nomi di autostrade: es. *A24*
- nomi di ospedali, ecc. Nel caso di *Clinica S. Lucia*, *clinica* è stato lemmatizzato come sostantivo e S. Lucia come nome proprio sintagmatico.
- nomi di organismi economici e finanziari: es. *cipe, consob, bankitalia, istat*

Non sono stati considerati nomi propri i nomi di feste, i nomi di secoli ed epoche; il *seicento*, *l'ottocento* ecc. sono dunque considerati Numerali.

4. Nomi di ditte

Nel LIP esisteva anche una categoria D, che indicava Nomi di ditte. Poiché la categoria *Ditta* non è una categoria grammaticale, questa categoria è stata eliminata dal novero della categorie considerate per CoLFIS. Ciò che il lemmatizzatore classificava come Ditta è stato ricategorizzato come Nome Proprio.

Si è tenuto presente che alcune ditte si chiamano con il cognome del proprietario o fondatore: es. Ferrari, Olivetti. Inoltre, anche il prodotto tipico della ditta può essere chiamato con tale nome: una Ferrari, una Olivetti (macchina da scrivere). In questi casi sono state separate le ricorrenze come sostantivo da quelle come nome proprio (di ditta o di persona, che confluiscono entrambi nella categoria generale di nome proprio). Quindi:

la Ferrari ha sede a Maranello, il mitico Enzo Ferrari,la Ferrari ha vinto a Imola: Ferrari è stato considerato nome proprio;
la Ferrari di Mansell è ferma ai box: Ferrari è stato considerato un sostantivo;
una vecchia fiat rossa: Fiat è stato considerato un sostantivo;
una vecchia panda rossa: Panda è stato considerato un nome proprio;
una Fiat Panda / Lancia Thema / Mercedes 500: Fiat è stato considerato sostantivo e *Panda* nome proprio, e analogamente nei restanti casi.

5 Altri casi

Nei cognomi col genitivo sassone, `s è parte del nome, per cui *Christie's* è stato considerato nome proprio.

Nei nomi comuni la forma con genitivo sassone è stata invece considerata una forma flessa del lemma corrispondente: *house's* è stato dunque ricondotto a HOUSE.

Cognomi tipo O'Hara e McDonald sono stati lemmatizzati come nomi propri uguali a se stessi. Nel caso in cui si è trovato scritto Mac Donald con due parole separate, il cognome è stato trattato come un nome sintagmatico i cui singoli membri sono stati categorizzati entrambi come E.

Casi tipo *Tangentopoli, Vaticano, Pentagono, Translantico, Quercia, Edera, Lega, Polo, Rifondazione, Garofano, Scudo crociato, Bianco fiore, Fiamma, e Verdi* come partito sono stati trattati tutti come nomi propri.

In *la Statale di Milano*, Statale è stata considerata nome proprio.
In *laureata alla Cattolica*, Cattolica è stata considerata nome proprio.
In *dipendente delle Nord* (nel senso di ferrovie Nord), Nord è stato considerato nome proprio.
G7 è stato lemmatizzato come nome proprio.
In un *kalashnikov AK47*, AK47 è stato considerato nome proprio in quanto si tratta del nome di un tipo di arma.

6. Nomi propri sintagmatici

Uno dei principali tipi di intervento di correzione manuale ha riguardato i nomi propri sintagmatici. Per questi nomi, era interessante la frequenza dell'unità politematica e dei singoli membri. Il lemmatizzatore automatico non riconosceva i nomi sintagmatici e quindi si è dovuto ricorrere all'intervento manuale.

I principali tipi di nomi propri sintagmatici sono toponimi composti e cognomi composti (es. *Castel di Sangro, De Mauro, Di Pietro*).

6.1 Cognomi

I cognomi italiani composti vanno classificati come Nomi propri sintagmatici, per es. *Di Pietro, Del Turco, Dalla Chiesa*. Non abbiamo sciolto nei singoli componenti le preposizioni articolate incontrate (come negli ultimi due cognomi qui citati), ma le abbiamo trattate come un tutt'uno. Per es. in *Del Turco* si ha:

DEL TURCO : nome proprio sintagmatico i cui componenti sono Del nome proprio e Turco nome proprio.

Si ricordi che il lemmatizzatore scorpora sempre *Del* in DI preposizione e IL articolo . Nel caso in cui *Del* si fosse presentato come componente di un cognome, i due componenti sono stati riaccorpati.

Anche i cognomi doppi, come per es. Sforza Pallavicini sono stati trattati come nomi propri poliramatici.

Allo stesso modo sono stati trattati anche i cognomi doppi di donne, in cui uno dei cognomi è quello del marito .

I cognomi stranieri sono stati trattati alla stessa stregua di quelli italiani.

6.2 Nomi di battesimo sintagmatici

Sono stati lemmatizzati come sintagmatici solo i nomi propri composti molto diffusi: per es.: *Anna Maria, Maria Teresa, Pier Paolo, Giovanni Paolo II, Pio XII* ecc. ma non *Alessandro Carmine, Silvia Piera* ecc.

6.3 Nomi nobiliari

I nomi nobiliari non sono stati considerati sintagmatici, quindi i singoli componenti sono stati lemmatizzati uno per uno: in *duca di Wellington*,duca è stato lemmatizzato come sostantivo, di come preposizione e Wellington come nome proprio.

6.4 Soprannomi

I sostantivi e gli aggettivi usati per designare in modo specifico un personaggio sono stati classificati come nomi propri sintagmatici, per es. *Il Temporeggiatore, Il Venerabile G, il Molleggiato G, l'Innominato G, la Pantera di Goro, Sua Emittenza*.

Eventuali nomi storpiati sono stati ricondotti alla forma corretta, es. *Berlusca* in BERLUSCONI .

6.5 Nomi storici

In casi come *Erasmo da Rotterdam, Leonardo da Vinci, Alberto da Giussano, Lorenzo il Magnifico* si è classificato il tutto come nome proprio sintagmatico, ma con da come preposizione e il come articolo.

6.6 Nomi di divinità

Sono stati lemmatizzati come nomi propri anche i casi di nomi/appellativi di divinità, come *Madonna, Cristo, Dio, Signore, Vergine, Allah, Messia*. Non sono stati trattati invece come nomi propri i seguenti casi: *papa, santo padre, sua santità, padreterno, chiesa, islam, imam*.

6.7 Toponimi

I toponimi sono stati classificati come nomi propri e lemmatizzati uguali a sé stessi.

Es. *Palma di Montechiaro*, l' uscita della lemmatizzazione automatica era *palma* sostantivo, *di* preposizione e *Montechiaro* sostantivo.

Noi abbiamo ricomposto l'unità e l'abbiamo lemmatizzata uguale a se stessa categorizzandola come nome proprio sintagmatico e categorizzando ogni singolo componente.

Sono stati considerati come nomi propri sintagmatici anche i casi di *Sud America, America del Sud, Centro America, Nord Europa, Irlanda del Nord , Africa del Nord, Germania est* , ecc.

6.8 Nomi di strade (allo stesso modo i nomi di Piazze, larghi ecc.)

Sono stati considerati nomi propri sintagmatici e i singoli componenti sono stati lemmatizzati: es. *Via Oreste de Gaspari*: è stato classificato come nome proprio sintagmatico con via classificato come Sostantivo, Oreste come nome proprio, *De* come nome proprio e *Gaspari* come nome proprio.

6.9 Nomi di opere

Se sintagmatici, sono stati classificati come tali. Esempio*: Il sabato del villaggio*, dove si ha *il* articolo, *sabato* sostantivo, *del* preposizione (con DE e -L articolo) e *villaggio* sostantivo.

Sono stati trattati come nomi propri i nomi dei seguenti testi: *Bibbia, Corano, Vangelo, Gazzetta Ufficiale*.

Non sono stati considerati nomi propri : *Costituzione, Nuovo Concordato, Codice civile*.

6.10 Nomi di associazioni, movimenti, ecc.

Anche se derivano da un uso metonimico sono stati classificati come nomi propri: per es. *Casa Bianca, Nazioni Unite, Mani pulite, Botteghe oscure* ecc.

7. Nomi di mostre, manifestazioni e simili

Di fronte al proliferare di titoli molto lunghi (magari anche con sottotitoli), si è deciso di considerare come sintagmatici solo i titoli di mostre, manifestazioni e sim. periodiche (tipo *Pitti uomo, Biennale di Venezia* e sim.). Quindi, i titoli di manifestazioni non cicliche, non permanenti e occasionali non sono stati considerati sintagmatici. Un ulteriore criterio di scelta è dovuto alla presenza di una struttura frasale piuttosto che sintagmatica (per es. in *La moda a Milano*).

7. 1. Nomi di feste, periodi storici, ecc.

Natale, Pasqua, Fascismo, Rinascimento, fascio, ventennio, capodanno, immacolata concezione, tutti i santi, non sono stati considerati nomi propri.

7.2. Nomi di santi

Tutte le forme di santo (Santo, San, S.) che compaiono seguite da un nome proprio sono state trattate come polirematiche in cui il componente san, santa, è categorizzato come sostantivo e il nome che segue come nome proprio. Ad esempio, *San Francesco* è un nome proprio sintagmatico in cui *san* è un sostantivo e *Francesco* un nome proprio.

Quando la parola Santo ecc. è abbreviata S., il punto è stato lemmatizzato come parte dell'abbreviazione e non come uno dei segni di interpunzione.

Non sono stati classificati come nomi propri le seguenti classi di nomi:

il **sessantotto, quarantotto** e sim., trattati come come Numerali;
nomi di **istituzioni** e **organismi militari**, es*. finanza, carabinieri, fiamme gialle, milite ignoto* (quest'ultimo è nome proprio se indica il monumento);
nomi di **monete** es. *dollaro, rublo, ecu*, di titoli es. *future, bot, btp,* di tasse es. *iva, irpef, irpeg*;
nomi propri di persone usati come antonomasie: *si credeva un ercole, era un giuda*.

NUMERALI Num.

Il LIP fa rientrare in questa categoria tutti i numeri, cardinali e ordinali, senza distinguere l'uso aggettivale o sostantivale degli ordinali.

Il trattamento di questa categoria è quello che ha posto più problemi di adattamento del lemmatizzatore automatico. Infatti, il lemmatizzatore, tarato sul parlato, riconosce i numerali come tali solo se sono scritti

in forma ortografica, ad esempio come millenovecentonovantasei. Nei testi scritti, però, la maggior parte dei numeri non ricorre in forma ortografica, ma in forma di cifre. In tal caso, il lemmatizzatore produceva un output a caso.

Abbiamo quindi assegnato automaticamente il codice NU a tutte le forme costituite interamente di cifre. Questa scelta ha implicato una perdita di informazione, in quanto non si distingue tra funzione sostantivale e funzione aggettivale dei numerali; tuttavia, ha permesso di risparmiare molto lavoro di correzione manuale e di rendere il risultato della nostra lemmatizzazione confrontabile con quello del LIP.

Gli **Aggettivi Numerali Ordinali** sono stati trattati come Aggettivi e non come Numerali.

La parola **mila**, nel caso in cui sia stata scritta staccata dal numero che la precede (apparendo quindi su un'altra riga nell'output del lemmatizzatore), è stata classificata Numerale e non Aggettivo e non unita al numero che la precede.

Anche **milione**, **miliardo** ecc. sono stati classificati come Numerale (non però *paio, decina, dozzina, centinaio, migliaio* ecc. che sono stati classificati come Sostantivi).

1. I Numeri romani sono stati classificati come numerali.

2. Altri casi :

- in *un anno e mezzo, un* è Numerale
- in *più di una volta, una* è Numerale
-i numeri abbreviati alle ultime due cifre come '97 sia con apice sia senza, sono stati classificati come numerali e lemmatizzati uguali a se stessi.
-Nel *cinque e seicento: cinque* è stato classificato come Numerale CINQUECENTO e *seicento* come Num. SEICENTO.

PREPOSIZIONI Prep.

Sono stati adottati gli stessi criteri del LIP, sia per le semplici sia per le articolate. Le preposizioni semplici sono state lemmatizzate uguali a se stesse e le articolate sdoppiate, di modo che l'articolo contenuto nella preposizione venisse lemmatizzato sotto la forma singolare maschile dell'articolo e la preposizione uguale alla preposizione semplice. Ad esempio, *nel* è lemmatizzato sotto IN, e la sua parte *-l* è lemmatizzata sotto IL.
La preposizione apocopata *de'* è stata classifica come forma di DI.
Le preposizioni articolate francesi **au, du** sono state classificate come preposizioni senza scorporare l'articolo.

Dopo e **Prima** sono stati classificati come Preposizioni se reggono sostantivi (es. *Dopo cena...*) o verbi di modo non finito (infinito, participio, gerundio), ossia in proposizioni implicite (es. *Dopo mangiato, dopo aver mangiato*); sono invece stati classificati come Avverbi se usati in senso assoluto (es. *Ci vediamo dopo; potevi pensarci prima*).

Nell'espressione via fax.... **via** è stato classificato come preposizione.

Nelle espressioni sintagmatiche che finiscono con preposizione articolata l'articolo è stato considerato esterno alle medesime; es. *nei confronti delle* è classificata come preposizione sintagmatica NEI CONFRONTI DI, in cui *nei* si riconduce alla preposizione IN e *–i* all'articolo IL, e infine *confronti* al sostantivo CONFRONTO.

E’ stata anche individuata una lista di preposizioni sintagmatiche.

PRONOMI Pron.

1. Pronomi possessivi

Sono stati considerati pronomi solo se preceduti da articolo e non seguiti da nome. Sono lemmatizzati, come gli aggettivi, sotto il maschile singolare.

2. Pronomi dimostrativi

Sono lemmatizzati, come gli aggettivi, sotto il maschile singolare.

3. Pronomi relativi

Il quale, i quali, la quale, le quali

Sono lemmatizzati scomposti (il = articolo, quale = pronome) nel LIP, ma in Colfis abbiamo rimandato a **IL QUALE**, trattandolo quindi come una parola sintagmatica.

Nel caso di **DEL QUALE** e sim., si è lemmatizzato sdoppiando la preposizione articolata in di e -l articolo, più quale ricondotto a IL QUALE.

Cui è lemmatizzato uguale a se stesso.

Nel caso di un intreccio di sintagmatiche che comprendesse il pronome IL QUALE, si è lemmatizzato nel modo seguente:
es. *davanti al quale: davanti a*, preposizione sintagmatica in cui davanti è un avverbio in sintagmatica, *al* è preposizione articolata in sintagmatica da sdoppiarsi in *a* preposizione semplice e *–l* pronome e *quale* pronome nel pronome sintagmatico IL QUALE .

Che cosa è stato trattato come pronome sintagmatico anche in frasi interrogative sia dirette che indirette e i due componenti sono stati analizzati rispettivamente come aggettivo *che* e sostantivo *cosa*.

4. Pronomi personali

Nel LIP i pronomi personali, nelle forme sia toniche sia atone, sono lemmatizzati sotto le forme del nominativo e le terze persone al nominativo maschile. I lemmi sono quindi IO, TU, EGLI, NOI, VOI, ESSI. Pertanto, *me* è stato lemmatizzato sotto IO, lo sotto EGLI e gli come forma di EGLI o di ESSI .

Questo tipo di scelta appariva insoddisfacente, perché non permette di recuperare la distinzione tra pronomi di terza persona animati e inanimati (sia *lo* che *lui*, ad esempio sono classificati comunque come forme di EGLI). Tuttavia, con le risorse umane e finanziarie a disposizione, era impossibile intervenire con correzioni manuali su questa categoria. Una riclassificazione delle diverse ricorrenze dei pronomi per genere, numero e animatezza, da operarsi sull'output della lemmatizzazione, potrebbe essere svolta come lavoro autonomo in un secondo momento (chiunque intendesse collaborare è invitato a contattarci).

Come nel LIP, anche in COLfis il **si** riflessivo è stato riportato di volta in volta a EGLI o ESSI e il ***si*** impersonale ha costituito un lemma a sé. Sono lemmi separati anche i pronomi **ci** e **ne**.

I pronomi di cortesia sono stati trattati come gli altri pronomi personali: *lei* è stato lemmatizzato come pronome e ricondotto a EGLI, ecc.

Coloro è stato classificato come pronome e ricondotto al lemma COLUI, **Costoro** è stato classificato come pronome e ricondotto al lemma COSTUI.

Questi e **quegli**, quando hanno valore singolare, sono stati ricondotti ai lemmi QUESTI e QUEGLI.

Sono stati classificati come pronomi, ovviamente nei casi in cui ne svolgevano le funzioni, anche le seguenti parole:

alcuni, alquanto, altro, chi, chichessia, chiunque, ciascuno, ciò, colui, cosa, costui, granché, molteplice, molto, nessuno, niente, nulla, ognuno, parecchio, poco, qualcosa, qualcuno, quale, qualunque, quanto, quegli, quello, questi, questo, stesso, tale, tanto, troppo, tutto, uno.

Trattamento di **ci**, **si**, e **vi**.

In fase di correzione, sono state incontrate alcune difficoltà di lemmatizzazione riguardo a questi pronomi. Si è deciso di operare nel seguente modo:

In costruzioni del tipo *ci si rompe una gamba*,ci è stato lemmatizzato SI e *si* SI, dal momento che *ci* è frutto di una regola morfofonologica che cambia *si* in *ci* quando due si si trovano di seguito (uno impersonale o passivante e l'altro riflessivo); si rappresenterebbe l' oggetto (per così dire) del ci, che ha, in quanto impersonale, la funzione di soggetto.

Dunque, *ci* può essere forma di cinque lemmi diversi:

a) Pronome CI in frasi tipo *ci penserò, ci dormirò sopra*;
b) Avverbio CI in frasi del tipo c'è, *ci vado,* dal momento che ha valore locativo;
c) Pronome NOI in frasi del tipo *ci vestiamo*, *ci stiamo impegnando;*
d) Pronome SI in frasi del tipo *ci si rompe una gamba, ci si arrabbia*.
e) Pronome EGLI in frasi del tipo *con lui ci parlo bene, a lui ci dico sempre la verità*.

Parallelo a ci è il comportamento di *vi*:

a) Pronome VI in *vi farò caso*;
b) Avverbio VI in *vi vado*;
c) Pronome VOI in *vi dico*.

Analogamente **si** può essere forma di tre lemmi diversi: EGLI, ESSI, SI.

a) Pronome riflessivo ricondotto a EGLI in frasi tipo *è la prima volta che il pool milanese si spacca, Mario si arrabbia*;
b) Pronome riflessivo ricondotto a ESSI in frasi del tipo *si ruppero una gamba, Mario e Paolo si arrabbiarono*;
c) Pronome nei dativi etici, ricondotto a EGLI o a ESSI, in frasi del tipo *Mario si beve un bicchiere di latte*;
d) Pronome impersonale passivante, ricondotto a SI, in frasi tipo *in questa piazza si incontra tanta gente, si è visto Ugo, Bisogna farsi coraggio, Si riempie la vasca*;
e) Pronome negli infiniti pronominali terminanti in *-si* (che si possono sciogliere in una locuzione con *ci si*...) ricondotto a SI; per esempio, *questo permette di districarsi*... (*questo permette che ci si districhi*...); bisogna *liberarsi*... (*bisogna che ci si liberi*...).

In generale, da un punto di vista pratico, quando il *si* italiano è traducibile con *on* francese, *one* inglese e *man* tedesco, si è considerato *si* pronome e lo si è ricondotto al lemma SI.

Quando *si* è coreferente con *chi* o altri indefiniti (es: *chi si accorge*...), è stato lemmatizzato come pronome EGLI.

Infine **sè** con referente singolare è stato lemmatizzato sotto EGLI, con referente plurale sotto ESSI, come generico o impersonale è rimasto uguale a sé stesso SE':

a) *Lo fa per sé*: pronome EGLI ;
b) *Lo fanno per sé*: Pronome ESSI ;
c) *Essere se stessi*: Pronome SE', con stesso aggettivo ricondotto a EGLI;

I clitici sono stati lemmatizzati come gli altri pronomi personali, ovvero separati dalla forma verbale con cui ricorrono e rimandati al nominativo a cui si riferiscono; ad es., *parlami* è lemmatizzato *parla* verbo PARLARE e *-mi* pronome IO.

Anche nelle forme doppie il trattamento è uguale a quello dei pronomi non clitici. I clitici composti glielo, melo, etc. sono stati considerati come due pronomi, con ogni membro della coppia lemmatizzato sotto il nominativo corrispondente. *Glielo* è quindi stato ricondotto a *–gli* EGLI e *–lo* EGLI.

Nel caso di clitici con raddoppiamento si è deciso di non correggere sistematicamente l'output del lemmatizzatore (che non registra i raddoppiamenti). In *dammene* si avrà quindi Verbo DARE , pronome *-me* lemmatizzato IO e Pronome *-ne* lemmatizzato NE.

I casi di clitici attaccati a non-verbi, come *eccoci, eccone, rieccola* e sim., sono stati trattati come segue: *eccoci* è stato ricondotto a ECCO avverbio e *-ci* pronome NOI, *rieccola* è stato ricondotto a RIECCO avverbio e *–la* pronome EGLI. Nel caso di infiniti sostantivati, quali *l'anarchico moltiplicarsi*, *moltiplicarsi* è stato trattato come sostantivo MOLTIPLICARE e *–si* come pronome SI.

5. Altri casi :

In *se ne vedono troppi di drammi*, *troppi* è stato classificato come pronome;
In *da un centro ortopedico all'altro*, *altro* è stato classificato come pronome;
In *dire la sua, sua* è è stato classificato come pronome;
In *Forlani invita i suoi alla calma, suoi* è stato classificato come pronome.
In *da poco, da molto* e simili*, poco, molto* sono stati classificati come pronomi, ma in *contare poco/niente*, *poco* e *niente* sono stati considerati avverbi.
In niente *di male, nulla di nuovo, niente di meglio, niente* e *nulla* sono stati considerati pronomi; ma in *niente scuole, niente* modifica il nome e pertanto è stato considerato aggettivo.
In *non ha nulla/niente a che fare/vedere con...*, si è considerato *niente* pronome e *che* congiunzione;
In *un bel niente, un bel nulla* e sim., si è considerato *niente* pronome e *bel* aggettivo ricondotto a BELLO;
In *nient'altro* si è considerato pronome *niente* e aggettivo *altro*;
In *è quanto di peggio tu possa fare*, si è considerato *quanto* pronome;
In *tutti quanti, tutti* è stato considerato pronome e *quanti* è stato considerato aggettivo;
In *vennero tutti e due, c'erano tutt'e tre i bambini,tutti* è stato considerato pronome e ricondotto a TUTTO e *due* numerale;
In *c'è qualcun altro*, *qualcun altro dei presenti*, *qualcun* è stato considerato pronome e ricondotto a QUALCUNO e *altro* è stato considerato aggettivo;

In *cos'altro posso fare?cos'* è stato considerato pronome e lemmatizzato come COSA e *altro* aggettivo;
In *con il PDS e con chi altro ci sta*, si è considerato *chi* pronome e *altro* aggettivo;
In *chissà chi altri c'era*, si è considerato *chi* pronome e *altro* aggettivo;
In *quale che sia, quali che siano,* si è considerato *quale* pronome e *che* pronome;
In *nel qual caso*, si è considerato *quale* aggettivo;
In *modificato quel tanto che basta*, si è considerato *quel* aggettivo, *tanto* pronome e che pronome;
In frasi del tipo *ti piace? - tutt'altro, tutt'altro* è stato considerato avverbio sintagmatico in cui *tutto* è aggettivo e *altro* è pronome.

PUNTEGGIATURA Punt.

Per punteggiatura si intendono i seguenti segni: . (punto), , (virgola), ; (punto e virgola), : (due punti), ! (punto esclamativo), ? (punto interrogativo), i diversi tipi di parentesi (quadre, tonde, graffe) e i diversi tipi di virgolette (", ', < , >).

Sono stati lemmatizzati uguali e se stessi.

In caso di elenchi con a) b) c) e 1) 2) e simili, si è lemmatizzato ogni elemento separatamente

SIMBOLI Simb.

Questa categoria non esiste nel LIP, perché nel parlato non possono ovviamente ricorrere simboli. E' stata introdotta per lemmatizzare simboli come $, £, %, e in generale tutti i simboli che si trovano su una tastiera, compresi i simboli delle quattro operazioni, che sono stati lemmatizzati uguali a se stessi.

Il caso di & è stato trattato come simbolo e non come congiunzione.

SINTAGMATICHE (PAROLE) Poli

Un'espressione sintagmatica o i tempi composti dei verbi possono contenere parole che non appartengono alla sintagmatica o al verbo, ma che possono essere elementi singoli di una frase o elementi di un'altra sintagmatica.

In questi casi, sono state introdotte due informazioni:

a) ... (tre puntini di sospensione), a significare per es.che in mezzo a ci ... essere (sintagmatica ESSERCI) vi sono altri elementi (può, deve ecc.);

b) #1 (cancelletto e numero), a significare che si tratta di un inserimento all'interno della sintagmatica, i cui componenti possono essere a loro volta elementi singoli o elementi di un'altra frase; con il numero è possibile quindi individuare eventuali altre sintagmatiche annidate, che avrebbero una numerazione sequenziale diversa.

Pertanto in non ci può essere, si ha ci…essere e può#1, ad indicare che si tratta di una forma all'interno di una polirematica.

Nel caso di è, come abbiamo detto, quasi del tutto ignorata si ha invece è…ignorata#1 Verbo, come Congiunzione, abbiamo detto#2 Verbo, quasi Avverbio e del tutto#3 Avverbio.

SOSTANTIVI Sost.

1 Femminili

1.1 Femminili animati

Il lemmatizzatore assegnava tutti i sostantivi alla loro forma singolare. Le ricorrenze dei sostantivi femminili animati, anche quando esista una forma maschile corrispondente, erano quindi riportate a lemma; ad esempio, *organizzatrice* è lemmatizzato uguale a sé stesso.

Si è adottato lo stesso criterio anche in CoLFIS, correggendo manualmente laddove il lemmatizzatore riportava al maschile; per es., attrice ATTORE.

1.2 Altri lemmi femminili

In casi come *le politiche, le europee, le comunali* (sottinteso elezioni) si è deciso di lemmatizzare al femminile plurale ('pluralia tantum').

2.Sostantivi alterati

I sostantivi alterati, come gli aggettivi alterati, sono stati lemmatizzati inizialmente come lemmi autonomi. In una seconda fase di correzione manuale si è aggiunto ad ogni alterato (aggettivi e sostantivi) un rimando al lemma positivo corrispondente.

I sostantivi in -issimo, quali *partitissima, finalissima, governissimo* sono stati lemmatizzati come tali, e non ricondotti ai sostantivi *partita, finale e governo.*

3. Sostantivi più comuni al plurale

I sostantivi usati più comunemente al plurale, come *mutande, pantaloni*, ma il cui singolare esiste, sono stati lemmatizzati sotto il singolare.

4. Altre categorie sostantivate

Qualsiasi forma funzionante come sostantivo è stata categorizzata come sostantivo (aggettivi sostantivati, infiniti sostantivati e altre parti del discorso sostantivate, esclusi i numerali che hanno un trattamento a sé).

5. Nomi di colore

Per alcuni composti indicanti nomi di colore, si è deciso di operare nel seguente modo: *Una gonna grigio ferro, Un gilè grigio perla, Una giacca color rosso, ferro, perla, rosso* sono stati considerati sostantivi e grigio aggettivo.

In *Dei ricami avorio*, *avorio* ha la funzione di Aggettivo mentre in *Dei ricami color avorio*, *avorio* ha la funzione di sostantivo.

In *vestire di/in bianco, bianco* è stato considerato sostantivo.

In *di un bel blu elettrico, blu* è stato considerato sostantivo e elettrico aggettivo.

6. Composti con trattino

I composti scritti con trattino sono stati trattati come due parole separate. Ad esempio, *baby-estorsori* è trattato come *Baby* Sostantivo BABY, - come punteggiatura, *estorsori* come sostantivo ESTORSORE.

Fanno eccezione i casi in cui uno dei membri del composto con trattino non è un lemma semanticamente autonomo, ma un modificatore. Quindi *catto-leghista* è stato trattato come un unico lemma, ossia CATTO-LEGHISTA.

I composti scritti staccati sono stati lemmatizzati separatamente: *decreto salva potenti* , sostantivo, verbo, sostantivo;

Invece composti verbo-nome con trattino sono stati riportati a lemma: *porta-agenda*, PORTA-AGENDA sostantivo.

7. Vari tipi di prefissati

Un caso particolare è costituito dai **prefissi "factored out"**. Esempio: in *pre e* post-scuola, *pre* e *post* sono stati considerati aggettivi e lemmatizzati uguali a se stessi.

Altri prefissati scritti con trattino sono stati lemmatizzati unitariamente: *anti-Cossiga* è assegnato al lemma aggettivale ANTI-COSSIGA.

8. Casi particolari

Cosa, quando introduce interrogative dirette o indirette, è pronome. Nella sintagmatica che cosa è invece trattato come sostantivo , mentre il che è un aggettivo .

In *ieri mattina /pomeriggio / sera, mattina, pomeriggio e sera* sono classificati come sostantivi.
In casi tipo *la serie A, A* è stato considerato sostantivo.
In *i nostri 007, 007* è stato considerato sostantivo.

9. Casi che potrebbero essere interpretati come numerali

In *il primo della classe, il primo della lista, primo* è in stato categorizzato come un sostantivo.
In *l'una di notte* e sim., *una* è stato classificato come sostantivo.

Paio, Decina, Dozzina, Centinaio, Migliaio e sim.: siccome non denotano quantità precise, sono stati qualificati come sostantivi e non come Numerali.

10. Dialettalismi

Le forme di dialetti italoromanzi (esclusi il sardo e il friulano che sono tradizionalemnte considerate lingue a sé) sono state riportate al lemma italiano etimologicamente corrispondente.

Per esempio:

Totò Riina detto *`u curtu: `u* è stato riportato a IL e *curtu* a CORTO.
Lo Spedale degli Innocenti: Spedale è stato riportato a OSPEDALE.
Marescia' è stato trattato come MARESCIALLO.
Il Senatur: Senatur è stato trattato come SENATORE.

Si è lemmatizzato con asterisco anche il dialettalismo *dane'* in DANARO.

11. Esotismi

Il LIP lemmatizza come Esotismi oltre 120 parole che per la stragrande maggioranza sono e funzionano come sostantivi (es. aikido, bagarre, body,..). Alcuni sono molto acclimatati in italiano (es. box, club, depliant, extra, fax, ...). Inoltre, la classificazione come esotismo non era coerentemente applicata dal lemmatizzatore: lo stesso LIP classifica come Sostantivi e non come Esotismi, per es., pullman e ticket.

Dal punto di vista logico, l'essere esotismo o meno è una proprietà indipendente dalla categoria grammaticale di una parola. Nella base di dati ideale, questa informazione andrebbe in un campo che indica l'etimologia della parola, non in quello che indica la categoria grammaticale. Quindi nella correzione della lemmatizzazione è stata abolita la categoria Esotismo e ogni parola straniera è stata assegnata alla categoria grammaticale appropriata. Poiché questa categoria consiste quasi sempre di Sostantivi, il codice O indicante gli esotismi è stato automaticamente trasformato nel codice S dei sostantivi prima di fornire l'output del lemmatizzatore alla correzione manuale. Questo ha ridotto il numero di interventi di correzione manuale.

I sostantivi inglesi con **genitivo sassone** vengono lemmatizzati come forme del lemma corrispondente. Esempio: book's diventa BOOK, mentre, quando il genitivo sassone è parte di un nome proprio, esso fa parte del lemma; es., CHRISTIE'S.

Le **locuzioni in lingue straniere** sono state trattate come le sintagmatiche italiane, coi singoli lemmi riportati alla forma di citazione della lingua corrispondente (i plurali ai singolari, i passati agli infiniti ecc.). Per esempio, nelle locuzioni latine *a priori*, *a posteriori* i costituenti sono classificati secondo la categoria latina: *a* come preposizione, *priori* come aggettivo.

Minimum tax non è sintagmatica e i singoli lemmi sono stati riportati a MINIMUS aggettivo e a TAX sostantivo.

12. Sostantivi sintagmatici

Si è introdotto il tipo dei sostantivi sintagmatici, specialmente quando essi si incrociano con gli esotismi e quando ci sono più ortografie possibili. Per es., *tira e molla, faccia a faccia, tête-à-tête, week-end* ecc. In casi come *week-end* è possibile trovare sia la forma col trattino, sia quella senza trattino (*week end*).

VERBI Verb.

Sono stati lemmatizzati sotto l'infinito. Il lemmatizzatore automatico registra in lemmi separati le occorrenze dei verbi essere, avere, venire, andare come ausiliari: ad esempio, sono partito é trattato come *sono* verbo ESSERE e *partito* come verbo PARTIRE. In COLFIS, i tempi composti sono stati lemmatizzati come una unica unità. Questa operazione è stata svolta in una fase di correzione manuale.

Gli infiniti con clitico, anche se pronominali, sono ricondotti al lemma senza clitico: *arrabbiarsi* a ARRABBIARE.

Gli infiniti in funzione di sostantivi sono stati classificati come sostantivi. Nei casi in cui l'infinito in funzione di sostantivo fosse unito a un clitico, il clitico è stato scorporato e lemmatizzato come pronome e l'infinito classificato come sostantivo; ad esempio, il farlo, *farlo* FARE sostantivo e *-lo* pronome EGLI.

1. Participio passato

Il participio passato ha posto molti problemi. In generale, si è cercato di attenersi alla regola seguente: è considerato aggettivo o sostantivo nel caso in cui distribuzionalmente sia sostituibile con un aggettivo o un sostantivo, in caso contrario è classificato come verbo. Cfr. quanto detto in **Aggettivi** a proposito della distinzione tra uso aggettivale e uso participiale dei participi passati.

2. Sintagmatiche verbali

I casi dei verbi come *prendersela*, *andarsene* etc. sono stati lemmatizzati come forme sintagmatiche del lemma PRENDERSELA.

In generale, sono stati classificati come sintagmatici verbali i verbi che terminano col clitico *-la* (es. *cavarsela, avercela, giocarsela* ecc.), *-ci* (es. *esserci, volerci* ecc.) o *-ne* (es. *andarsene, saperne* ecc.).

PAROLA SCONOSCIUTA Sconosciuto

Nel caso di parole appartenenti a lingue per noi non immediatamente identificabili o di parole di cui non si sapeva riconoscere la categoria grammaticale di appartenenza si è assegnato arbitrariamente la categoria X. Per esempio, Ruan ling yu, titolo di un film cinese, è stato lemmatizzato come Nome Proprio in sintagmatica in quanto titolo di film, e i suoi componenti, uno per uno, uguali a se stessi sotto la categoria X.

Gli errori di stampa sono stati corretti. Inoltre, le lettere accentate non sono rappresentate con il loro carattere ASCII, ma con la lettera corrispondente senza segni, seguita da apostrofo. Quindi, è viene rappresentata come e'. La necessità di questa modifica si è presentata perché alcune parole, per es. quelle francesi tipo e'lite, venivano scorrettamente spezzate su più righe dal lemmatizzatore, che interpretava gli apostrofi interni alla parola come demarcatori di fine parola.