# LINEE GUIDA

Lo scopo della lemmatizzazione è di ridurre ogni parola presente nel testo alla sua forma citazione, ovvero alla corrispettiva entrata del dizionario.

## 1. FORMATO

Il lavoro si svolge su fogli di calcolo. I token e le parole sintattiche (vedi Sezione 3) sono nella colonna B, mentre i lemmi nella colonna C. Ai lemmi vengono associate delle feature da aggiungere nella colonna F e H, mentre nella colonna K va aggiunta, se presente, la forma antica di eventuali lemmi antichi e la forma alterata degli alterati: alcuni lemmi antichi non si trovano nei dizionari moderni, altri hanno un'entrata che rimanda al lemma attuale. Si consiglia di consultare il dizionario De Mauro in cui i lemmi obsoleti o letterari hanno un rimando alla versione attuale: vedi, ad esempio, https://dizionario.internazionale.it/parola/respondere.

ATTENZIONE: nei file annotati dei capitoli 1, 8 e 23 mancano i lemmi nella forma alterata nella colonna K. Chiara Febbraro è incaricata di aggiungere tali lemmi → FATTO

## 2. DIVISIONE IN FRASI

Le frasi sono già segmentate. In caso di dubbi, contattare Rachele Sprugnoli prima di procedere con l'annotazione indicando i punti problematici.

In generale, il periodo NON si spezza:

- dopo il *verbum dicendi*;
- dopo punti esclamativi, interrogativi o puntini di sospensione se questi sono seguiti da una lettera minuscola.

Al contrario, si divide dopo punto fermo o altra interpunzione forte ( ! ? …) se seguiti da maiuscola:

- split: "Venimmo a lei: o anima lombarda, come ti stavi altera e disdegnosa e nel mover de li occhi onesta **e tarda! // Ella** non ci dicea alcuna cosa, ma lasciavane gir, solo sguardando a guisa di leon quando si posa."
- no split: "Quivi mi cinse sì com' altrui piacque: oh **maraviglia! ché** qual elli scelse l' umile pianta, cotal si rinacque subitamente là onde l' avelse."

## 3. TOKENIZZAZIONE E DIVISIONE IN PAROLE SINTATTICHE

Le frasi sono già tokenizzate, cioè ogni unità minima (quali parole e punteggiatura) è già su righe diverse (una riga per ciascun token). Un token, però, può essere diviso in due o più parti corrispondenti a parole sintattiche. Per una panoramica generale: https://universaldependencies.org/u/overview/tokenization.html.

Esempi di divisione:

- nella → in + la
- alla → a + la
- amarsi → amare + si

- accostarsegli → accostare + se + gli
- de' → di + '
- co' → con + i
- pel → per + i
- colla → con + la
- nel → in + il
- nell' → in + l'
- del → di + il
- dell' → di + l'

Come si può intuire dagli esempi sopra riportati, il senso è ricostruire le forme che compongono parole complesse, come le preposizioni articolate e le parole con clitici:

1. dicendogli: dicendo gli
2. fecersi: fecero si
3. al: a il
4. nol: non lo
5. co': con il

In caso di errori nella tokenizzazione automatica, bisogna correggere a mano dividendo tali token direttamente sul foglio di calcolo: si aggiungono sotto il token in questione tante righe quante le forme che lo compongono (ciascuna forma andrà quindi a occupare una riga) e si sistema la numerazione dei token per tutta la frase. Ricordarsi di aggiungere gli underscore (_) nelle celle che non hanno annotazione. Segnalare i punti su cui si è tokenizzato a mano evidenziandoli con il colore rosso.

## 4. LEMMATIZZAZIONE

Dopo aver controllato la divisione in frasi e la tokenizzazione (vedi Sezioni 1 e 2), si procede ad assegnare la forma citazionale. La lemmatizzazione è stata già effettuata automaticamente ma con un sistema che non è stato addestrato a lemmatizzare testi storici, per cui il lavoro consiste nel controllo e nella eventuale correzione dei lemmi pre-assegnati automaticamente nella colonna C del foglio di calcolo.

In generale, si usa il lemma moderno delle parole, il quale viene indicato nella terza colonna (C). Se le parole si presentano in una forma arcaica, il lemma arcaico, cioè quello che segue la forma antica originale, si aggiunge nella colonna K.

Esempio:

ufizi --> colonna C: ufficio --> colonna K: ufizio

### ARTICOLI

Gli articoli determinativi (*il*, *lo*, *la*, *i*, *gli*, *le* e le corrispondenti forme con apostrofo) si lemmatizzano tutti con il lemma *il*.

Gli articoli indeterminativi (*un*, *uno*, *una*) si lemmatizzano tutti con il lemma *uno*.

## AGGETTIVI

Tutti gli aggettivi si riconducono al maschile singolare e nella loro forma positiva.

Gli aggettivi tronchi (*bel*, *gran* ecc.) sono lemmatizzati come forme del lemma rappresentato dalla forma di citazione non tronca (*bello*, *grande* ecc.).

Gli aggettivi alterati (ad esempio, *belline*) sono stati ricondotti al lemma di grado positivo corrispondente (*bello*). Il lemma corrispondente alla forma alterata va aggiunto alla colonna K (*bellino*).

Vengono portate a lemma corrispondente tutte le forme sintetiche di comparativo e superlativo: per esempio, *migliore* ha come lemma *migliore*, *pessime* ha come lemma *pessimo*.

La differenza tra aggettivi e participi può non creare problemi, ma questa differenza è fondamentale perché le due parti del discorso richiedono lemmi diversi: l'aggettivo si riconduce al maschile singolare, mentre il participio all'infinito del verbo.
Per il presente lavoro di lemmatizzazione, si fa riferimento alla *Grande grammatica italiana di consultazione* (vol. 2), e in particolare al capitolo dedicato al sintagma aggettivale (curato da M.T. Guasti), dove vengono offerti 3 test per la discriminazione tra aggettivi e participi:

1. i participi NON possono essere modificati con il suffisso *-issimo* o con gli avverbi rafforzativi *molto*, *assai* ecc., mentre gli aggettivi sì.
2. solo il participio passato può essere sede di un clitico: per esempio, se abbiamo "la notizia comunicata" e siamo in dubbio sulla natura di "comunicata", possiamo provare ad aggiungergli un clitico, per esempio "la notizia comunicatagli". Essendo accettabile, consideriamo "comunicata" come participio e non come aggettivo. Al contrario, nel caso di "la ragazza simpatica" e "il nome sconosciuto", aggiungengo un clitico avremmo "la ragazza *simpaticagli", "il nome sconosciutagli". Come si vede, "simpatica" e "sconosciuto" non possono essere sede di un clitico, quindi sono aggettivi.
3. i participi sono compatibili con gli ausiliari "essere" e "venire", gli aggettivi solo con "essere".

## AVVERBI

Quando gli avverbi ricorrono con forme alterate, sono stati riportati alla forma positiva: ad esempio, *lucidissimamente* è lemmatizzato come *lucidamente*. Il lemma corrispondente alla forma alterata (*lucidissimamente)* va aggiunto alla colonna K.

"non" deve essere ADV (non PART) con PronType=Neg

## CONGIUNZIONI

Si lemmatizzano uguali a loro stesse. Ad esempio, *e* si lemmatizza "e".

## INTERIEZIONI

Si lemmatizzano uguali a loro stesse.

## PRONOMI

I pronomi personali si lemmatizzano uguali a loro stessi. *Lei* con *lei*, *noi* con *noi*.
Costui → costui
Costei → costei
Costoro → costoro
*colui*, *costui* e forme del paradigma sono dimostrativi, e sempre `PRON`
*proprio* annotato come determinante possessivo e riflessivo di terza persona

## PUNTEGGIATURA

Si lemmatizza uguale a se stessa: "," ha come lemma ",".

## SOSTANTIVI

Si segue la lemmatizzazione standard (maschile singolare) ricordandosi che, come per gli aggettivi, eventuali alterati vanno ricondotti alla forma base. Quindi, *coltellacci* si lemmatizza con "coltello". Il lemma corrispondente alla forma alterata va nella colonna K (*coltellaccio*).

## NOMI PROPRI

Si lemmatizzano uguali a loro stessi e sono gli unici lemmi che hanno l'iniziale maiuscola.

## VERBI

I verbi sono da lemmatizzare sotto l'infinito presente attivo corrispondente.
I verbi con clitici vanno divisi in parole sintattiche (vedi Sezione 3), ognuna con il proprio lemma.

Gli ausiliari non si limitano a *essere* e *avere*, ma comprendono anche altri verbi quando si combinano con infiniti/converbi senza intermediari. In particolare ho annotato come ausiliari:

- *sapere*: sa fare vs. so che fa
- *potere*: può fare vs. può molto
- *volere*: vuol fare vs. voglio che faccia
- *dovere*: deve fare vs. mi deve soldi
- *fare*: fa fare vs. fa male
- *stare*: sta facendo vs. sta attento
- *venire*: viene facendo vs. viene a fare
- *andare*: va facendo vs. va a fare
- *lasciare*: lascia fare vs. lascia l'oggetto
- *bisognare*: bisogna fare

## PAROLE IN LINGUE DIVERSE DALL'ITALIANO

Si lemmatizzano uguali a loro stesse.

inventus → iuventus
est → est

## 5. FEATURE

ATTENZIONE: l'ordine delle feature è quello alfabetico! Es. Degree è sempre prima di Style.

Quando si lemmatizza si controllano le feature già assegnate automaticamente e si aggiungono nella colonna J, se necessario, quelle che seguono (relative alla forma originale del token):

- Style=Arch: forme **arcaiche**. Alcune forme richiedono sia un doppio lemma (contemporaneo nella colonna C e arcaico nella colonna K) che questa feature: ad es. "annunzio". Altri hanno solo la feature: ad esempio, *chieggio* non ha un lemma arcaico legato alla forma (*chieggiere), ma il lemma è solo "chiedere", cui si aggiunge nella colonna J questa feature per indicare che la forma è arcaica. Alcuni lemmi antichi non si trovano nei dizionari moderni, altri hanno un'entrata che rimanda al lemma attuale. Si consiglia di consultare il dizionario del De Mauro in cui i lemmi obsoleti o letterari hanno un rimando alla versione attuale: vedi, ad esempio, https://dizionario.internazionale.it/parola/respondere
- Degree=Dim/Pej/Aug/End: forme **alterate**. Quindi Degree=Dim (diminutive, diminutivo, es. *casina*), =Pej (pejorative, peggiorativo, es. *cagnaccio*), =Aug (augmentative, accrescitivo, es. *bestione*), =End (endearment, vezzeggiativo, *poveretta* - si noti che "etta/etto" è sempre vezzeggiativo e non diminutivo). Ricordarsi che il lemma da mettere nella colonna C è la forma non alterata: per una parola come *pecorelle* è "pecora", per *tavolino* è "tavolo", e così via. Il lemma della forma alterata (*pecorella*, *tavolino*) va, invece, nella colonna K. Sono lemmatizzati come lemmi solo alterati quelli il cui significato non sia ricavabile dalla base della forma suffissata e quelli che hanno un uso più comune del lemma base: si veda, ad esempio, "libretto" https://dizionario.internazionale.it/parola/libretto.
- Variant=Apoc: forme **apocopate**. Le uniche forme apocopate che non si annotano sono quelle degli articoli (ad esempio, *un*). Di solito l'apocope non ha l'apostrofo ma ci sono delle eccezioni come nella parola *po'*. Per ulteriori dettagli, si veda: https://it.wikipedia.org/wiki/Apocope.

Per tutte le altre feature, si rimanda a:
https://github.com/UniversalDependencies/docs/tree/pages-source/_it/feat.

Attenzione che in alcuni casi è necessario modificare le feature assegnate automaticamente. Questo succede, ad esempio, se un participio viene considerato aggettivo o viceversa. Verbi e aggettivi, infatti, hanno feature molto diverse (ad esempio un aggettivo non ha il Tense o il Mood).

### Aspect

Questa feature non è ancora accettata in italiano quindi non deve essere usata.

I casi di imperfetto devono avere: Tense=Imp (nell'annotazione precedente c'erano Aspect=Imp|Tense=Past).

- es. diceva: Mood=Ind|Number=Sing|Person=\1|Tense=Imp|VerbForm=Fin

I participi NON devono avere l'Aspect:

- es. “bastati” = Gender=Masc|Number=Plur|Tense=Past|VerbForm=Part → ATTENZIONE, anche Voice=Pass NON va usato nei participi.

### Polarity versus PronType

*nemmeno*, *neppure*, *no*, *nessuno*, *niente*, *nulla*, *né*, *non* → hanno PronType=Neg e NON Polarity=Neg.
Polarity si usa solo per le INTJ “no” e “sì”: No con Polarity=Neg; Sì con Polarity= Pos.

### PronType=Tot

In molte delle treebank attuali non è usato ma è un errore quindi noi lo usiamo con *tutto*, *ogni*, *ognuno*. → sicuramente da controllare e correggere a mano.

### VerbForm=Ger versus VerbForm=Conv

In molte delle treebank attuali il valore Conv non è usato ma è un errore quindi noi lo usiamo. Va usato con i GERUNDI al posto di Ger. → sicuramente da controllare e correggere a mano.

### `Degree=Cmp`

È aggiunto ai seguenti lemmi: *più*, *meno*, *meglio*, *peggio*, *inferiore*, *maggiore*, *minore*, *peggiore*, *migliore*, *superiore*, *posteriore*, *anteriore*.

### Degree=Sup versus Degree=Abs

Sup non è usato ma è sostituito ovunque con `Abs` per i superlativi. Da aggiungere ai seguenti lemmi: *ottimo*, *pessimo*.

### Person

Da rimuovere da:

- determinanti possessivi: *mio*, *tuo*, *suo*, *nostro*, *vostro*, *loro*

### Gender

Da rimuovere da:

- pronomi personali: *mi*, *me*, *io*, *ci*, *noi*, *vi*, *voi*
- determinante possessivo *loro*
- modificatori (`ADJ`/`DET`) terminanti in *-e*/*-i* (es. *grande*)

### Number

Da rimuovere da:

- determinante possessivo *loro*
- nomi tronchi (es. *città*)

## 6. ERRORI RICORRENTI DA CONTROLLARE

- Prima persona dei verbi
- Verbi al passato remoto
- Congiuntivi: il presente scambiato per l’indicativo e l’imperfetto per il passato remoto indicativo.

- Niente: a meno che non sia sostantivo (per cui ha come feature: Gender=Masc | Number = Sing), togliere Gender e Number.
- vi - ve - v':
    - pronome seconda persona plurale "accasatevi": PRON con feature Clitic=Yes|Number=Plur|Person=2|PronType=Prs
    - se avverbio riferito a un luogo: no feature
- che: può avere varie annotazioni in base al contesto:
    - pronome relativo: si può sostituire con "il quale", "la quale", "i quali".... Si annota come PRON e con feature PronType=Rel.
    - pronome interrogativo: si può sostituire con "cosa…?". Si annota come PRON e feature PronType=Int.
    - aggettivo interrogativo: ha la struttura "che + nome …?". Si annota come DET e feature PronType=Int.
    - pronome esclamativo: si può sostituire con "cosa…!". Si annota come PRON e feature PronType=Exc.
    - aggettivo esclamativo: ha la struttura "che + nome …!". Si annota come DET e feature PronType=Exc.
    - pronome indefinito: esempio "un certo che", "un che di …". Si annota come PRON e feature PronType=Ind.
    - congiunzione che introduce proposizioni subordinate. Si annota come SCONJ e non ha feature.
- lemmi:
    - certe volte sono inventati perché non conosciuti dal sistema automatico
    - certe volte sono sbagliati anche se l'annotazione della parte del discorso è corretta
- feature:
    - aggiungere feature Reflex=Yes a "si", "sé", "proprio" (o se quando in "se stesso")

## 7. CONSIGLI OPERATIVI.

In caso di dubbi, evidenziare la riga in giallo e aggiungere una nota nella colonna L.

## 8. PoS TAGGING

Per una panoramica generale delle etichette di UPOS:
https://universaldependencies.org/u/pos/index.html

Per essere in linea con le scelte dell'UD italiano, si segue un'annotazione contestuale:

- Infiniti sostantivati = NOUN (non hanno feature)
- Participi usati come aggettivi = ADJ

Le parole in lingua diversa rispetto all'italiano si annotano come segue:

- primor primor X SW Foreign=Yes _ _ _ _

Alcuni casi importanti:

- innominato: PROPN (anche se ha l'iniziale minuscola, no feature)
- Dio: PROPN (ma altre forme che si riferiscono a Dio e scritte con la maiuscola come "Colui", "Lui", "Egli" NON sono PROPN)
- signor curato: NOUN NOUN
- Don/don Abbondio: NOUN PROPN
- Giacomo da Lentini: PROPN ADP PROPN

**DET versus PRON**

I DET sono in funzione aggettivale, i PRON in funzione pronominale. I numerali sono determinanti ma hanno un categoria a sé stante: NUM.

Alcuni lemmi possono avere funzione diversa a seconda del contesto.

- Quel bambino → DET
- Ho visto quello → PRON
- Proposte diverse (nel senso di dissimili) → ADJ
- Diverse proposte sono arrivate agli organizzatori (valore indefinito) → DET, PronType=Ind

**CLITICI**

Ci/vi/ne: sono pronomi clitici e non avverbi (ad eccezione di quando significano "in quel luogo"):

- V'era un aria umida → ADV
- C'era un bel tramonto → ADV
- Pietro ne parla → PRON Clitic=Yes|PronType=Prs

**AVVERBI**

Alcuni lemmi hanno funzione avverbiale e vanno annotati come ADV anche se nei dizionari non hanno questa categoria. Ad esempio, possono trovarsi tra ausiliare e participio, non coordinano frasi dello stesso livello sintattico, hanno una posizione molto mobile nella frase.

- anche / soprattutto / almeno / dunque / perciò / anzi / però / tuttavia / pure / nemmeno / pertanto / inoltre

**INTERIEZIONI**

Si riferiscono alla sfera emotiva e costituiscono un atto linguistico completo (sono dette olofrasi) per cui non hanno relazioni sintattiche con altre parole all'interno della frase. Includono le profrasi (sì, no quando risposte a domande più o meno esplicite). Si tratta di olofrasi.

**PREPOSIZIONI**

Si annotano come ADP sia le preposizioni proprie (di, a, da, in, con, su, per, tra, fra) che quelle improprie:

- dentro / dietro / contro / sotto / sopra / accanto

Questi lemmi possono essere annotati come ADV a seconda del contesto:

- Dietro la siepe → ADP (segnalano un rapporto con un sintagma di natura nominale)
- Sono andato dietro → ADV (compaiono in isolamento)

**CASI AMBIGUI**

DOPO, a seconda del contesto è SCONJ, ADP o ADV:

- DOPO + PROPOSIZIONE: Sarai più tranquillo dopo aver sostenuto l'esame —> SCONJ
- DOPO + SINTAGMA NOMINALE: Sarai più tranquillo dopo l'esame → ADP
- DOPO (senza instaurare alcun rapporto sintattico ma modificando la frase): Se sosterrai l'esame dopo sarai più tranquillo → ADV