ISTRUZIONI PROGETTO FASE 3

Data consegna: 22 Giugno

Questo documento contiene le istruzione per l'esecuzione della fase 3 del progetto (analisi sintattica). Per informazioni o domande scrivere a pennacchiotti@info.uniroma2.it.
Le terza fase deve essere eseguite da tutti i membri del gruppo, ciascuno dei quali dovrà analizzare la sezione del testo precedentemente annotata nella fase 1 e 2.
L'**analisi sintattica** deve essere eseguita in due passi distinti: *analisi dei costituenti* e *analisi delle dipendenze*.
ATTEZIONE: prima di iniziare la fase 3 del progetto, è necessario apportare le modifiche a Chaos, come indicato nelle slides della lezione *NLP_7*. Inoltre tali slides contengono alcuni esempi di costrutti sintattici complessi nel formalismo di Chaos che devono essere seguiti durante il processo di annotazione.

1. Analisi dei costituenti

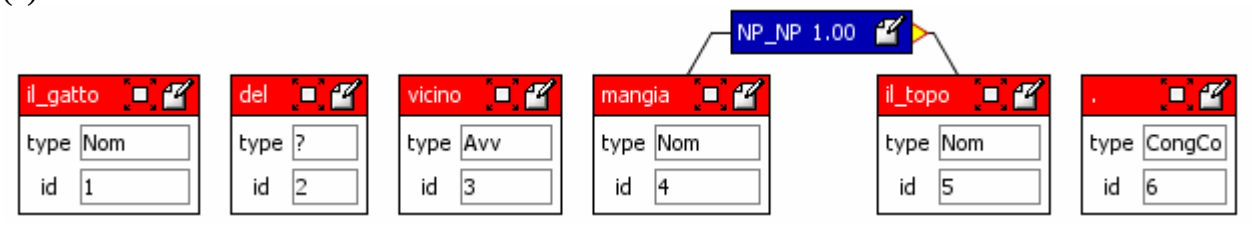
Scopo dell'analisi dei costituenti è quella di correggere gli errori prodotti da Chaos, e valutare la *precision*, la *recall*, la *labeled precision* e la *labeled recall* (per maggiori informazioni su queste misure fare riferimento alle slide della lezione *NLP_7* presenti sul sito del corso).
Ogni annotatore dovrà effettuare la correzione su:

1. ***chunks*:** costituenti indicati *in rosso*. Due tipi di correzione devono essere effettuate:
   a. ***identificazione dei chunk***: l'annotatore dovrà correggere gli errori di Chaos (*Fig.1a*) nell'identificare i chunk (*Fig.1b*)
   b. ***tipo del chunk*:** l'annotatore dovrà assegnare ad ogni chunk il tipo corretto (*Fig.1c*).

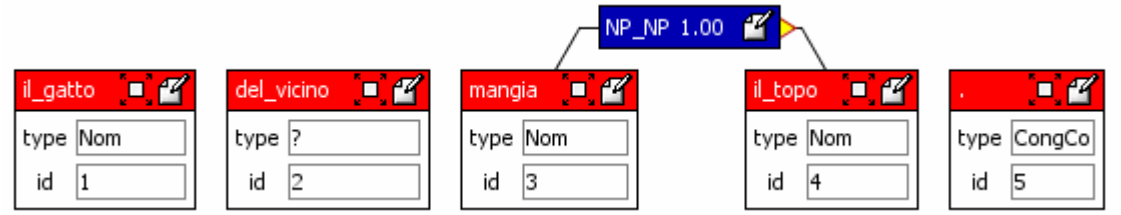
2. ***sotto-costituenti***: per ogni chunk l'annotatore dovrà (*Fig.2*) ***identificare la head e il governor:*** selezionando la corretta testa sintattica (head) ed il governatore (governor) per il chunk in esame.

Per dettagli sulle definizioni di chunk, costituente, governor, head, fare riferimento alle slides del corso.
La fase di valutazione delle prestazioni consiste nel calcolare *precision*, la *recall*, la *labeled precision* e la *labeled recall*, solo per i chunk (quindi non è necessario calcolare tali misure per i sotto-costituenti). Per il calcolo delle misure, seguire la procedura indicata nelle slides del corso.

(a)



(b)



(c)

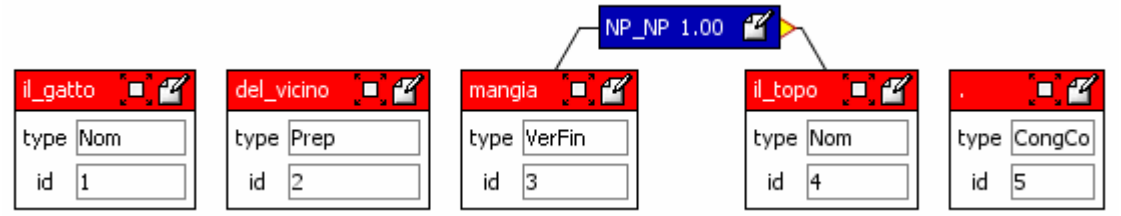
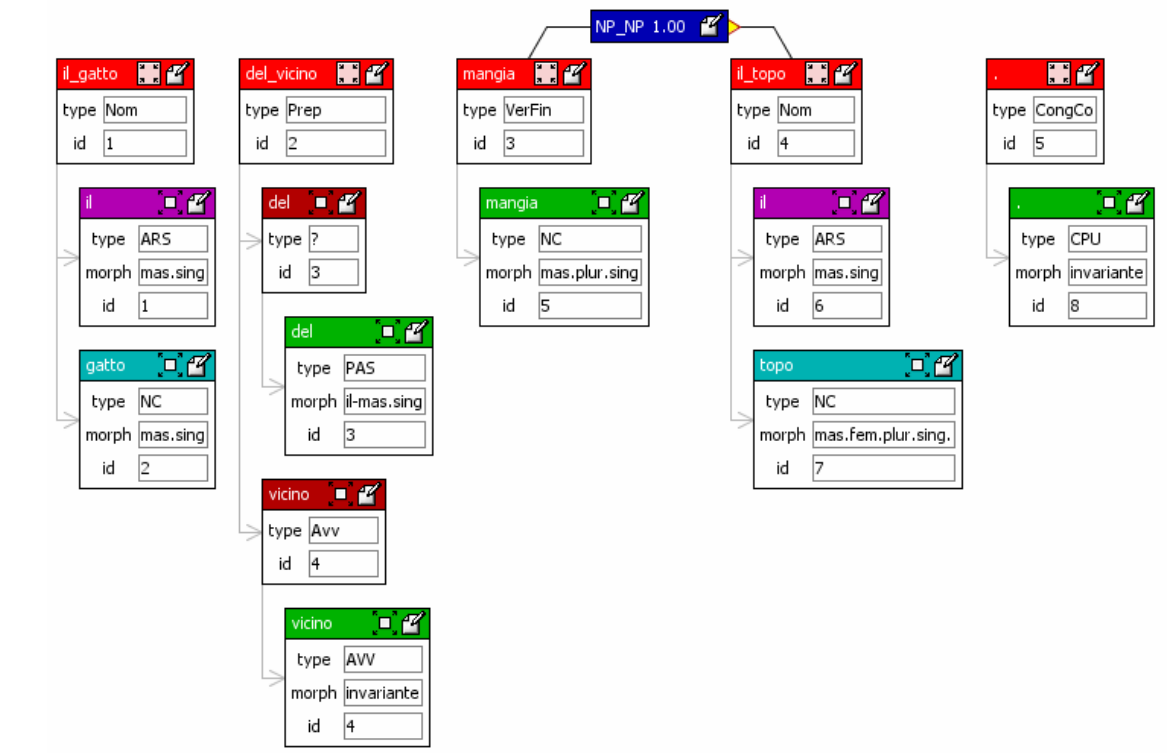
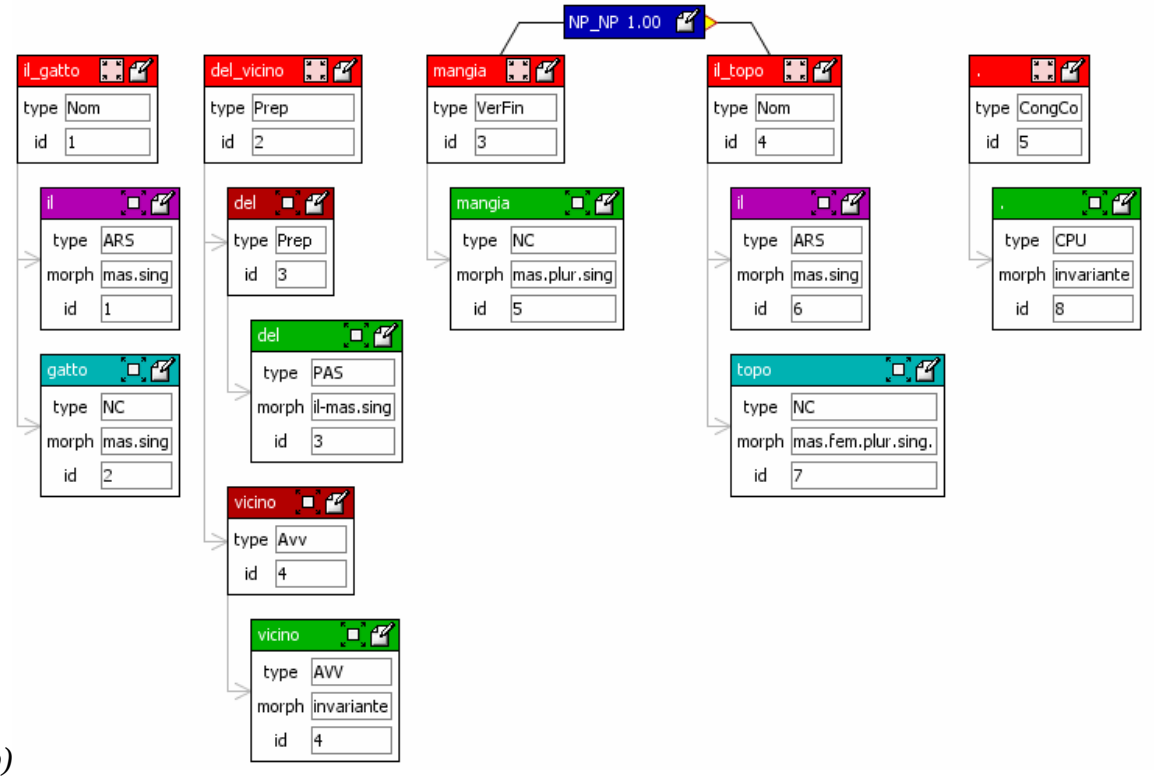


Fig. 1 . Esempio di analisi dei costituenti, punto 1 (*chunks*)

(a)





(b)

Fig. 2 . Esempio di analisi dei costituenti, punto 2 (*sotto-costituenti*). *(a)*: prima del punto 2. *(b)*: dopo punto 2.

La procedura per l'analisi dei costituenti è la seguente:

1. Per ogni file prodotto al termine della fase 2 (`[nome_frammento]_[num_paragrafo]_pos.cha` :
    a. Caricare il file dall'interfaccia grafica di Chaos (*File→Open Chaos File*)
    b. Verranno visualizzati gli XDG del paragrafo
    c. Effettuare l'analisi dei chunk per ogni XDG (differenti finestre *Fig.3-1*):
        - ***1a (identificazione dei chunk)***, utilizzando gli operatori di *unione* (*Fig.3-4*) e di *separazione* (*Fig.3-5*) dei chunk:
            - *unione*: per unire due chunk, selezionarli con il bottone sinistro del mouse; i chunk diventeranno di colore verde; quindi premere il bottone (*Fig.3-4*)
            - *separazione*: per separare un chunk in due chunk differenti, selezionarlo con il bottone sinistro del mouse; il chunk diventerà di colore verde; quindi premere il bottone (*Fig.3-5*)

        - ***1b (tipo del chunk)***: selezionare il corretto tipo del chunk, aprendo la *Edit Window* con il bottone (*Fig.3-8*), e servendosi del menù a tendina (*Fig.3-9*). Utilizzare la tabella dei tipi di chunk in appendice come riferimento.
        - ***2 (identificazione della head e del governor)***: selezionare il sotto-costituente con il bottone sinistro del mouse; utilizzare il bottone (*Fig.3-6*) per identificarlo come head, oppure il bottone (*Fig.3-7*) per identificarlo come governor. I colori dei costituenti sono:
            - viola : head
            - celeste: governor
            - verde : head + governor
            - grigio: altri costituenti

    d. Terminata l'analisi di tutti gli XDG, salvare la struttura create: *File→SaveAs* dando al nuovo file il nome: `[nome_frammento]_[num_paragrafo]_`**`cost`**`.cha`
    e. Valutare le prestazioni (la *precision*, la *recall*, la *labeled precision* e la *labeled recall*) sui tutti i chunk del file, comparando gli XDG appena annotati con quelli prodotti da Chaos.
2. Calcolare le prestazioni (la *precision*, la *recall*, la *labeled precision* e la *labeled recall*) di Chaos su tutti i file della pagina web, come media dei valori ottenuti sui singoli file.

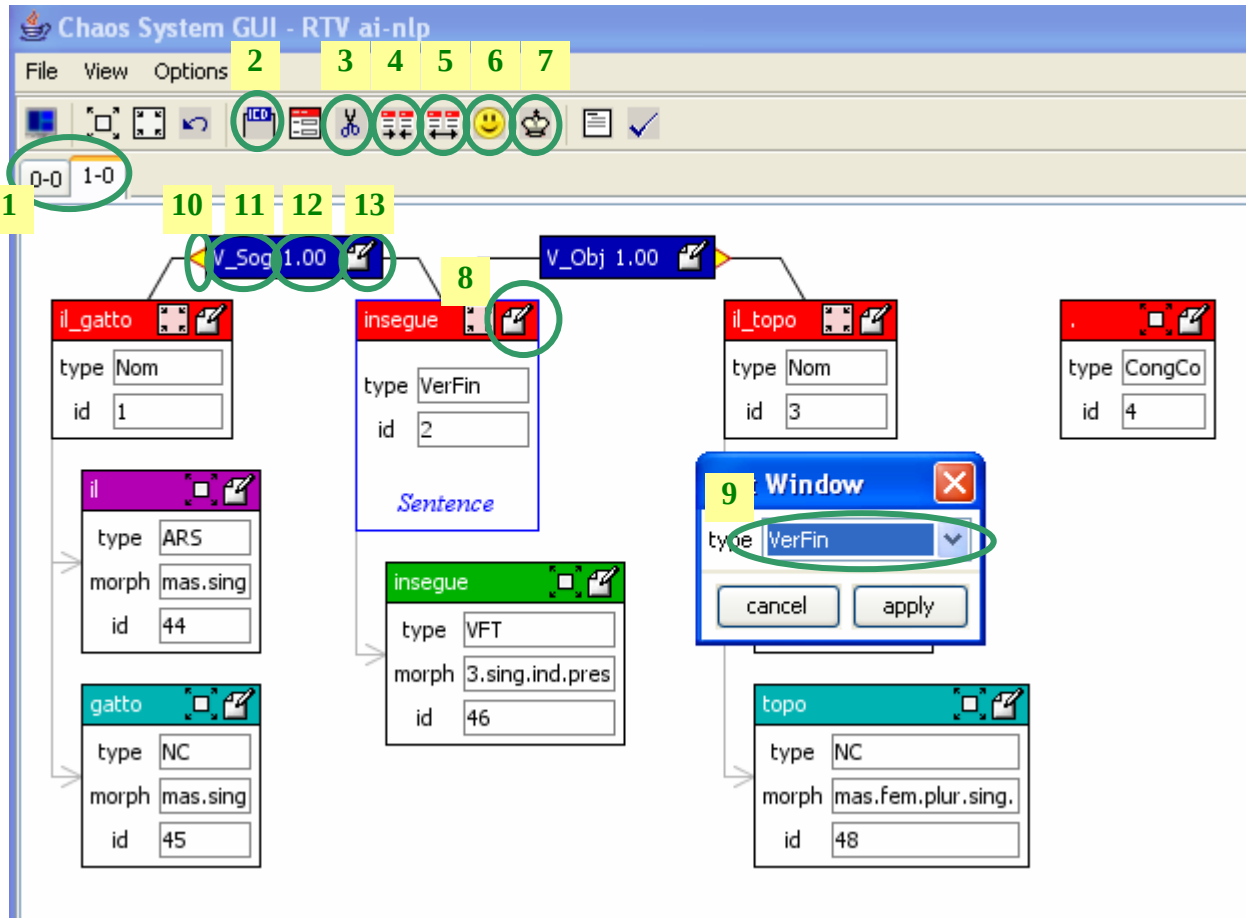


Fig.3 Strumenti per l'analisi dei costituenti e delle dipendenze

2. Analisi delle dipendenze (ICD)

Scopo dell'analisi delle dipendenze (ICD) è quello di correggere gli errori prodotti da Chaos, e valutare la *precision* e la *recall* (per maggiori informazioni su queste misure fare riferimento alle slide della lezione *NLP_7* presenti sul sito del corso).
Un *ICD* indica una relazione sintagmatica *orientata* tra due costituenti. L'etichetta dell'ICD indica il tipo (*Fig.3-11*) ed il verso (*Fig.3-10*) della relazione. Ad esempio un ICD *V_sog* indica una relazione *verbo-soggetto*, il cui verso va dal costituente verbale a quello nominale (*Fig.3-10*). Ogni ICD ha assegnato un grado di *plausibilità* (*Fig.3-12*) indicante il grado di certezza della relazione (valori da 0 a 1).
Ogni annotatore dovrà:

1. **cancellare gli ICD *errati***, ovvero gli ICD che connettono due costituenti che non sono in relazione, oppure che connettono due costituenti in relazione, ma con il verso sbagliato (per esempio un ICD di tipo *V_sog* che inizia sul soggetto e termina sul verbo, invece del contrario);
2. **inserire nuovi ICD**, quando sussiste una relazione tra due costituenti non identificata da Chaos;
3. **cambiare l'etichetta degli ICD**, se Chaos ha identificato una relazione tra due costituenti correttamente, ma il tipo di ICD è sbagliato.
4. **Settare la plausibilità degli ICD**: per ogni ICD, l'annotatore dovrà porre il valore di plausibilità a 1 (certezza). Solo in caso di ambiguità, l'annotatore potrà produrre più ICD e porre il valore di plausibilità a differenti valori, la cui somma dovrà comunque essere pari a 1.

La procedura per l'analisi delle dipendenze è la seguente:

1. Per ogni file prodotto al termine della fase precedente (`[nome_frammento]_[num_paragrafo]_cost.cha`:
   a. Caricare il file dall'interfaccia grafica di Chaos (*File→Open Chaos File*)
   b. Verranno visualizzati gli XDG del paragrafo
   c. Effettuare l'analisi delle dipendenze per ogni XDG (differenti finestre *Fig.3-1*):
      - ***cancellazione ICD:*** selezionare l'ICD da cancellare con il bottone sinistro del mouse; l'ICD diventerà di colore verde; quindi premere il bottone (*Fig.3-3*)
      - ***inserimento nuovi ICD:*** selezionare i due chunk da connettere; i chunk diventeranno di colore verde; utilizzare il bottone (*Fig.3-2*) per creare l'ICD. Selezionare dal pop-up la direzione dell'ICD; selezionare il corretto tipo del ICD, aprendo la *Edit Window* con il bottone (*Fig.3-13*), e servendosi del menù a tendina. Utilizzare la tabella dei tipi di ICD in appendice come riferimento
      - ***cambiare etichetta agli ICD a selezione della plausibilità:*** selezionare il corretto tipo di ICD e il valore di plausibilità, aprendo la *Edit Window* con il bottone (*Fig.3-13*), e servendosi del menù a tendina. Utilizzare la tabella dei tipi di ICD in appendice come riferimento.
   d. Terminata l'analisi di tutti gli XDG, salvare la struttura create: *File→SaveAs* dando al nuovo file il nome: `[nome_frammento]_[num_paragrafo]_synt.cha`
   e. Valutare le prestazioni (*precision*, *recall*) sui tutti gli ICD del file, comparando gli XDG appena annotati con quelli prodotti da Chaos (e sorpavvisuti alle modifiche della fase precedente) nel file (`[nome_frammento]_[num_paragrafo]_const.cha`)

2. Calcolare le prestazioni (*precision recall*) di Chaos su tutti i file della pagina web, come media dei valori ottenuti sui singoli file.

3. *Prima di iniziare*: agreement

Prima di iniziare l'annotazione complessiva, tutti i membri del gruppo dovranno annotare autonomamente (analisi dei costituenti e delle dipendenze) il frammento di agreement `agreement_fragment_[idGruppo].txt` prodotto a valle della fase 1 e 2, seguendo la stessa procedura utilizzata per l'annotazione complessiva presentata nelle precedenti sezioni.
L'annotazione finale del gruppo del frammento di agreement dovrà essere infine ottenuta confrontando le annotazioni prodotte dai singoli membri, che si accorderanno sulla interpretazioni corretta finale. Questa deve essere salvata nel file:

`agreement_fragment_synt_[idGruppo].cha`

4. Invio risultati

Inviare tutti i file prodotti utilizzando la e-mail del gruppo a *pennacchiotti@info.uniroma2.it*.
Indicare come soggetto della e-mail: "*Progetto parte 3: gruppo [Id gruppo]*".
Il messaggio deve contenere i valori di precision e recall in questo formato:

precisionChunk = ??
recallChunk = ??
precisionLabChunk = ??
RecallLabChunk = ??
precisionIcd = ??
RecallIcd = ??

Allegare un file compresso (*zip* o *rar*), contenenti i file prodotti:

`agreement_fragment_synt_[idGruppo].cha`
`[nome_frammento].html`

per ogni paragrafo:

`[nome_frammento]_[num_paragrafo].txt`
`[nome_frammento]_[num_paragrafo].cha`
`[nome_frammento]_[num_paragrafo]_`**`morp`**`.cha`
`[nome_frammento]_[num_paragrafo]_`**`pos`**`.cha`
`[nome_frammento]_[num_paragrafo]_`**`cost`**`.cha`
`[nome_frammento]_[num_paragrafo]_`**`synt`**`.cha`

TIPI DI *CHUNK* PER L'ITALIANO

TIPO	SIGNIFICATO
Agg	Chunk aggettivali
Avv	Chunk avverbiale
CongCo	Chunk coordinativo
CongSub	Chunk subordinativo
Nom	Chunk nominale
Prep	Chunk preposizionale
VerFin	Chunk verbale finito
VerGer	Chunk verbale gerundivo
VerInf	Chunk verbale infinito
VerPart	Chunk verbale participio
VerPred	Chunk verbale predicativo aggettivale
VerNom	Chunk verbale nominale
VerPrep	Chunk verbale preposizionale
?	Chunk sconosciuti

TIPI DI *ICD* PER L'ITALIANO

TIPO	SIGNIFICATO
V_Sog	Grammatical Subject
V_Obj	Grammatical Object
V_NP	Indirect Object
V_PP	Verb Preposition Modifier
V_Adv	Verb Adverb Modifier
NP_NP	gruppo Nominale Nominale
NP_PP	gruppo Nominale Preposizionale
PP_PP	gruppo Preposizionale Preposizionale
NP_Adj	gruppo Nominale Aggettivo
NP_VPart	gruppo Nominale Participio
PP_Adj	gruppo Preposizionale Aggettivo
PP_VPart	gruppo Preposizionale Participio
Adj_PP	gruppo Aggettivo Preposizionale
Adv_PP	gruppo Avverbio Preposizionale
x_Cong_x	Congiunzione coordinativa tra costituenti
x_Cong	gruppo Costituente Congiunzione
x_Sub	gruppo Costituente Subordinata
V_CSub	gruppo Verbo Congiunzione Subordinativa