La dtd
di
AN.ANA.S. 4
(Annotazione e analisi sintattica)

Miriam Voghera
Annamaria Landolfi
Carmela Sammarco

AN.ANA.S. 4

1. La dtd

L'applicazione AN.ANA.S. è stata creata per l' annotazione e l' analisi sintattica di tutti i tipi di testo facendo uso del linguaggio XML. AN.ANA.S. si avvale del software manuale XGate che prevede varie funzionalità tra cui quella di Editor che serve ad etichettare, modificare e infine creare un database di testi in formato XML avvalendosi di una struttura chiamata DTD (*Document Type Definition*).
La DTD è uno schema di regole grammaticali che definisce la struttura ad albero del testo. La rappresentazione per alberi viene generata attraverso l'annotazione in formato XML del testo trascritto. La DTD contiene tutte le informazioni su tutti gli elementi che costituiscono un nodo dell'albero. Le informazioni riguardano:

- il nome dell'elemento;
- i nomi degli elementi che possono dipendere da esso o, se si vuole, da cui può essere costituito;
- la lista di attributi associati a quel determinato elemento e i loro possibili valori.

La definizione di un'entità può avvenire in base ad una lista ben definita di attributi ('ATTLIST') oppure semplicemente in base ad un modello di contenuti (CDATA). Gli attributi, a loro volta, possono avere un valore scelto all'interno di un determinato set definito o essere costituiti da una stringa di testo.

1)
```
<!ELEMENT turn (sentence+ | subturn+)>
<!ATTLIST turn
turn_id CDATA #REQUIRED
compl (t | f) #REQUIRED
```

Nell'esempio 1) vediamo riportate le informazioni relative all'elemento TURN. Da questo elemento possono dipendere i due elementi SUBTURN e SENTENCE in numero infinito (+); gli attributi dell'elemento *turn* sono:

- TURN ID che corrisponde al numero progressivo di turni presenti in un testo;
- COMPL che indica se si tratta di un turno di completamento, i cui valori possono essere solamente *true* (t) o *false* (f).

Entrambi gli attributi qui esemplificati sono obbligatori (REQUIRED). Come vedremo, esistono anche attributi non obbligatori (IMPLIED).

1. **I livelli di analisi**

Poiché l'idea da cui nasce AN.ANA.S. è quella di costruire un programma di analisi sintattica adeguato a tutti i tipi di testo, si è cercato di mettere a punto una dtd che fosse adattabile sia al parlato che allo scritto.
In questo paragrafo daremo una descrizione dettagliata della dtd, indicando i nomi delle entità, gli attributi e i loro valori. Per comodità la descrizione verrà suddivisa in livelli. Bisogna però avvertire che la dtd prevede un alto livello di ricorsione e che i vari livelli non devono essere intesi come rigidamente fissati, ma possono contenere anche entità di livello superiore. La ricorsione è prevista per ovvi motivi sintattici, ma anche perché AN.ANA.S. è parte di un progetto integrato di annotazione multilivello che rende necessario il mantenimento dell'allineamento con la linearità del testo. Ciò fa sì che l'etichettatura ottenuta possa essere rappresentata come una sorta di parentesizzazione sintattica.
Per un uso oramai diffuso in tutti i programmi di annotazione e di *parsing*, la terminologia usata è in inglese.

Per ogni livello indichiamo il nome degli elementi previsti e i loro attributi. I valori di ogni attributo sono messi tra parentesi. Nel caso di attributi booleani i valori previsti sono T=true e F= false.

2. 1. Primo elemento: TEXT

Il primo livello è costituito dall' ELEMENT TEXT e ha lo scopo di fornire informazioni generali sulla macrostruttura testuale del testo parlato o scritto. Questo livello prevede i seguenti attributi:
TEXT_ID: numero o sigla identificativi del testo.
TYPE (N | D | E | A | O) : indica il tipo testuale e ha i seguenti valori: n=narrative, d=descriptive, e= explicative, a= argumentative, o=other.
PRODUCTION (MG | DG): indica il contesto enunciativo del testo e ha i seguenti valori: m=monologue, d=dialogue.
FREE_INT (T | F) : indica se si tratta di un testo con presa di parola è libera (free interaction) o no.
TASK (T | F) : indica se si tratta di un testo elicitato in un contesto prestabilito (ad es. test delle differenze, *maptask*).

2. 2. Secondo livello: element PARAGRAPH e TURN

Il secondo livello prevede due possibili entità: PARAGRAPH e TURN. Si userà il primo nel caso di testi monologici, parlati o scritti, il secondo nel caso di testi dialogici parlati o scritti..
Se scegliamo paragraph, bisognerà compilare solo un attributo:
PARAGRAPH_ID : numero identificativo del paragrafo; ogni paragrafo sarà numerato in sequenza.

Se scegliamo turn, avremo due attributi:
TURN_ID: identificativo del turno; ogni turno sarà numerato in sequenza.
COMPL (T | F) : indica se si tratta di un completamento sintattico di un turno precedente.

2. 3. Terzo livello: element SUBPARAGRAPH e SUBTURN

Come abbiamo già anticipato, l'etichettatura sintattica mantiene l'allineamento con il testo e quindi, nel caso di testi parlati, con il segnale audio. Per evitare di avere file audio troppo pesanti, si rende talvolta utile spezzare i paragrafi o i turni troppo lunghi. In tal caso abbiamo previsto SUBPAR e SUBTURN. Entrambi questi elementi prevedono un'identificazione:
SUBPAR_ID : numero identificativo del sottoparagrafo; ogni sottoparagrafo sarà numerato in sequenza;
SUBTURN_ID : numero identificativo del sottoturno; ogni sottoturno sarà numerato in sequenza.

2. 4. Quarto livello: element SENTENCE

Il quarto livello è costituito dall'ELEMENT SENTENCE, che prevede i seguenti attributi:
SPLIT (START | MID | END): segnala una frase 'spezzata' (splitted) che continua in un altro turno; i tre valori indicano se si tratta della parte iniziale, centrale o finale della frase.
N_OF_CLAUSES : indica il numero delle clausole (number of clauses) che costituiscono la frase.

2. 5. Quinto livello: element CLAUSE

Il quinto livello è costituito dall'ELEMENT CLAUSE, che prevede i seguenti attributi: .
TYPE (M | DEP | VL) : indica se si tratta di una clausola principale (m=main), dipendente (dep=dependent), o a nodo centrale non verbale (vl=verbless).
N_OF_PHRASES : indica il numero di sintagmi (number of phrases) che costituiscono la clausola.
LINK (S_CONJ | S_PREP | NULL | REL|NULL_REL): questo attributo va riempito solo in caso di clausola DEP (è, dunque, IMPLIED) ed indica se la relazione di dipendenza tra detta clausola e la clausola principale reggente è realizzata tramite una congiunzione subordinante, (subordinating conjunction), una preposizione subordinante (subordinating preposition); senza subordinatore (null); con subordinatore che introduce una relativa (relativizer) o è una relativa senza subordinatore.
ARG (T | F) : indica se la clausola subordinata è di tipo argomentale.

2. 6. Sesto livello

Il sesto livello è quello in cui abbiamo inserito il testo. Per questo motivo si è deciso di marcare non solo i sintagmi, ma anche parti di testo che non costituiscono sintagmi:. A questo livello ci deve essere dunque necessariamente almeno una delle entità-foglia dell'albero a cui si associa il testo vero e proprio. Le entità previste sono:

- sintagma nominale <NP>
- sintagma verbale <VP>
- sintagma preposizionale <PP>
- sintagma predicativo nominale <PredP>
- congiunzioni <COORD |SUB>
- stringhe 'isolate' <ISO>
- esitazioni <HES>
- stringhe ripetute <REP>
- segnali discorsivi <DM>
- retrace-and-repair sequences: <RR>

NP (Noun Phrase): attributi e loro valori

Ciascun sintagma nominale presenterà una stringa con i seguenti attributi:

LEXEME: indica il lessema testa del sintagma nominale nella sua forma di citazione.
MW (T | F): indica se la testa NP è una polirematica (multiword).
N (T | F): indica se la testa di NP è un nome (noun).
CL (T): se la testa di NP è un pronome clitico (clitic).
ARG (T | F):: indica se la testa di NP è un nome argomentale (argumental).
SUB (T | F): indica se NP è soggetto (subject).
OBJ (T | F): indica se NP è oggetto (object).
DICONTINOUS (T | F) : viene marcato nel caso in cui il sintagma è discontinuo (viene interrotto da materiale linguistico che non appartiene al sintagma stesso, ad es. in *un momento di aggregazione molto importante*)
DIS. ID.: serve alla tracciabilità del sintagma discontinuo, e va riempito con un indice, che poi verrà ripetuto in CONTIN (v. sotto)
POSITION (PRE | POST | NULL | INFRA): indica la posizione NP rispetto al verbo.
DET (T | F): indica se NP è preceduto da un determinante[1] (determinant).
MOD (T | F): indica se la testa dell'NP ha un modificatore[2] (modifier)
MULTIPLE (T | F): segnala costituenti multipli (che hanno cioè livelli di ricorsione al loro interno).
MULT_N: indica il numero dei livelli di ricorsione in costituenti multipli. Il numero include anche il primo livello ed è decrescente: ad esempio, nella stringa *La casa del ragazzo di mia sorella*, *la casa* avrà Multiple= 3 e poi via via a scalare, *del ragazzo* 2 e *di mia sorella* 1.

[1] Sono considerati determinanti: gli articoli, gli aggettivi dimostrativi, gli aggettivi indefiniti, i partitivi
[2] Sono considerati modificatori: gli aggettivi qualificativi, gli aggettivi possessivi, le apposizioni, i numerali, i quantificatori polirematici (come *un po' di*).

WEIGHT: va riempito con un numero e indica il peso del sintagma, secondo una scala che tiene conto della presenza nel sintagma di determinanti, modificatori e livelli di ricorsione costituiti da sintagmi o intere clausole relative.

VP (verb phrase): attributi e loro valori

Ciascun sintagma verbale presenterà una stringa con i seguenti attributi:

LEXEME: indica il lessema testa del sintagma verbale nella sua forma di citazione
MW (T | F): indica se la testa del VP è una poliremática (multi word).
COP_VB (T | F) : indica se si tratta di una copula o di un verbo copulativo (copular verb).
N_OF_ARG (0 | 1 | 2 | 3): indica il numero degli argomenti del VP (number of arguments).
SAT (T | F): indica se le valenze di VP sono saturate (saturated); nella valutazione non si considera l'argomento soggetto, la cui presenza o assenza viene ricavata dai valori dell'attributo SUB. L'unica eccezione riguarda i verbi monovalenti, in cui, se il soggetto non è espresso, l'attributo SAT avrà valore F.
MOD (T | F): indica se la testa del VP è modificata da avverbi (modifier).
SUB (T | F | NULL): indica se il soggetto di VP è espresso o no (subject).
SUB_TYPE (N | PRON | O): indica se il soggetto di VP è un nome, pronome o altro (subject type).
POSITION (PRE | POST|DIS|INFRA): indica la posizione del verbo rispetto al soggetto.
DICONTINOUS (T | F) : viene marcato nel caso in cui il sintagma è discontinuo (viene interrotto da materiale linguistico che non appartiene al sintagma stesso, ad es. in *un momento di aggregazione molto importante*)
DIS. ID.: serve alla tracciabilità del sintagma discontinuo, e va riempito con un indice, che poi verrà ripetuto in CONTIN (v. sotto)
MULTIPLE (T | F): segnala costituenti multipli (che hanno cioè livelli di ricorsione al loro interno)
MULT_N: indica il numero dei livelli di ricorsione in costituenti multipli. Il numero include anche il primo livello ed è decrescente: ad esempio, nella stringa *La casa del ragazzo di mia sorella*, *la casa* avrà Multiple= 3 e poi via via a scalare, *del ragazzo* 2 e *di mia sorella* 1.
WEIGHT: va riempito con un numero e indica il peso del sintagma, secondo una scala che tiene conto della saturazione del verbo e/o della presenza di modificatori avverbiali e/o di verbi servili, modali o causativi.

PP (prepositional phrase): attributi e loro valori

Ciascun sintagma preposizionale presenterà una stringa con i seguenti attributi:

PREP: indica la preposizione che introduce il PP (preposition).
LEXEME: indica il lessema introdotto dalla preposizione nella sua forma di citazione.

MW (T | F): indica se il lessema introdotto dalla preposizione è una polirematica (multiword).
N (T | F): indica se il lessema introdotto dalla preposizione è un nome (noun).
ARG (T): indica se il lessema introdotto dalla preposizione è un nome argomentale (argumental).
CL (T): indica se il PP è costituito da un clitico (clitic).
MP (NP | VP | PP | PREDP | NULL): indica il tipo di sintagma modificato dal PP (modified phrase).
POSITION (PRE | POST | INFRA): indica la posizione (position) di PP rispetto al sintagma che modifica
CIRC: quest'attributo è del tipo IMPLIED, segnala un PP circostanziale e va riempito solo se MP=NULL
DICONTINOUS (T | F) : viene marcato nel caso in cui il sintagma è discontinuo (viene interrotto da materiale linguistico che non appartiene al sintagma stesso, ad es. in *un momento di aggregazione molto importante*)
DIS. ID.: serve alla tracciabilità del sintagma discontinuo, e va riempito con un indice, che poi verrà ripetuto in CONTIN (v. sotto)
DET (T | F): indica se nel PP c'è un determinante (determinant).
MOD (T | F): indica se nel PP c'è un modificatore (modifier).
MULTIPLE (T | F): segnala costituenti multipli (che hanno cioè livelli di ricorsione al loro interno)
MULT_N: indica il numero dei livelli di ricorsione in costituenti multipli. Il numero include anche il primo livello ed è decrescente: ad esempio, nella stringa *La casa del ragazzo di mia sorella*, *la casa* avrà Multiple= 3 e poi via via a scalare, *del ragazzo* 2 e *di mia sorella* 1.
WEIGHT: va riempito con un numero e indica il peso del sintagma, secondo una scala che tiene conto della presenza nel sintagma di determinanti, modificatori e livelli di ricorsione costituiti da sintagmi o intere clausole relative.

PredP (Prediicative phrase) : attributi e loro valori

Consideriamo PredP tutti i sintagmi predicativi nominali sia quando seguono la copula sia quando svolgono la loro funzione predicativa all'interno di una clausola senza verbo. Etichettiamo come PredP , quindi, le parti in corsivo nei due esempi seguenti:

2) La cucina è *grande*
3) *Grande* la cucina!

Ciascun sintagma predicativo presenterà una stringa con i seguenti attributi:

LEXEME: indica il lessema del predicato nominale nella sua forma di citazione
MW (T | F): indica se il lessema è una polirematica (multiword).
P_OF_SPEECH (N | ADJ | PRON | O): indica a quale parte del discorso (part of speech) appartiene il predicato nominale, nome (noun), aggettivo (adjective), pronome (pronoun), altro (other).

ARG (T): indica se il lessema è un nome argomentale (argumental).
CL (T): indica se il predicato nominale è un pronome clitico (clitic).
POSITION (PRE | POST | NULL|INFRA): indica la posizione del PredP rispetto al verbo copulativo, quando questo è presente; nel caso il sintagma predicativo non sia accompagnato da nessun verbo indicheremo sempre la posizione null.
DICONTINOUS (T | F) : viene marcato nel caso in cui il sintagma è discontinuo (viene interrotto da materiale linguistico che non appartiene al sintagma stesso, ad es. in *un momento di aggregazione molto importante*)
DIS. ID.: serve alla tracciabilità del sintagma discontinuo, e va riempito con un indice, che poi verrà ripetuto in CONTIN (v. sotto)
DET (T | F): indica se il PredP è preceduto da un determinante (determinant).
MOD (T | F): indica se il PredP è accompagnato da un modificatore (modifier).
MULTIPLE (T | F): segnala costituenti multipli (che hanno cioè livelli di ricorsione al loro interno)
MULT_N: indica il numero dei livelli di ricorsione in costituenti multipli. Il numero include anche il primo livello ed è decrescente: ad esempio, nella stringa *La casa del ragazzo di mia sorella*, *la casa* avrà Multiple= 3 e poi via via a scalare, *del ragazzo* 2 e *di mia sorella* 1.
WEIGHT: va riempito con un numero e indica il peso del sintagma, secondo una scala che tiene conto della presenza nel sintagma di determinanti, modificatori e livelli di ricorsione costituiti da sintagmi o intere clausole relative.

DM (Discourse marker): attributi e loro valori

L'etichetta DM serve a indicare tutti i tipi di segnali discorsivi.

INFRA (T): indica se DM è incluso in un altro sintagma ma solo se esso è un elemento linearmente interposto e non parte del costituente (infraposed).

CONJ (Conjunction)

L'elemento ***CONJ*** indica la presenza di congiunzioni coordinanti o subordinanti tra clausole e sintagmi.

TYPE (COORD| SUB): indica il tipo di congiunzione (coordinante o subordinante).

ISO (Isolated): attributi e loro valori

Indica tutte le stringhe che non costituiscono sintagma perché sequenze interrotte, isolate (per esempio le interiezioni o i fonosimboli).
L'elemento ISO prevede l' attributo TYPE che indica la categoria di appartenenza di questa stringa non processata; questo attributo può avere i seguenti valori:

ADV: avverbio (adverb)
ADJ: aggettivo (adjective)
PREP: preposizione (preposition)
CONJ: congiunzione (conjunction)
N: nome (noun)
V: verbo (verb)
PRON: pronome (pronoun)
ART: articolo (article)
INT: interiezione (interjection)
PH: fonosimbolo (phonosymbol)

INFRA (T): indica se ISO è incluso in un altro sintagma ma solo se esso è un elemento linearmente interposto e non parte del costituente (infraposed).

HES (Hesitation)
L'elemento HES indica i vari tipi di esitazione nel parlato e non prevede una lista di attributi, ma solo la trascrizione del testo.

REP (Repetition)
L'elemento REP indica parti del testo ripetute e non prevede nessun attributo, ma solo l'indicazione della stringa ripetuta.

RR (Retrait-and -repear sequences)
L'elemento RR indica correzioni e false partenze con cambiamento o meno del progetto indipendentemente da come sono costituite.

INFRA (T): indica se RR è incluso in un altro sintagma ma solo se esso è un elemento linearmente interposto e non parte del costituente (infraposed).

CONTIN
Quest'elemento riporta, nel nodo foglia, il testo della seconda parte di un costituente discontinuo. Possiede l'attributo DIS_ID che va riempito con lo stesso indice riportato nella prima parte del costituente.