Progetto CLIPS
Corpora e Lessici di Italiano Parlato e Scritto

W1- a5
(**S**trumenti **H**ardware e **S**oftware)

Title: ***Un Database per CLIPS***

Document No: CLIPS/W1-a5/SHS-DB/004

Document Name: NADL011

Status: **pubblico**

Date: 10/3/2006

Authors: *Pasquale Basile, Francesco Cutugno*

1 SCOPI E FINALITÀ 3

1.1 Archiviazione 3
1.2 Consultazione 3
1.3 Le entità fondamentali 3

2 IL MATERIALE 4

2.1 Corpora 4
2.1.1 Frequenza di campionamento 4
2.2 Struttura del materiale 4
2.3 Tipi di file 5
2.4 Dislocazione fisica del materiale 7
2.4.1 Classificazione del materiale e denominazione directory 9
2.4.2 Esempi 11

3 VINCOLI DI ETICHETTATURA 11

3.1.1 WRD <-> STD 11
3.1.2 WRD|STD <-> PHN 11
3.1.3 PHN <-> ACS 13
3.1.4 ADD 13
3.2 Tipologia del materiale 14

4 ARCHITETTURA DEL SISTEMA 14

4.1 Configurazione adottata 16

5 STRUTTURA DELLA BASE DATI 17

5.1 Commenti alla struttura 19
5.1.1 Vincoli 19
5.1.2 MATERIALE 19
5.1.3 Speaker 20
5.1.4 UDT (Unità di Trascrizione) o Turni 20
5.1.5 Etichettature 21
5.1.6 [ADD] 21
5.1.7 [WRD] 22
5.1.8 [PHN] 22
5.1.9 [ACS] 23
5.1.10 Tabelle Supplementari 23
5.2 Popolazione della base dati 24
5.3 Validazione 27

6 LE RICERCHE 27

6.1 Ricerca sull'intero corpus 27
6.1.1 Risultato della ricerca 28
6.2 Ricerca sul materiale etichettato 28
6.2.1 Il risultato della ricerca 29

7 CONCLUSIONI 29

PHN <-> WRD 30

1 Scopi e finalità

dbCLIPS è un database il cui scopo è un'adeguata rappresentazione del corpus CLIPS. Gli scopi primari sono relativi all'archiviazione ed alle modalità di consultazione dei dati.

1.1 Archiviazione

Memorizzare i dati del corpus CLIPS in una struttura semplice, altamente efficiente, flessibile nella consultazione e capace di interagire facilmente con altre strutture di rappresentazione e in particolare con quelle *XML-Based*.

Per "dati CLIPS" si intendono file di registrazioni, file segmentati, file di trascrizione e file contenenti indicazioni di etichettatura. La descrizione dettagliata è riportata nel capitolo [2]. Tali file sono affiancati da un database, relativo al solo materiale segmentato ed etichettato, in cui sono memorizzati una serie di informazioni in forma strutturata finalizzati all'utilizzo da parte di ricercatori e studiosi.

1.2 Consultazione

Consentire agli utenti del sistema la possibilità di effettuare ricerche rapide e precise su un fenomeno o su un insieme di fenomeni, a qualsiasi livello di etichettatura, avendo la possibilità di reperire tutte le informazioni relative al fenomeno stesso e al suo corretto posizionamento anche in relazione agli altri livelli di descrizione.

In aggiunta è possibile interrogare il database per avere indicazioni sul materiale disponibile che risponde a particolari criteri.

1.3 Le entità fondamentali

Sono diverse le entità che devono trovare adeguata rappresentazione in dbClips. In particolare sono rappresentati il materiale, dialogico e di altro genere, e tutti i livelli di etichettatura previsti.

Non tutte le informazioni presenti nel lavoro di trascrizione ed etichettatura sono affidabili. Le entità di dbCLIPS, invece, fanno riferimento soltanto ad informazioni, misure e trascrizioni di sicura affidabilità[1].

[1] Sono ovviamente tenuti in conto errori non critici – e comunque individuabili secondo criteri ben definiti e recuperabili – quali disallineamenti temporali fra episodi che invece devono essere allineati, etc...

2 Il materiale

Con tale termine viene identificata l'unità fondamentale da cui vengono derivate tutte le entità che costituiscono alla formazione del corpus nonché alla sua rappresentazione.

Il materiale, per come organizzato, è strutturato in file i cui dettagli sono riportati nel paragrafo seguente.

2.1 Corpora

L'intero materiale di CLIPS è suddiviso nei seguenti corpora:

(a) Radiotelevisivo

(b) Dialogico

(c) Letture

(d) Telefonico

(e) Ortofonico

2.1.1 Frequenza di campionamento

Il file originari di tutti i corpora, ad eccezione di quello telefonico, sono campionati a 22 KHz e sono in formato WAV con codifica standard a 16 bit. Nel caso del telefonico la frequenza di campionamento è di 8 KHz e la codifica è μLaw. La stessa osservazione vale anche per i file segmentati.

2.2 Struttura del materiale

I dettagli di analisi, quali trascrizioni ed etichettatura, non sono disponibili per tutto il materiale costituente l'intero CLIPS ma solo per un suo sottoinsieme selezionato. Per tale motivo la gestione dei dati CLIPS dipende proprio dalla disponibilità delle informazioni addizionali a partire dalla trascrizione e dalla segmentazione.

In sintesi, il materiale disponibile si può suddividere nelle quattro categorie seguenti:

- **Raccolto (R)**. Materiale di cui esiste la registrazione sonora nella sua integrità.
- **Trascritto (T)**. Materiale di cui esiste una trascrizione (TXT).

- **Segmentato (S)**. Materiale di cui è stata effettuata la segmentazione (in turni, ad esempio). Esistono di conseguenza tanti file acustici (WAV, ad eccezione del telefonico) quanti sono i segmenti prodotti.

- **Etichettato (E)**. Materiale di cui esistono un insieme di file che lo etichettano ai diversi livelli previsti dai documenti progettuali.

A seconda dei corpus si possono verificare delle condizioni in cui una categoria può includere o coincidere con quella immediatamente adiacente.

In particolare, occorre precisare che per ciò che riguarda le relazioni di inclusioni relative a tali suddivisioni vale quanto di seguito specificato.

Per i corpora (a) (b) e (c) accade che

$$\mathbf{R} \supseteq \mathbf{T} \supseteq \mathbf{S} \equiv \mathbf{E}$$

Per i corpora (d) ed (e)

$$\mathbf{R} \equiv \mathbf{T} \equiv \mathbf{S} \supseteq \mathbf{E}$$

Di conseguenze per ogni corpus è possibile che si verifichi, in generale, la situazione seguente:

(a) sia presente un solo file WAV[2] che costituisce la registrazione sonora nella sua globalità.

(b) sia presente un file WAV, registrazione integrale, e un file TXT che è la relativa trascrizione.

(c) sia presente il file WAV della registrazione integrale, il file TXT e i file risultanti dall'operazione di segmentazione che sono, tipicamente, dello stesso formato di quello della registrazione iniziale.

(d) sia presente il file WAV della registrazione integrale, il file TXT della trascrizione e i file di segmentazione unitamente a quelli relativi all'etichettatura

2.3 Tipi di file

Da un punto di vista fisico un "materiale" è costituito da un certo numero di file i cui dettagli sono riportati nei punti seguenti.

[2] il file è di tipo RAW nel caso del corpus telefonico, ma la sostanza delle considerazioni è identica.

- Un file acustico (WAV o RAW) che costituisce la registrazione sonora nella sua globalità e il cui nome identifica univocamente il materiale stesso con un'opportuna codifica.

- Un file *.TXT, con il medesimo nome del precedente, contenente la descrizione del materiale e la relativa trascrizione. Tale file, ove previsto, incorpora anche il contenuto delle turni dialogici che compongono il materiale. E' estremamente importante per il reperimento delle informazioni di base del materiale.

- Un file *_TEMPI.TXT contenente la temporizzazione delle interazioni con accuratezza al millesimo di secondo. L'affidabilità di tali informazioni non è tale da poterle adottare come riferimento attendibile, per cui nel processo di popolazione della base dati tale file verrà ignorato e non ne sarà richiesta neanche la presenza.

- Nel caso di materiali segmentati, saranno presenti tanti file sonori quanti sono i turni dialogici. È a questi ultimi che fanno riferimento i valori temporali – in campioni - riportati nei file di etichettatura.

Nel caso di materiale etichettato, per ciascuna interazione (parola, turno dialogico, etc...) sono presenti i file riportati di seguito. Il contenuto di ogni tipo di file è descritto nel documento **W1-a4** ["Protocolli di Trascrizione e Annotazione" a cura di *Renata Savy*].

- Un file WR_ [*opzionale*] che, derivato dal TXT iniziale e prodotto automaticamente, contiene la *trascrizione ortografica* di ogni singolo turno o interazione. Il file contiene anche informazioni sulla sovrapposizione di interazioni, non quantificabili in quanto privi di riferimenti temporali.

- Un file WRD contenente l'etichettatura ortografica del singolo turno. L'etichettatura è effettuata manualmente e tipicamente è segmentata per "parole" con indicazione anche di eventuali fenomeni di incertezza di confine tra parole adiacenti. Il simbolo che individua l'incertezza è [%]. C'è la possibilità, quindi, che in una singola cella si faccia riferimento a due parole consecutive.

- Un file ST_ [*opzionale*] che contiene la trascrizione fonologica corrispondente a WR_. Il file è generato automaticamente.

- Un file STD che contiene la trascrizione fonologica corrispondente a ogni singolo evento in WRD.

- Un file PHN che rappresenta l'etichettatura a livello fonetico.

- Un file ACS che, invece, rappresenta l'etichettatura a livello acustico.

- Un file WAV, che è in realtà un'estrazione di quello descritto in precedenza, che è riferito al singolo turno.

In aggiunta a questi potrebbero essere presenti altri file:

- Un file di etichettattura ADD [*opzionale*] che contiene indicazioni supplementari e in particolare indicazioni circa la sovrapposizione fra livelli. Non esiste un corrispondente file AD_.

- Un file FRM che contiene l'elenco delle formanti ricavate con un intervallo prefissato e relativo all'intero dialogo.

2.4 Dislocazione fisica del materiale

La dislocazione fisica del materiale, sia quello raccolto che risultato dall'elaborazione, riveste un'importanza notevole per quanto riguarda le eventuali estrazioni di corpus ridotti secondo criteri di varia natura.

L'organizzazione adottata deriva dalla proposta *Savy-D'Anna* riassunta nella tabella seguente:

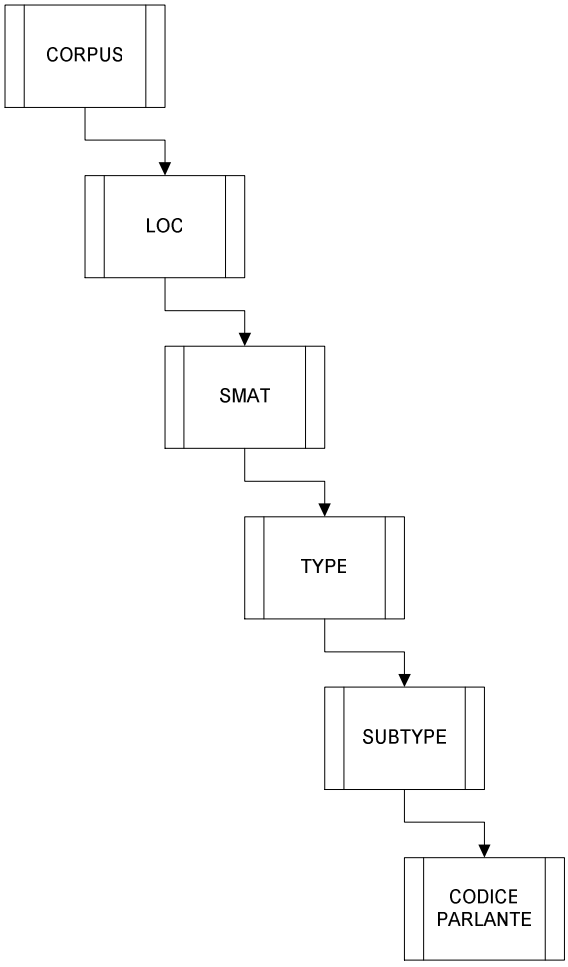
Liv.	Denominazione Livello					
1°	**<CORPUS>**	**(a) radiotelevisivo**	**(b) dialogico**	**(c) letto**	**(d) telefonico**	**(e) ortofonico**
2°	**<LOC>**	ROMA (R) MILANO (M) NAPOLI (N) …… NAZIONALE (Z)	ROMA (R) MILANO (M) NAPOLI (N) ……	ROMA (R) MILANO (M) NAPOLI (N) ……	ROMA (R) MILANO (M) NAPOLI (N) ……	NAZIONALE (Z)
3°	**<SMAT>**	SEGMENTATO (Turni / Enunciati) NON SEGMENTATO	SEGMENTATO (Turni) NON SEGMENTATO	SEGMENTATO (Frasi) NON SEGMENTATO	SEGMENTATO (Turni)	SEGMENTATO (Frasi)
4 °	**<TYPE>**	RADIO TV	MAPTASK TEST_DIFFERENZE	LISTA_FRASI LISTA_OGGETTI_MAPTASK LISTA_OGGETTI_TESTDIFF	AUTOMATICO WOZ (Manuale)	LISTA_FRASI_PROFESSIONISTI LISTA_FRASI_BILANCIATE
5°	**<SUBTYPE>**	DC IT IS PB	//	//	MASCHI FEMMINE	MASCHI FEMMINE
6°	**<CODICE PARLANTE>**	//	//	//	Numero di 5 cifre	F1,F2,…,F10 M1,M2,…,M10

Tabella 1 - Struttura del materiale CLIPS

Nella struttura complessiva del sistema, di conseguenza, si è assunto che il materiale sia classificato, dal punto di vista della sua memorizzazione, e organizzato gerarchicamente secondo il criterio illustrato nella figura seguente e derivato, per l'appunto, dalla tabella precedente.

A ogni livello gerarchico corrisponde una directory fisica sul dispositivo di memoria di massa. Opportune *utilities* consentiranno di estrarre o un intero corpus o tutto i corpora relativi ad una determinata località. Non sono emerse necessità di estrazione di

materiale con criteri diversi dal corpus o dalla località per le quali valga la pena di predisporre automatismi.



La radice gerarchica è costituita dal corpus (RADIOTELEVISIVO | DIALOGICO | ...) che quindi – in questi termini - è anche visto come un'unità autonoma con una propria strutturazione.

La località di registrazione ([LOC]) costituisce il secondo livello della struttura ed è immediatamente seguita dall'identificazione del materiale ([SMAT][3]) sulla base del fatto che sia o meno segmentato. Si distingue, in tale livello, la possibilità che si abbia a disposizione il materiale non segmentato (lunghe registrazioni integrali) e quello segmentato che, quindi, è stato trascritto o etichettato. La denominazione del materiale dipende, ovviamente, dal corpus di appartenenza.

Si noti inoltre che nella cartella del materiale NON SEGMENTATO (NSEG), in realtà, sono presenti TUTTE le registrazioni integrali[4] comprese, quindi, anche quelle che sono state successivamente segmentate.

Nella cartella SEGMENTATO (SEG), di converso, sarà presente soltanto il materiale prodotto dalla segmentazione. In altri termini conterrà tutto il materiale acustico (WAV o RAW nel caso del telefonico) segmentato ed eventualmente etichettato, unitamente ai file addizionali.

Il IV livello è relativo alla tipologia del materiale ([TYP]). A parte alcune denominazioni comuni (come "lista frasi" nel *Letto* e nell'*Ortofonico*), in generale le tipologie dipendono dal corpus.

Nel V livello si ha la distinzione in sottotipi, ove questa sia presente mentre il VI e ultimo livello sarà differenziato sulla base del codice del parlante (per i corpora per i quali questo è previsto).

[3] [SMAT] ::= SEGMENTATO | NON-SEGMENTATO.

[4] In altri termini contiene tutto il materiale acustico (WAV o muLaw)

2.4.1 Classificazione del materiale e denominazione directory

CORPUS ::=

- Radiotelevisivo (RT)
- Dialogico (DG)
- Letto (LT)
- Telefonico (TL)
- Ortofonico (O)

I nomi delle directory relative sono per esteso e coincidono coi nomi della lista precedente.

LOCALITÀ

LOC ::= B | C | D | .... | Z

CITTA'	SIGLA
Bari	B
Cagliari	C
Bergamo	D
Parma	E
Firenze	F
Genova	I
Catanzaro	H
Lecce	L
Milano	M
Napoli	N
Perugia	O
Palermo	P
Roma	R
Torino	T
Venezia	V
Nazionale	Z

Anche nel caso delle località i nomi delle relative directory sono riportati per esteso.

SEGMENTAZIONE MATERIALE

SMAT ::= SEG | NSEG

Il nome della directory relativa consiste esattamente in uno dei due valori appena riportati.

TIPO MATERIALE [TYPE]

Come già anticipato, dipende dal corpus secondo la tabella seguente:

Corpus	Valori di [TYPE]
Radiotelevisivo	TYPE ::= RADIO (**RD**) | **TV**
Dialogico	TYPE ::= MAPTASTK (**MT**) | TEST_DIFFERENZE (**TD**)
Letto	TYPE ::= FRASI (**LF**) | OGGETTI_MAPTASK (**LM**) | OGGETTI_TESTDIFFERENZE (**LT**)
Telefonico	TYPE ::= AUTOMATICO (**AUTO**) | WOZ (**WOZ**)
Ortofonico	TYPE ::= FRASI (**LP**) | FRASI_BILANCIATE (**LB**)

In questo caso i nomi utilizzati, per le directory relative, sono quelli delle sigle riportati in grassetto.

SUBTYPE

Anche in questo caso c'è una forte dipendenza dal corpus. I valori possibili sono riportati nella tabella seguente.

Corpus	Valori di [SUBTYPE]
Radiotelevisivo	SUBTYPE ::= **DC** | **IT** | **IS** | **PB**
Telefonico	SUBTYPE ::= **M** | **F**
Ortofonico	SUBTYPE ::= **M** | **F**

Anche per il *subtype* come per il *type*, per i nomi delle directory si utilizzeranno le sigle.

CODICE PARLANTE

Il codice parlante è presente solo nel caso dell'ortofonico e del telefonico. La tabella seguente riassume i possibili valori

Corpus	
Telefonico	CODICEPARLANTE ::= XXXXX X ::= [0-9]
Ortofonico	CODICEPARLANTE ::= [F1 ... F10] | [M1 ... M10]

In tal caso occorrerà valutare la necessità di directory separate denominate utilizzando il codice del parlante.

2.4.2 Esempi

Nel presente paragrafo sono riportati alcuni esempi di nomi di files che soddisfano le regole di classificazione e dislocazione enunciate.

- /dialogico/Roma/NSEG/TD/DGtdB04R.WAV
- /radiotelevisivo/Palermo/SEG/RADIO/DC/RDdc_01P.TXT

3 Vincoli di etichettatura

I materiali sono segmentati ed etichettati manualmente da personale specializzato secondo modalità specifiche eventualmente utilizzando specifici tools.

Uno dei problemi più importanti, ai fini di una corretta archiviazione ed il recupero delle informazioni relative all'etichettatura, è quello relativo alle condizioni di allineamento dei segmenti prodotti ai vari livelli.

Nei paragrafi seguenti sono riportati i vincoli emersi dalle specifiche e dall'analisi successiva del materiale.

3.1.1 WRD <-> STD

I livelli WRD ed STD sono sempre *allineati* fra di loro: i marker di inizio e fine coincidono per ogni singola etichetta.

3.1.2 WRD|STD <-> PHN

La condizione generale, in questo caso, è che a un evento WRD corrispondono uno o più segmenti del livello PHN.

Tuttavia, per ciò che concerne l'allineamento tra gli eventi correlati, ci sono da fare alcune considerazioni in quanto, in generale, un evento WRD|STD non è sempre allineato con quelli relativi nel livello PHN.

Parliamo di *allineamento* quando un evento WRD è allineato con gli eventi PHN nei quali può essere "scomposto". È questo il caso illustrato nella Figura 1 nella quale un'etichetta di livello WRD|STD ([Giordano]) corrisponde a più etichette al livello PHN. Esistono, per l'intero gruppo di etichette, delle condizioni di coincidenza dei marker di inizio e fine, rispettivamente, per la prima e l'ultima etichetta del livello PHN corrispondenti a quella del livello WRD.

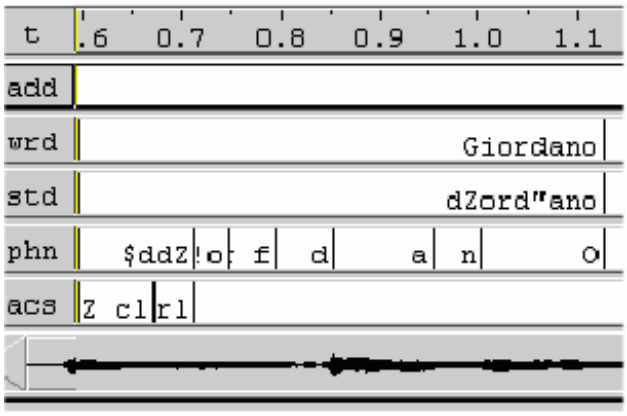


Figura 1 - Allineamento tra livelli WRD e PHN

Ci sono dei casi, però, per i quali la condizione di allineamento, per come definita nel capoverso precedente, viene a mancare. Si tratta dei casi di incertezza del confine tra due parole a livello WRD che danno origine a situazioni simili a quella riportata di seguito, in cui due etichette del livello WRD ([va'%] e [%allora]) :

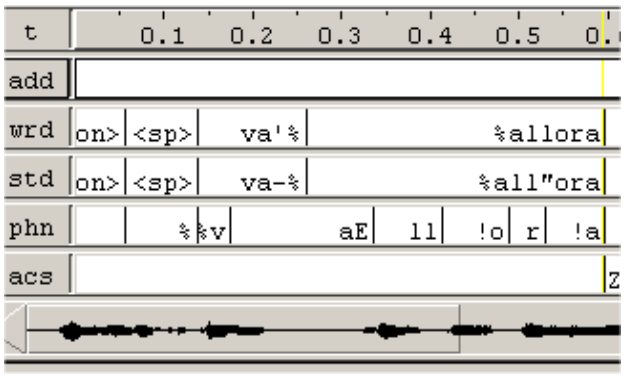


Figura 2 - "Non allineamento" tra WRD e PHN

In tali casi, occorre notare, che comunque – pur trattandosi di casi di non allineamento – è possibile trovare dei criteri di coincidenza in cui due – e in teoria anche più -

etichette a livello WRD sono *globalmente* allineate (all'inizio e alla fine della sequenza) con un certo numero di etichette a livello PHN. In questa situazione, ovviamente, uno o più marker WRD risultano compresi tra coppie di marker a livello PHN, nessuno dei quali coincide con uno di quelli di partenza.

È sempre possibile, con un algoritmo peraltro non complesso, recuperare le informazioni relative all'allineamento "minimo" da rispettare e credo che tali informazioni potrebbero, magari non in prima realizzazione, essere utili come risultato ulteriore di una ricerca se uno degli eventi trovati è indicativo di una situazione simile.

3.1.3 PHN <-> ACS

I file descrittivi del livello ACS, nonché i relativi descrittori interni, sono presenti soltanto nel caso in cui nel turno dialogico corrispondente sono presenti fenomeni di occlusive e affricate. In caso contrario non esiste il file .ACS.

Solitamente c'è una coincidenza tra un'etichetta a livello PHN e uno o due etichette al livello ACS. Occorre però tenere presente che per poter indicare il punto di inizio della prima etichettatura è necessario inserire un marker di inizio che corrisponda con il marker di fine di qualcosa non direttamente pertinente al fenomeno ACS sotto esame. Tale fenomeno, utilizzato solo a scopi pratici, è indicato come contenuto "vuoto" col simbolo "[_]".

Nei casi ordinari, nel livello ACS – in corrispondenza di un evento PHN – sono presenti due etichette: una relativa alla *closure* e una relativa alla *release*, come esemplificato nella figura seguente.

Non è esclusa, sebbene rara, la presenza di casi in cui esista solo un etichetta, relativa o alla *closure* o alla *release.*

3.1.4 ADD

Il livello ADD, opzionale, è destinato a inglobare un certo numero di informazioni addizionali, come ad esempio <noise> o comunque delimitati mediante <>|[], nel caso in cui queste identificano fenomeni in sovrapposizioni con altri eventi[5].

[5] Nei casi in cui questo non accade, ossia nel caso in cui sono presenti tali eventi ma non in sovrapposizione, questi avranno un loro "spazio" sul segnale e quindi nell'etichettatura WRD.

Inoltre sono segnalate informazioni circa la sovrapposizione di turni (##), quando si tratta di interazioni fra più parlanti.

Infine esso può spaziare su più parole del livello WRD|STD e in tali casi risulta allineato, globalmente, con questo.

3.2 Tipologia del materiale

Le principali tipologie di materiale sono riportate nella tabella seguente unitamente ai codici con i quali sono identificate.

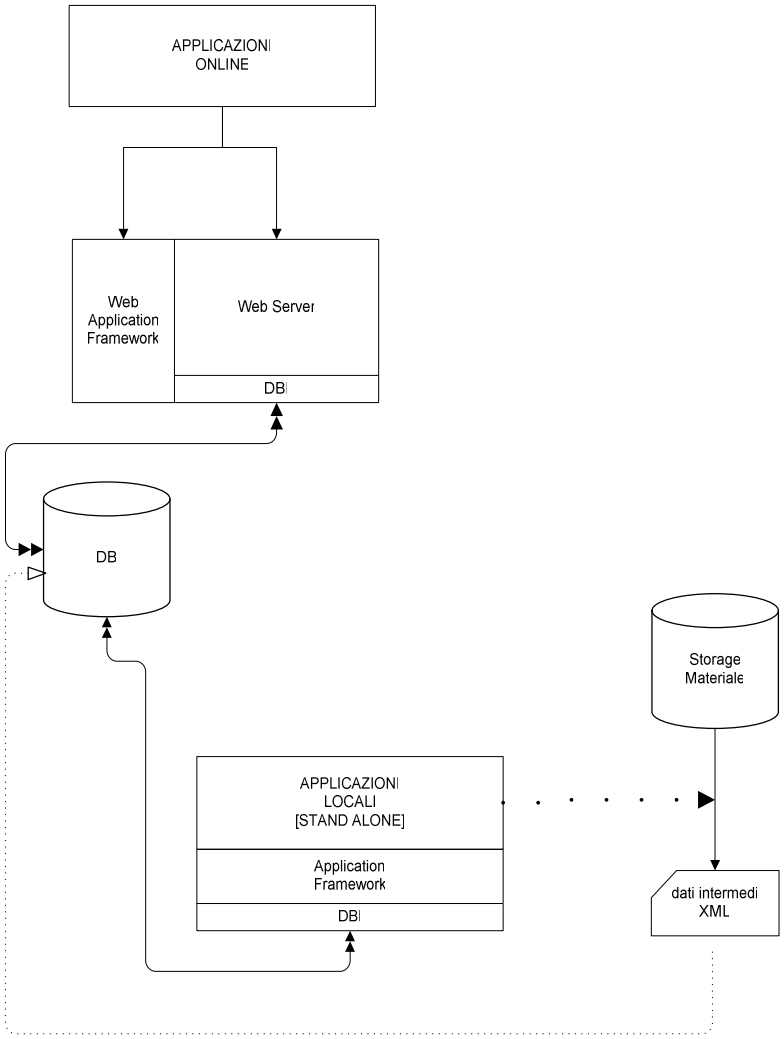
Le sigle MAT e TRM fanno riferimento a quanto riportato nel file di trascrizione .TXT.

CORPUS	**MAT**	**TRM**	
Dialogico	DG	MT	Dialoghi:
Dialogico	DG	TD	Dialoghi:
Letto	LF		Letture [Frasi] Letto
Letto	LM		Letture [Parole, Map Task] Letto
Letto	LT		Letture [Parole, Test alle differenze]
Ortofonico	LP	M|F	Ortofonico [Lista Frasi Professionisti]
Ortofonico	LB		Ortofonico [Lista Frasi Bilanciate]
Radiotelevisivo	RD	DC	Radiofonico: [Divulgazione e cultura]
Radiotelevisivo	RD	IS	Radiofonico: [Informazione e Servizi]
Radiotelevisivo	RD	IT	Radiofonico: [Intrattenimento]
Radiotelevisivo	TV	DC	Televisivo: [Divulgazione e cultura]
Radiotelevisivo	TV	IS	Televisivo: [Informazione e Servizi]
Radiotelevisivo	TV	IT	Televisivo: [Intrattenimento]
Telefonico	TL	M|F	Telefonico

Come già, anticipato, la tipologia di materiale – ovvero il corpus di appartenenza dello stesso - determina la sua frequenza di campionamento.

4 Architettura del sistema

L'architettura del sistema, nella sua struttura generale, è riportata nella figura seguente.



L'interfaccia verso l'utenza esterna è costituita da un insieme di *applicazioni online* che consentono di effettuare le ricerche descritte nel capitolo dedicato. Tali applicazioni girano nel contesto costituito da un *Web Server* e da un *Application Framework* – ovviamente compatibile con web server in questione – che ne condizionano, ovviamente, il linguaggio in cui sono scritte.

Le applicazioni online effettuano tipicamente ricerche sul database (DB) in cui sono memorizzate le entità che costituiscono la realtà CLIPS. La connessione al database avviene mediante un'interfaccia dipendente dal sistema operativo e dall'*application framework* scelto.

Il DBMS che ospita il database CLIPS è stato scelto tra quelli che dispongono di un'interfaccia, possibilmente standard, verso i web server e gli application server più diffusi, oltre che per le funzionalità offerte e per le prestazioni.

Tutto il materiale relativo ai corpora è memorizzato su un file system raggiungibile dal web server e da un insieme di applicazioni locali che sono finalizzate principalmente al parsing del materiale stesso e al popolamento della base dati. Il risultato di tali applicazioni è la produzione di un certo numero di documenti XML che rappresentano, in

maniera intermedia, la base dati. I dati rappresentati mediante tali documenti possono essere importati successivamente nel DB.

La presenza di tali documenti XML consente, nel caso ve ne fosse necessità, di modificare in tempi sufficientemente rapidi il DBMS utilizzato ripopolandolo con procedure che non hanno bisogno di interventi se non di tipo marginale.

Il file system su cui è memorizzato il materiale può non essere, in linea di principio, locale a quello delle applicazioni online e al web server. Dato che, però, le applicazioni possono accedere a tale materiale è consigliabile che il file system relativo sia locale in maniera da non degradare le performance dell'intero sistema.

L'architettura qui descritta è abbastanza flessibile da poter inglobare diverse tecnologie anche eterogenee. Ovviamente non tutte le tecnologie sono, al momento, direttamente interscambiabili: ad esempio non è possibile, ma lo sarà in un futuro piuttosto vicino, utilizzare *Apache* come web server e *.NET Framework* come application server.

4.1 Configurazione adottata

In considerazione delle risorse disponibili e dei tempi di sviluppo piuttosto compressi è stata decisa la configurazione seguente.

Sistema Operativo Host. Windows 2000 Server. Il sistema è già disponibile e non costituisce, quindi, un aggravio economico.

Web Server. Il web server è quello incluso in Windows 2000 Server noto col nome di Internet Information Server.

Application Server. Come server applicativo è stato scelto .NET Framework e di conseguenza le applicazioni online saranno costituite da moduli ASP.NET che si interfacciano direttamente col resto del sistema mediante il .NET Framework. Anche in questo caso la scelta non costituisce un aggravio di costi. Tale framework consente, tra l'altro, lo sviluppo di applicazioni stand-alone utilizzando le stesse classi e gli stessi namespace anche in questo tipo di applicazioni.

C'è da dire che la scelta di .NET Framework vincola, allo stato attuale, il sistema operativo host tra quelli di classe Windows 2000 o Windows 2003. Tuttavia lo stato di avanzamento del progetto *Mono* (http://go-mono.com) e del progetto DotGNU (http://ww.gnu.org) fa ritenere che entro breve tempo, il .NET Framework sarà disponibile in release stabile anche su sistemi operativi Unix e Linux consentendo, a questo punto, la totale interscambiabilità dei componenti del sistema CLIPS.

DATABASE. Considerando quanto pubblicato di recente sui canali specializzati, la scelta del DBMS è stata indirizzata verso SAP-DB un DBMS open source il cui sviluppo e supporto sono assicurati direttamente da una grande azienda (SAP Ag). Nel corso dello sviluppo tale DBMS è passato sotto la gestione e la distrubizione di MySQL AB col nome

di MaxBD. Per ciò che concerne la semplice memorizzazione dei dati[6] occorre sottolineare che il DBMS è facilmente scambiabile con altri open source. In entrambi i casi l'interfaccia privilegiata è quella standard ODBC e tutti e due i DBMS sono disponibili su sistema operativo Linux.

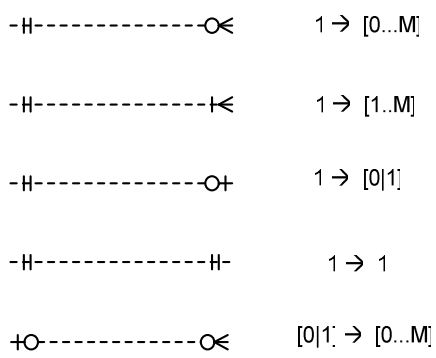
5 Struttura della base dati

La struttura logico-concettuale della base dati è riportata nella Figura 3. Nel diagramma ER raffigurato sono riportate tutte le entità coinvolte in dbCLIPS.

La struttura elaborata consente di inglobare, in una singola struttura coerente, sia il materiale non segmentato che quello segmentato, trascritto ed etichettato secondo quanto già premesso in [2.2].

I dettagli relativi alle singole tabelle, attributi, relazioni e indici sono riportati nel "*DBMS Report*" allegato al presente documento.

La cardinalità dettagliata delle relazioni è indicata secondo il criterio riportato nella figura seguente:



Sono inoltre riportate – sempre nel medesimo diagramma - gli attributi col ruolo di chiave primaria (PK), di chiave esterna (FK) e quelli che contribuiscono alla definizione di un indice (I).

[6] escludendo quindi eventuali interfacce applicative del DBMS (stored procedures, ad esempio).

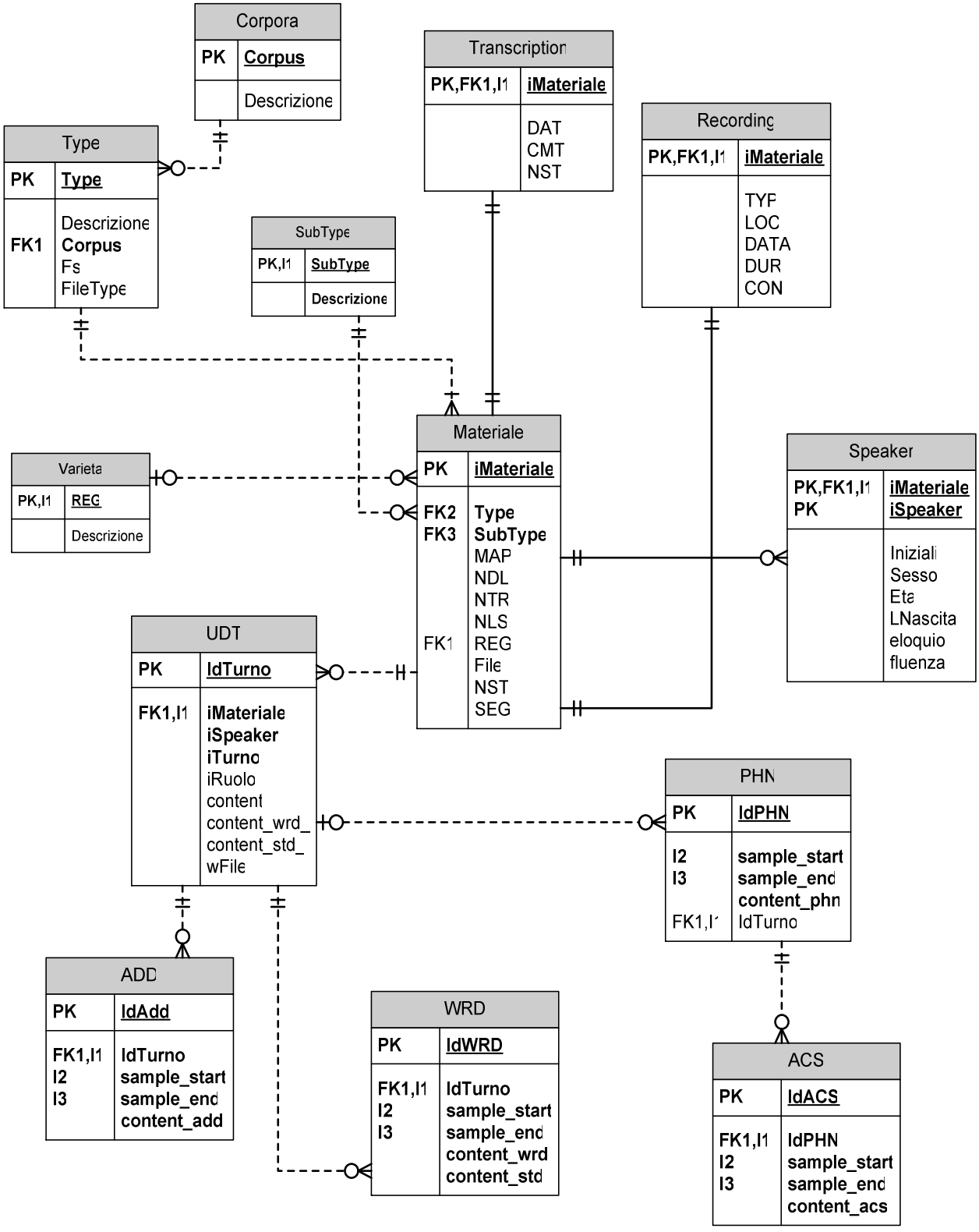


Figura 3 - Struttura logico-concettuale di dbCLIPS

Oltre alle relazioni esplicitamente indicate nel diagramma è possibile, occasionalmente, instaurare altre relazioni (come ad esempio tra gli attributi 'sample_start') oltre a quelle previste dal modello citato.

5.1 Commenti alla struttura

In questo paragrafo sono i commenti essenziali alla struttura con riferimento alle entità coinvolte. Commenti dettagliati sui campi sono riportati nel DBMS Report allegato.

5.1.1 Vincoli

Le tabelle **`Varieta`**, **`Type`** e **`SubType`** sono utilizzate anche come riferimenti di vincoli di valore per i corrispondenti attributi dell'entità **`Materiale`**.

In Type, tra l'altro, compare anche la frequenza di campionamento del materiale in questione nonché il tipo di file (WAV | RAW) relativo.

5.1.2 MATERIALE

Ogni materiale è identificato completamente da un record di [Materiale]. Questa tabella riporta le informazioni essenziali per la caratterizzazione.

Le informazioni essenziali sono:

Type	Tipologia del materiale. Fa riferimento ai valori della tabella **`Type`** e a quanto specificato nello schema a pagina [**Errore. Il segnalibro non è definito.**]. In pratica si tratta delle suddivisioni del materiale di un certo corpus. Quest'ultimo è specificato, per l'appunto, tra gli attributi di **`Type`**.
SubType	Tipo trasmissione o ulteriore tipologia del materiale. Non è presente per tutti i corpus. Anche per tale attributo si faccia riferimento alla tabella a pagina [**Errore. Il segnalibro non è definito.**].
MAP	Mappa
NDL	Numero Dialogo
NTR	Numero Trasmissione
NLS	Numero Lista
REG	Luogo di registrazione
NST	Numero Interazioni[7]
SEG	Segmentato (1) – Non Segmentato (0)
File	Nome del file del materiale

[7] A rigore, questo è un elemento di [Transcription] ma poiché i dati di questa entità non sono considerati vincolanti e, in prima realizzazione, possono anche essere omessi, si è deciso di duplicare tale informazione per comodità e semplicità di trattamento.

5.1.3 Speaker

Per ogni record di [Materiale] ci sono uno o più [Speaker]. In particolare deve essere presente almeno un record che identifica, anche con informazioni parziali, uno speaker – eventualmente fittizio – coinvolto nei turni o nelle interazioni o letture.

Le informazioni che caratterizzano uno speaker sono riportate nella tabella seguente:

Iniziali	Iniziali dello Speaker
SESSO	M | F
LNascita	Luogo di Nascita
eta	Età
eloquio	
fluenza	

5.1.4 UDT (Unità di Trascrizione) o Turni

Tale entità è stata così denominata sulla base del citato documento di progetto **W1-a4**.

In dbCLIPS, come nell'intero CLIPS, un materiale può comparire soltanto come file sonoro – non segmentato né, ovviamente, trascritto – e come tale esso sarà presente come un singolo record in [Materiale].

Ogni [Materiale] segmentato può essere composto da un certo numero di unità di trascrizione. Ogni unità di trascrizione può essere identificata univocamente mediante gli attributi seguenti:

- [Materiale] di appartenenza
- [Speaker]
- [Ruolo] dello Speaker
- [Numero dell'interazione o Turno]

Queste informazioni sono sempre disponibili – nel file di trascrizione o nel nome stesso del file – ad eccezione dell'ultima che non è riportata nel caso di elenchi di oggetti. In tal caso viene ricostruita automaticamente mediante un numero progressivo.

Ad ogni modo si è preferito utilizzare, per l'identificazione interna alla base di dati, un numero univoco – che potrebbe anche essere calcolato con un'opportuna codifica - in maniera tale da non pregiudicare l'efficienza della struttura.

IdTurno	Identificativo numerico del turno
iMateriale	Identifica il materiale
iSpeaker	Identifica lo speaker
iTurno	Numero progressivo di ogni singola interazione
iRuolo	F|G, nel caso di Map Task
Content	testo della trascrizione, per come riportato nel file di trascrizione.
content_wrd	Trascrizione completa dell'interazione secondo quanto contenuto in WR_.
content_std	Trascrizione completa dell'interazione secondo quanto contenuto in ST_. Non necessario, in realtà, né previsto nei documenti.
wFile	Nome del file WAV associato all'interazione

5.1.5 Etichettature

Le tabelle descritte nei paragrafi successivi sono destinate a contenere le etichettature ai vari livelli.

La stringa di etichettatura è contenuta in un campo con il prefisso content_. Il valore di tale campo in alcuni casi può essere nullo. In tal caso può essere riempito con il simbolo "__" oppure, più semplicemente, lasciato vuoto. Questo caso capita spesso, ad esempio, a livello ACS.

5.1.6 [ADD]

Riporta le singole informazioni addizionali per ogni interazione. In relazione 1-[0..M] con UDT. [iAdd] è un identificativo - progressivo o calcolato - per ogni singolo elemento di ADD.

IdAdd	Identificativo numerico [PK]
IdTurno	IdTurno corrispondente [FK]
sample_start	Campione iniziale
sample_end	Campione finale
content_add	Contenuto etichettatura

5.1.7 [WRD]

Riporta i singoli elementi dei livelli WRD e STD[8] relativi ad ogni interazione. In relazione 1-[0..M] con UDT. [IdWRD] è l'identificativo numerico univoco.

IdWRD	Identificativo numerico [PK]
IdTurno	IdTurno corrispondente [FK]
sample_start	Campione iniziale
sample_end	Campione finale
content_wrd	Contenuto etichettatura WRD
content_std	Contenuto etichettatura STD

5.1.8 [PHN]

Riporta i singoli elementi di PHN relativi ad ogni interazione che, si ricorda, non sono generalmente coincidenti – come delimitazione temporale – con WRD/STD. In relazione 1-M con UDT[9]. [IdPHN] è l'identificativo numerico univoco.

Si tenga presente che un record X di PHN è in relazione con un record W di WRD se:

sample_start(X) = sample_start(W)

oppure se

sample_start(W) <= sample_start(X) <= sample_end(W)

IdPHN	Identificativo Numerico [PK]
IdTurno	IdTurno corrispondente [FK]
sample_start	Campione iniziale
sample_end	Campione finale
content_phn	Contenuto etichettatura PHN

Tra PHN e WRD esiste comunque una "relazione" precisa che è descritta in appendice [0].

[8] Ricordo che sono riportati entrambi i contenuti in quanto sono, per definizione, sempre allineati a livello di campioni.

[9] Non può essere in relazione diretta con WRD a causa del disallineamento.

5.1.9 [ACS]

Riporta i singoli eventuali elementi di ACS relativi a un elemento PHN. In relazione 1-[0..M] con PHN. [IdACS] è l'identificativo numerico univoco.

Si tenga presente che un record X di ACS è in relazione con un record P di PHN se:

sample_start(X) = sample_start(P) OR sample_end(X) = sample_end(P)

IdACS	Identificativo Numerico [PK]
IdPHN	Record PHN corrispondente [FK]
sample_start	Campione iniziale
sample_end	Campione finale
content_acs	Contenuto ACS

5.1.10 Tabelle Supplementari

Le informazioni relative alle due tabelle seguenti, pur presenti, non sono considerate di importanza primaria. In particolare tali informazioni possono anche mancare senza pregiudicare in alcun modo la coerenza minima del database.

Entrambe le tabelle contengono un campo [iMateriale] che identifica il materiale cui fanno riferimento e col quale sono in relazione rigida 1-1.

[Recording]

Contiene le informazioni relative alla registrazione.

iMateriale	Identifica il materiale [PK, FK]
TYP	Tipologia Mezzo
LOC	Luogo di registrazione
DAT	Data di registrazione
DUR	Durata
CON	Condizioni di registrazione

[Transcription]

Contiene informazioni relative alla trascrizione

iMateriale	Identificativo Materiale [PK, FK]
DAT	Data
CMT	Commenti

NST	Numero Turni

5.2 Popolazione della base dati

La popolazione del database avviene mediante delle procedure automatiche che estraggono i dati dai file di cui al paragrafo [2.1].

E' preferibile che tali procedure piuttosto che essere a trasformazione diretta (dai file al database) introducano un livello intermedio. In particolare si ipotizza che una rappresentazione intermedia in XML dei dati strutturati sia particolarmente indicata perché da una parte introduce un livello di elasticità e di indipendenza dal modello adottato consentendo di avere a disposizione i dati "grezzi" – trattati ed elaborati in modo da ridurre le eventuali incoerenze e gli eventuali errori - e in un formato immediatamente importabile in un altro modello. Questo si traduce anche in una riduzione delle risorse necessarie anche in funzione di eventuali ripensamenti sulla struttura del database o sulla scelta del motore stesso.

Sulla base di quanto premesso, la popolazione della base dati avviene in due fasi. Nella prima fase i dati presenti nei file relativi al materiale vengono trasformati in documenti XML ben formati con un opportuno analizzatore.

I dati realmente utili alle fase di ricerca sono riportati nel paragrafo successivo.

La struttura di tali documenti non è necessariamente conforme alla struttura della base dati. Tale step viene effettuato in modo da segmentare il processo e avere una struttura dati intermedia insensibile alla struttura fisica della base dati operazionale.

Questo assicura la disponibilità di una struttura importabile in qualsiasi sistema senza particolari problemi legati alla struttura stessa.

Tuttavia è ovvio che la struttura dei documenti XML è tale da rispecchiare le informazioni relative al materiale nonché i vincoli che le contraddistinguono, per cui l'analogia con la struttura della base dati è piuttosto forte.

Esistono documenti XML per ognuno dei file

- **`Type`**,
- **`SubType`**,
- **`Varieta`**

che sono a contenuto predefinito, ricavabile dalla specifiche di progetto nonché dalla **Errore. L'origine riferimento non è stata trovata.**.

Si prevede, quindi, per ogni materiale la produzione dei seguenti file:

- MAT.XML. File relativo all'identificazione del materiale. Coinvolge tutte le tabelle "Materiale", "Transcription", "Speaker", "Recording". Non tutte le informazioni hanno la medesima importanza. In particolare sono assolutamente indispensabili quelle relative alle tabelle "materiale" e "speaker".
  Un esempio della struttura di tale documento è riportato di seguito. Si noti che il corpus è determinato dall'attributo <TYPE>.

```
<Materiale>
      <iMateriale>DGMTA01N</iMateriale>
      <Type>MT</Type>
      <MAP>A</MAP>
      <SubType> </SubType>
      <NDL>1</NDL>
      <REG>N</REG>
      <File>DGMTA01N.WAV</File>
      <NST>133</NST>
      <SEG>1</SEG>
</Materiale>

<Recording>
      <TYP>Cassetta DAT</TYP>
      <LOC>Napoli / Università </LOC>
      <DATA>2000-12-12 00:00:00</DATA>
      <DUR>5,45</DUR>
      <CON>Buone</CON>
</Recording>

<Transcription>
      <iMateriale>DGMTA01N</iMateriale>
      <DAT>2001-05-07T00:00:00</DAT>
      <NST>133</NST>
</Transcription>


<SPEAKER>
      <iMateriale>DGMTA01N</iMateriale>
      <iSpeaker>P1</iSpeaker>
      <Iniziali>F.P.</Iniziali>
      <Sesso>M</Sesso>
      <LNascita>Napoli</LNascita>
      <eta>28</eta>
      <eloquio><![CDATA[fluente G>F]]></eloquio>
</SPEAKER>
<SPEAKER>
      <iMateriale>DGMTA01N</iMateriale>
      <iSpeaker>P2</iSpeaker>
      <Iniziali>P.G.</Iniziali>
      <Sesso>M</Sesso>
      <LNascita>Napoli</LNascita>
      <eta>25</eta>
      <eloquio><![CDATA[fluente F>G]]></eloquio>
</SPEAKER>
```

- UDT.XML. Contiene le trascrizioni relativi ad ogni singola interazione. La struttura base del documento è riportata di seguito. Si noti che <iMateriale> e <iSpeaker> hanno valori determinati da quelli che compaiono nel file precedentemente descritto, mentre <IdTurno> è un valore che potrebbe essere

diverso da quello che comparirà nella base dati in quanto, pur essendo generato automaticamente, potrebbe essere basato su tecniche diverse di generazione.

```
<udt>
      <iMateriale>      </iMateriale>
      <iSpeaker>        </iSpeaker>
      <iTurno>          </iTurno>
      <iRuolo>          </iRuolo>
      <content>         </content>
      <time_start>      </time_start>
      <time_end>        </time_end>
      <content_wrd_ >   </content_wrd_>
      <content_std_>    </content_std_>
      <wFile>           </wFile>
</udt>
```

- WRD.XML. Contiene gli elementi relativi ai livelli WRD e STD di ogni singola interazione.

```
<wrd>
      <IdWRD>           </IdWRD>
      <IdTurno>         </IdTurno>
      <iRuolo>          </iRuolo>
      <time_start>      </time_start>
      <time_end>        </time_end>
      <content_wrd>     </content_wrd>
      <content_std>     </content_std>
</wrd>
```

- PHN.XML. Contiene gli elementi relativi al livello PHN. La struttura di base è riportata di seguito. Occorre sottolineare che IdPHN non sono necessariamente gli stessi valori che poi saranno assegnati all'inserimento nel database in quanto la sua generazione può essere basata su tecniche diverse.

```
<PHN>
      <IdPHN>               </IdPHN>
      <IdTurno>             </IdTurno>
      <sample_start>        </sample_start>
      <sample_end>          </sample_end>
      <content_phn>         </content_phn>
</PHN>
```

- ACS.XML. Contiene gli elementi relativi al livello ACS. Per IdACS vale il medesimo discorso fatto a proposito degli analoghi Id dei file precedenti.

```
<ACS>
      <IdACS>               </IdACS>
      <IdPHN>               </IdPHN>
      <sample_start>        </sample_start>
      <sample_end>          </sample_end>
      <content_acs>         </content_acs>
 </ACS>
```

- ADD.XML. Contiene gli elementi relativi alle informazioni addizionali. Anche nel caso di IdADD, per questa struttura, vale il discorso fatto nel caso delle strutture precedenti.

```
<ADD>
      <IdADD>          </IdADD>
      <IdTurno>        </IdTurno>
      <sample_start>   </sample_start>
      <sample_end>     </sample_end>
      <content_add>    </content_add>
</ADD>
```

Una volta formata la base dei documenti XML che rappresentano l'intero corpus, o un suo sottoinsieme autonomo, si potrà procedere mediante un'apposita applicazione alla popolazione della base dati vera e propria.

E' conveniente che i file WRD, ADD, PHN, ACS siano prodotti come un'unica entità fisica per tutto il dialogo cui fanno riferimento.

Si tenga presente che in questa fase occorrerà tenere conto dei vincoli relativi ad alcuni campi e alle relazioni stesse.

5.3 Validazione

Le procedure di trasformazione in XML che di importazione diretta devono avere un alto livello diagnostico. In particolare devono essere segnalate le eventuali mancanze di file, disallineamenti temporali e, in particolar modo, problemi di integrità referenziale che hanno origine nei dati grezzi.

6 Le ricerche

Il sistema dbCLIPS è corredato di due classi di strumenti di ricerca ed interrogazione. Il primo consente di interrogare il materiale del corpus al fine di estrarre dei riferimenti a suoi sottoinsiemi, secondo dei filtri attivabili dall'utente.

Il secondo ha la finalità primaria di consentire all'utente di identificare ed estrarre, dalla base di dati, uno o più fenomeni linguistici secondo vari criteri.

6.1 Ricerca sull'intero corpus

La ricerca sul materiale è possibile secondo i criteri seguenti:

- Corpus;
- Località;
- Tipo di materiale;
- Eventuale sottotipo;

Tutti gli elementi precedenti sono rappresentati con combo box contenenti i valori del dominio di ognuno di tali attributi.

6.1.1 Risultato della ricerca

In tal caso il risultato della ricerca è costituito dall'elenco dei file sonori che soddisfano i criteri specificati dall'utente.

Per ognuno dei file sarà visualizzato il nome e la directory di memorizzazione. Andrà valutata la possibilità di un link che consentirà all'utente di scaricare sul proprio PC ogni singolo file.

Il risultato della ricerca sarà paginato per evitare la produzione di pagine di dimensioni eccessive.

6.2 Ricerca sul materiale etichettato

I criteri di ricerca previsti possono essere di due tipi: fissi e specificati integralmente dall'utente.

Criteri fissi, specificati mediante 'combo box' a valori prefissati:

- Corpus;
- Tipo Materiale;
- Località

Criteri specificati dall'utente. Si tratta principalmente di espressioni contenenti etichette da cercare nei vari livelli.

L'utente sarà in grado di selezionare quali livelli includere nella ricerca e quali espressioni ricercare.

Queste ultime sono espressioni regolari costruite a partire da rappresentazioni di etichette, secondo una sintassi specificata in appendice.

6.2.1 Il risultato della ricerca

Indipendentemente dagli strumenti di ricerca, il risultato di ogni operazione di ricerca sarà costituito da un insieme di fenomeni a un certo livello.

Il primo risultato che quindi l'utente ottiene è una lista in cui sono specificati:

- Livello di etichettatura desiderato
- Tipo di Materiale
- Contenuto dell'etichettatura [opzionale nel caso di ricerche specifiche]

Accanto a tali dati possono essere fornite, presumibilmente su richiesta per evitare sovraccarichi inutili, informazioni relative a

- Fenomeno globale di appartenenza (dialogo, interazione)
- Livelli superiori di "inquadramento" del fenomeno, nel caso di fenomeni di livello inferiore (PHN, ACS,...)
- Fenomeni adiacenti, nel caso di eventi non allineati ai vari livelli

Infine, per ognuno di questi, possono essere estratti e trasferiti all'utente gli eventi acustici corrispondenti derivati dal file globale del materiale. Credo che questa sia una funzionalità non eccessivamente costosa dal punto di vista delle risorse e che qualificherebbe gli strumenti di ricerca in maniera notevole.

7 Conclusioni

Si è descritto, in questo documento, un modello di rappresentazione dell'intero insieme del materiale CLIPS. La rappresentazione include sia il materiale non segmentato che quello segmentato ed etichettato.

Per quest'ultimo è proposto un modello che ne consenta la ricerca dettagliata sulla base delle informazioni disponibili dalla trascrizione e dall'etichettatura, oltre che per i parametri fissi come il corpus, la città, la tipologia del materiale e altro ancora.

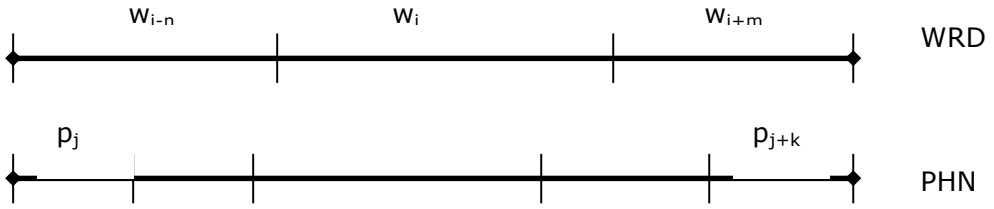
Il tutto è incastonato in un modello architetturale estremamente modulare in cui è stata posta la massima attenzione all'interscambiabilità dei vari moduli.

APPENDICI

PHN <-> WRD

Il livello PHN ha una sorta di dipendenza da quello WRD nel senso che dato un istante temporale t_i relativo a un evento phn_i, è sempre possibile trovare un limite inferiore di un evento PHN – precedente quello interessato - che coincide con quello di un evento WRD e analogamente per quello superiore, fino a realizzare punti di coincidenza.

In altri termini si possono verificare situazioni come quelle schematizzate qui in figura:



Più formalmente:

Per ogni *i* esistono *n,m,j,k* (maggiori o uguali a zero) tali che

$$start(wi-n) = start\ (pj)$$

e

$$end(wi+m) = end(pj+k)$$

Potrebbe quindi essere utile un meccanismo che:

1) Dato un evento a livello WRD|STD ricava gli eventi a livello PHN fino alla realizzazione della coincidenza

2) Dato un evento a livello PHN, ricava gli eventi a livello superiore fino alla realizzazione della coincidenza

Tali funzionalità possono trovare utile impiego non tanto nelle ricerche di eventi ma quanto nell'approfondimento delle stesse.

Package del sistema

Il package del sistema consiste in:

1. Set di DVD contenenti tutto il materiale suddiviso per corpus.
2. Struttura Applicazione online con documentazione per l'installazione e la configurazione. Il package contiene il .NET Frameword da installare e la documentazione relativa ai seguenti step:
    a. Setup di IIS;
    b. Installazione di .NET Framework;
    c. Installazione applicazione online;
3. Applicazioni locali per la preelaborazione del materiale. In particolare sono disponibili
    a. Applicazioni stand-alone per il parsing e la trasformazione del materiale in documenti XML
    b. Applicazioni per l'estrazione di un subset dell'intero storage del materiale
    c. Documentazione d'uso delle applicazioni
    d. Tutti i file accessori (DTD, etc.) necessari alle applicazioni.
4. DBMS e script accessori
    a. File necessari per l'installazione del DBMS. Documentazione per l'installazione del dbms in ambiente Windows.
    b. Data File già popolati, se consentito dal DBMS scelto.
    c. Script per la generazione della struttura dati (tabelle e quant'altro necessario)
    d. Applicazione stand-alone per il popolamento della base dati a partire dai documenti XML