Definizione ed annotazione prosodica di un database di parlato-letto usando il formalismo ToBI

Cinzia Avesani[+], Piero Cosi[+], Elisabetta Fauri*, Roberto Gretter*, Nadia Mana*, Silvia Rocchi*, Franca Rossi* e Fabio Tesser*

[+]ISTC-SFD
Istituto di Scienze e Tecnologie della Cognizione – Sezione di Fonetica e Dialettologia – CNR
Padova
e-mail:{avesani,cosi}@csrf.pd.cnr.it

*ITC-irst
Istituto Trentino di Cultura
Centro per la Ricerca Scientifica e Tecnologica
Povo, Trento
e-mail: {fauri,gretter,mana,sirocchi,frarossi,tesser}@itc.it

SOMMARIO

In questo articolo si descrive il lavoro realizzato per la definizione ed acquisizione di un database di parlato italiano letto, annotato a livello morfo-sintattico, a livello sintattico a costituenti e a livello prosodico secondo il formalismo ToBI. In particolare, vengono qui presentati il corpus, l'annotazione prosodica (convenzioni, metodologia e strumenti utilizzati), i risultati ottenuti ed alcune statistiche preliminari relative all'Intercoder Agreement.

INTRODUZIONE

Nell'ambito del Progetto europeo **M-PIRO** (Multilingual Personalised Information Objects[1]) è stato definito e acquisito un database di parlato italiano letto, annotato con ToBI (Tones and Break Indices). Partendo dall'assunto che l'applicazione di tale modello possa migliorare significativamente le prestazioni del sintetizzatore in termini di intonazione, espressività e naturalezza, lo scopo ultimo di tale lavoro è l'ottenimento di un modello prosodico utilizzabile da un sintetizzatore vocale. Postosi questo obiettivo, si è deciso di realizzare un database di parlato di circa un'ora, annotato a diversi livelli (Part Of Speech, sintattico e prosodico), per poi studiare le possibili relazioni tra le strutture morfo-sintattiche e quelle prosodiche e definire, sulla base di queste, il modello prosodico da utilizzare ai fini della sintesi.

[1] http://www.ltg.ed.ac.uk/mpiro/

Il database è stato realizzato a partire dalle registrazioni audio di tre racconti letti da uno speaker professionista, integrati da un insieme di frasi interrogative lette dallo stesso speaker, per un totale di 7285 parole - equivalenti a 67' 02" di audio.

Il corpus è stato annotato non solo livello prosodico secondo il sistema di trascrizione ToBIt (la versione ToBi per la lingua italiana), ma anche a livello Part Of Speech (POS) e sintattico.

In questo articolo illustriamo il lavoro svolto per la realizzazione di questo database, presentando obiettivi, definizione del corpus, criteri, metodologia e strumenti di annotazione, risultati ottenuti ed alcune statistiche.

1. OBIETTIVI

Lo studio e la ricostruzione dei modelli prosodici della lingua hanno un ruolo centrale nell'ambito delle tecnologie della voce, poiché la prosodia veicola informazioni importanti nel passaggio dallo scritto al parlato, informazioni di tipo linguistico, pragmatico, paralinguistico (affetti, attitudini) ed extralinguistico (sesso, età, stato di salute).

Come esempio dell'importanza dell'intonazione nel segnalare la struttura linguistica (sintattica e semantica) della frase, si considerino le seguenti frasi: (1) "*L'investigatore trovò la donna con i binocoli*" e (2) *"Giuseppe non beve perché è infelice"*. Le due frasi possono avere ciascuna due letture diverse benché sia la sequenza lessicale sia la struttura sintattica siano identiche[2]. Se la frase è scritta, le due letture non risultano disambiguabili a meno che il testo non sia corredato di annotazione POS. Se le frase viene pronunciata, invece, la prosodia dell'enunciato rende chiaro il significato che si vuole trasmettere (Avesani, 1999).

Lo scopo ultimo del lavoro qui descritto è l'ottenimento di un modello prosodico utilizzabile da un sintetizzatore vocale, partendo dall'assunto che l'applicazione di tale modello possa migliorare significativamente le prestazioni del sintetizzatore in termini di intonazione, espressività e naturalezza. Si è deciso quindi di costituire un database audio (di circa un ora) annotato a diversi livelli, per poi studiare le possibili relazioni tra le strutture morfo-sintattiche e quelle prosodiche e definire, sulla base di queste, il modello prosodico da utilizzare ai fini della sintesi. Inizialmente si è pensato di attingere all'archivio della RAI, in particolare ad alcune registrazioni dei giornalisti David Sassoli e Lilli Gruber. Tuttavia, dopo una prima esamina di questo materiale, lo si è ritenuto poco adatto agli scopi prefissati perchè caratterizzato da un'intonazione orientata a carpire l'attenzione dell'ascoltatore e quindi molto enfatizzata e poco naturale. In seconda istanza si è valutata l'ipotesi di lavorare su un database di fiabe registrate. Anche in questo caso però i problemi riscontrati sono stati di carattere linguistico e melodico: le fiabe infatti propongono un linguaggio che vuole essere semplice e suggestivo e l'inflessione del parlatore risulta inficiata – come nel caso precedente – dall'intento di catturare l'attenzione.
Alla luce di tutto ciò, la decisione finale ha portato all'utilizzo di un database letto da uno speaker professionista, costituito da alcuni racconti e da un gruppo di frasi interrogative.

[2] Le due letture sono, rispettivamente: (1a) L'investigatore riuscì a trovare la donna servendosi dei binocoli; (1b) L'investigatore riuscì a trovare la donna che aveva (sottratto) i binocoli; (2a) Giuseppe non beve e la ragione per cui non tocca alcool è dovuta alla sua infelicità; (2b) Giuseppe beve, ma la ragione della sua dipendenza dall'alcool non è legata alla sua infelicità.

2. DEFINIZIONE DEL CORPUS

Il database di parlato-letto è stato realizzato partendo dalle registrazioni audio di tre racconti letti da uno speaker professionista (Claudio Carini). I racconti utilizzati fanno parte di una serie di audiolibri disponibili in rete in formato audio (mp3) e testo (pdf), all'indirizzo http://www.ilnarratore.com/index.html.

Nella selezione del materiale audio a disposizione sul sito Internet si è prestata particolare attenzione alla qualità della voce dello speaker, affinché fosse ragionevolmente priva di inflessioni regionali. La scelta di attingere il materiale da tre testi uguali per genere e appartenenti allo stesso autore (per un totale di 371 frasi – 5335 parole) ha inoltre garantito una certa omogeneità di stile letterario e una buona varietà linguistica (quantomeno in termini di frasi dichiarative ed esclamative). Dal momento che nei testi selezionati non era presente la stessa varietà dal punto di vista quantitativo e qualitativo per le frasi interrogative, per quest'ultime si è deciso di definire un corpus "ad hoc", da far leggere allo stesso speaker dei racconti. Una prima fase di analisi sulle interrogative ha condotto alla definizione di un corpus di 180 interrogative parziali comunemente note in inglese come "**wh-** questions", equamente distribuite in 6 classi ("Chi", "Cosa/Che", "Quale/i", "Perché", "Quando", "Dove"). Ipotizzando che la lunghezza della frase possa in qualche modo influire sugli aspetti prosodici della frase stessa, ogni classe è stata differenziata in base alla lunghezza delle frasi (corte, medie, lunghe). Sono state registrate quindi 10 interrogative per ogni lunghezza e per ogni "classe" (vedi Tab. 1).

Domande	**Frase Corta (1-3 parole)**	**Frase Media (4-8 parole)**	**Frase Lunga (> 8 parole)**	**Totale**
Chi	10	10	10	30
Cosa/che/che_cosa	10	10	10	30
Quale/i - che	10	10	10	30
Perché	10	10	10	30
Quando	10	10	10	30
Dove	10	10	10	30
Totale	**60**	**60**	**60**	**180**

Tabella 1: Wh - questions

Questo primo corpus di interrogative è stato poi integrato con un ulteriore insieme di 104 frasi, nelle quali figurano interrogative diverse per tipologia (interrogative totali), per ordine delle parole (topicalizzate) e per struttura informativa (vedi Tab. 2).

Tipo di interrogativa	**Numero**
Topicalizzate	40
Scisse	12
Varie	52
Totale	**104**

Tabella 2: Altre interrogative

Costituzione e dimensione del corpus sono riassunti in Tabella 3:

Corpus	Frasi	Parole	Durata
Il Colombre	141	1909	16’ 45”
I Sette Messaggeri	67	1458	12’ 47”
La Giacca Stregata	163	1968	18’ 18”
Wh – questions	180	1150	12’ 10”
Altre interrogative	104	800	7’ 00”
Totale	**655**	**7285**	**67’ 02”**

Tabella 3: Dimensioni in termini di frasi, parole e durate

3. ANNOTAZIONE TOBI(T)

Durante l’annotazione con ToBIt è stato impiegato Praat (Boersma & Weenink, 2003), un programma di analisi e sintesi del segnale che offre una notevole varietà di strumenti per l'interazione con i dati vocali, inclusi alcuni strumenti per la trascrizione e l'annotazione a più livelli. Praat consente infatti di rappresentare sullo schermo: lo spettrogramma, la curva della frequenza fondamentale (F0), e altri parametri (intensità, formanti, etc.) allineati temporalmente con il parlato trascritto, segmentato in parole e in fonemi, e con altri livelli di annotazione, in cui (in questo caso) si é inserita l’annotazione ToBI.

ToBI è un sistema di trascrizione per l’intonazione ed altri aspetti della prosodia, concepito originariamente per la lingua inglese (Beckman *et alii*, 1993) e applicato, in un secondo momento, anche alla lingua italiana (ToBIt). Tale sistema è nato con l’intento di sopperire alla mancanza di univocità nella trascrizione prosodica all'interno della comunità scientifica. ToBI considera la distinzione delle unità prosodiche e dei contorni intonativi distintivi di una lingua; inoltre consente di catturare i fenomeni prosodici più importanti del parlato spontaneo. Non considera invece la registrazione dei fenomeni misurabili in termini quantitativi (per esempio la velocità di elocuzione) e di quegli aspetti della prosodia che, pur essendo di natura categoriale, sono predicibili in base ad altre parti della trascrizione o in base a strumenti ausiliari, come i dizionari (ad esempio, l’accento lessicale). Il sistema prevede alcuni elementi base per la trascrizione del parlato:

- Registrazione del testo
- Forma d’onda del segnale relativo (riferimento per collocare nel tempo gli eventi prosodici associati all’enunciato)
- Contorno di F0, ovvero la rappresentazione fonetica primaria della struttura intonativa.

ToBI è basato su una serie di simboli che designano specifici eventi prosodici organizzati in 4 livelli:

- **ortografico**: indica i confini di parola
- **miscellaneo**: indica fenomeni paralinguistici (tosse, risate, pause, discorso diretto…)
- **tonale**: descrive il contorno intonativo come una sequenza di toni alti (H) e bassi (L). In base alla loro funzione nel sistema intonativo, i toni sono marcati da differenti diacritici (*, -, %)

- **delle giunture**: indica il raggruppamento prosodico delle parole nell'enunciato, annotando il confine tra parola e parola (Break Indices).

Nella realizzazione orale di una frase la sequenza degli elementi linguistici viene scandita in unità prosodiche non necessariamente coincidenti con le unità sintattiche. Tale scansione prende il nome di **phrasing**. L'enunciato viene diviso in "sintagmi intermedi" (minori) e "intonativi" (maggiori, che includono uno o più sintagmi intermedi), identificati percettivamente e acusticamente da tratti ritmici (allungamenti segmentali, pause) e melodici (presenza di particolari movimenti intonativi realizzati sulle sillabe immediatamente precedenti il confine). Il grado di disgiuntura viene segnalato in ToBI con un indice numerico (Break Index) assegnato al confine di tutte le parole, sia che queste si trovino all'interno che al confine dei sintagmi intermedi e intonativi, contribuendo così a fornire una rappresentazione completa della forza che lega una parola all'altra entro la frase. Gli indici numerici, che vanno da 0 a 4 (a numero minore corrisponde forza di coesione maggiore) si utilizzano per segnalare i seguenti fenomeni:

0 - pronomi, clitici, elisioni;
1 - grado normale di disgiuntura tra una parola e l'altra entro il sintagma intermedio;
2 - esitazioni, interruzione ritmica ma non melodica o melodica ma non ritmica.
3 - allungamento dei segmenti finali, movimenti intonativi sulla sillaba finale, pausa non molto lunga e senso di continuazione;
4 - pausa lunga, movimenti intonativi sulla sillaba finale, senso di compiutezza.

Il livello tonale è la parte di trascrizione che corrisponde più da vicino all'analisi fonologica. I toni sono espressi con simboli (H, L) marcati con dei diacritici che ne indicano le funzioni intonative. Il tono alto H è realizzato come un picco o un massimo locale di frequenza fondamentale, il tono basso L è realizzato come un minimo locale di frequenza fondamentale. La loro differenza è paradigmatica: in una stessa posizione strutturale H sarà sempre piú alto di L. Vengono marcati due tipi di toni, quelli associati a sillabe accentate (**pitch accents**) e quelli associati a confini intonativi (**phrasal tones**: **phrase accents**, **boundary tones**), che delimitano costituenti prosodici di due livelli: intermediate phrase (sintagma intermedio) e intonational phrase (sintagma intonativo). Piú in dettaglio:

- **Pitch Accent**: accento intonativo. Il diacritico usato è "*" e indica l'associazione del tono alla sillaba accentata.
- **Phrase Accent**: letteralmente accento (tono) di sintagma, identifica un sintagma intermedio (composto da almeno un accento intonativo e uno di sintagma). Il diacritico usato è "-" e ha funzione delimitativa (L-, H-)
- **Boundary Tone**: tono di confine, identifica un sintagma intonativo (composto da uno o più sintagmi intermedi). Il diacritico usato è "%" e ha funzione delimitativa (L%, H%)

I toni di confine che si possono annotare con ToBI sono quindi i seguenti:

- **L-L%** discendente semplice (senso di conclusione)
- **H-H%** alto ascendente
- **L-H%** basso ascendente (senso di continuazione)
- **H-L%** medio
- **%H** inizio di contorno molto alto: in genere nelle esclamative

Per quanto concerne l'accento, ToBI ne marca di due tipi:

a) **Accenti monotonali:**
- **H*** si manifesta come un picco raggiunto dalla curva F0.
- **L*** target tonale realizzato nella parte più bassa dell'estensione melodica del parlatore, spesso senza variazioni considerevoli di F0.

b) **Accenti bitonali:**

Negli accenti bitonali un tono è associato alla sillaba tonica – e l'associazione esplicitata dalla notazione con l'asterisco - mentre l'altro tono che può precedere o seguire, non è associato ad alcuna sillaba, ma realizzato "con riferimento al tono asteriscato". Questo ha come conseguenza che la realizzazione fonetica di un tono asteriscato vedrà il target alto o basso del tono centrato stabilmente entro la sillaba tonica, mentre il target del tono non asteriscato sarà più variabile e potrà allinearsi a segmenti diversi della sillaba pretonica o postonica in dipendenza del contesto segmentale o della velocità di elocuzione.

- **H+L*** movimento discendente su sillabe contigue
- **L+H*** movimento ascendente su sillabe contigue
- **(H+L)*** movimento discendente nella stessa sillaba
- **(L+H)*** movimento ascendente nella stessa sillaba
- **H* !H** down-step: sequenza scalare di toni alti con andamento "a gradino".

La frequenza fondamentale viene marcata con la notazione "HiF0" nel punto più alto di ogni sintagma intermedio. I casi dubbi vengono invece codificati con la dicitura **X*?** quando si percepisce una prominenza, ma non si riconosce il tipo di accento e ***?** quando non si è certi dell'esistenza di una prominenza.

A titolo illustrativo riportiamo due esempi di annotazione ToBI(t). Il primo esempio preso in considerazione (Fig. 1) è un frammento di frase dichiarativa ("*No, no, niente. Dissi. Un lieve capogiro.*") tratta da *"La giacca stregata"* di Dino Buzzati.

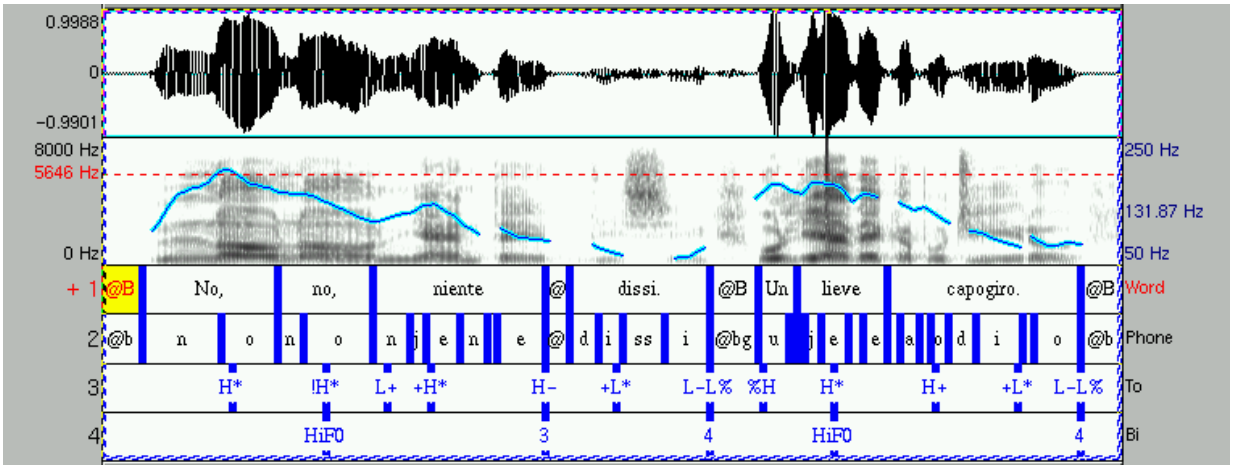

Figura 1: Frammento di dichiarativa tratto da *"La giacca stregata"*

Praat permette nella fascia alta di visualizzare il segnale vocale, il pitch e lo spettrogramma, mentre sui quattro livelli inferiori si trovano la trascrizione del parlato in parole e fonemi, l'annotazione a livello tonale e a livello di disgiunture. Analizzando la frase sino al primo tono di confine (*"No, no, niente dissi"*) vediamo che il parlatore inizia con un down-step **H* !H*** sulle parole "No, no" che trova un corrispettivo nella diminuzione progressiva del pitch. Realizza poi un movimento ascende sulla parola "niente" con un accento bitonale (**L+H***). Invece, l'accento di sintagma **H-**, posto prima della prima pausa (dopo *"dissi"),* è rappresentativo di un cambio di pitch dopo la stessa. La frase, per quanto affermativa, è stata letta con una certa enfasi e questo giustifica la presenza in più punti di una partenza alta da parte del parlatore, il quale riporta gradatamente il range ad un livello più basso. All'interno dell'intervallo, oltre ad accenti monotonali, è presente anche un accento bitonale **H+L*** ("capogiro"), tipicamente usato in posizione finale delle frasi dichiarative che veicolano informazione "nuova", atto a rappresentare un andamento discendente di F0 che parte dalla sillaba pretonica e raggiunge il target L sulla sillaba tonica. Le diverse **HiF0** indicano il punto più alto raggiunto dal pitch nel sintagma intermedio e i toni di confine vengono espressi da **L-L%**, indicando un movimento discendente semplice.
Il secondo esempio (Fig. 2) è un frammento tratto dalla sezione delle **Wh-**questions, composto da due frasi interrogative ("*Quando la sua innocenza è diventata evidente?*" – "*Quando ti deciderai a comprare una macchina?*").

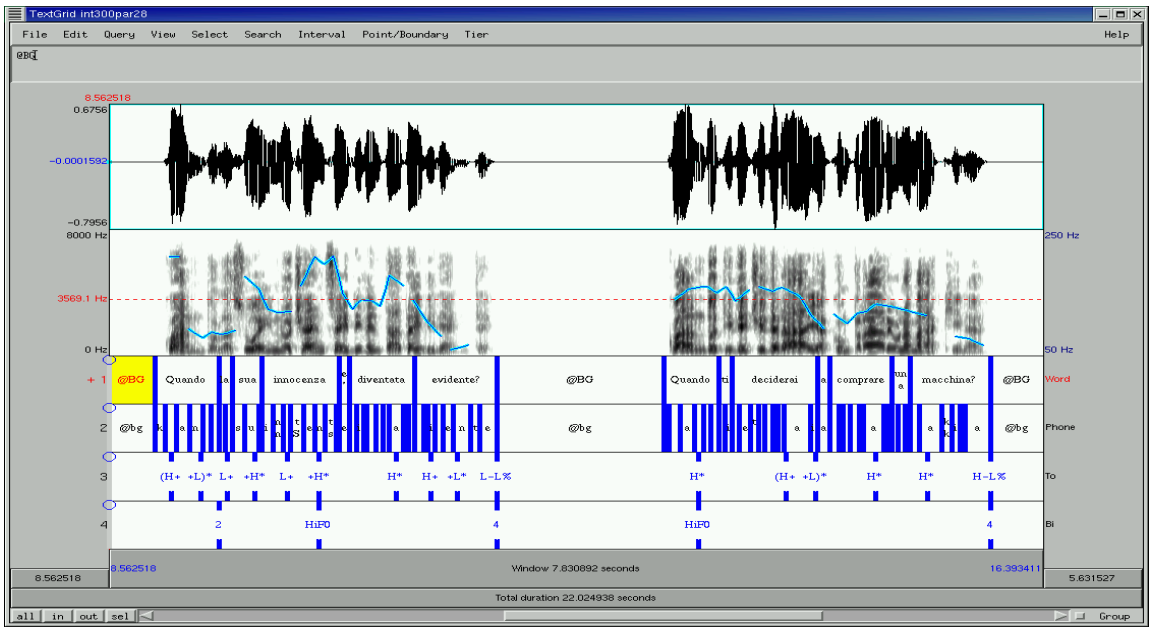

Figura 2: Frammento di frase interrogativa

Nella prima frase possiamo notare che la parola *"Quando"* è associata ad un accento intonativo bitonale i cui toni sono entrambi allineanti alla sillaba tonica **(H+L)***[3]; questa notazione sta a

[3] Diversamente dalla notazione ToBI standard, nel nostro database i due toni di un accento intonativo bitonale sono separati da due segni +. Questo artificio è stato dettato da ragioni empiriche: poiché, come abbiamo detto,

rappresentare un andamento melodico discendente in cui sia il target del tono H che il target del tono L sono raggiunti entro i confini della sillaba tonica. **(frase liminata)**. Il break index "**2"** cattura l'esitazione, o la breve pausa, durante l'elocuzione (senza che si verifichi un cambiamento melodico però nella stessa) da parte del parlatore. Infine, l'andamento andamento discendente al confine di frase è rappresentato da **L-L%.**
Nella seconda frase interrogativa la parola *"Quando"* presenta un accento monotonale che foneticamente rappresenta il valore massimo raggiunto entro il sintagma **(HiF0),** mentre l'andamento di chiusura è medio ed è marcato da **H-L%.**

4. METODOLOGIA

Il testo dei racconti è stato scaricato in formato .pdf e convertito in testo ASCII. Dopo alcune modifiche (allineamento tra testo e audio, suddivisione del testo in paragrafi, normalizzazione di alcuni caratteri), il testo è stato trasformato automaticamente in una famiglia di grammatiche lineari che sono state utilizzate per segmentare il segnale audio corrispondente (tutta la novella), prima in paragrafi e poi in parole e fonemi. Queste segmentazioni sono state effettuate in maniera automatica dai tool di riconoscimento e seguite da un controllo manuale per verificare la correttezza (ed eliminare eventuali errori) dell'allineamento. I paragrafi sono stati segmentati in tratti di audio che avessero una durata compresa all'incirca tra i 20" e i 30" ciascuno, che conservassero un'unità di senso e avessero pause ai bordi.

Questa suddivisione è sembrata la più adeguata agli aspetti più "pratici" dell'annotazione ToBIt, che considera l'andamento tonale all'interno di un sintagma intonativo. Segmentando l'audio nei punti in cui presentava pause evidenti e molto lunghe non si è rischiato di inficiare la successiva annotazione. Inoltre è risultato più agevole lavorare su tratti di audio relativamente brevi.

Associando le varie informazioni ottenute si sono ricavati dei file TextGrid con 4 Tier, due dei quali contengono i dati per la segmentazione in parole e la segmentazione in fonemi (entrambi con i tempi di inizio e fine del singolo elemento); gli altri due Tier erano vuoti e prevedevano l'inserimento dell'annotazione ToBI.

I file così ottenuti sono stati divisi tra tre annotatrici[4] che hanno proceduto nell'ordine all'annotazione dei tre racconti e poi delle frasi interrogative.
L'uso della convenzione ToBIt presenta una soggettività tale da poter causare una disomogeneità tra le annotazioni di persone diverse. Per limitare il più possibile questa disomogeneità e per circoscrivere i fenomeni più critici in termini di interpretazione e quelli di disaccordo, sistematicamente una parte del database è stata annotata da tutte e tre le annotatrici, e le diverse

---

l'allineamento del tono non asteriscato può dipendere dal contesto, ci aspettiamo che la distanza temporale tra il punto di allineamento del tono non asteriscato e quello asteriscato sia un parametro potenzialmente variabile. Non volendo trascurare nella nostra codificazione l'informazione relativa al punto di allineamento dei toni ai segmenti, abbiamo deciso di marcare il tono nel punto esatto in cui raggiunge il suo target e, dunque, di non seguire le indicazioni di Beckman et al. (1997) di marcare l'accento bitonale in un punto centrale della sillaba tonica. Con questo criterio, il tono non associato che precede è marcato T+ e il tono associato che segue è marcato +T*.

[4] A scopo formativo le persone coinvolte nel lavoro di definizione e annotazione del database avevano seguito inizialmente un corso sull'annotazione ToBI, tenuto dalla Dottoressa Cinzia Avesani presso l' ISTC-SFD di Padova.

versioni sono state confrontate tra loro, per una verifica delle divergenze e degli eventuali dubbi[5]. Su tutte le annotazioni sono stati poi svolti dei controlli incrociati dalle annotatrici stesse, in modo che ciascuna facesse una revisione ed eventualmente proponesse delle modifiche alla versione dell'altra: in questo modo ogni file è stato visto da almeno due annotatrici. Infine, tutti i TextGrid ottenuti sono stati sottoposti ad un controllo sintattico tramite uno script e sono stati corretti errori di vario tipo, quali l'utilizzo scorretto o inesatto della simbologia ToBI, dimenticanze o semplici errori di battitura.

Il database è stato annotato anche a livello linguistico (ortografico, POS e sintattico a costituenti) dal Dipartimento di Linguistica dell'Università di Venezia.

5. INTERCODER AGREEMENT

Dopo una revisione della versione definitiva dei file annotati prosodicamente, si è proceduto al calcolo del livello di accordo tra le tre annotatrici.
Lo scopo era quello di valutare il possibile impiego per un'annotazione su larga scala dei fenomeni prosodici riguardanti database di parlato con un buon margine di uniformità. Per valutare statisticamente il livello di accordo tra le annotazioni prosodiche si è preso in esame un sottocorpus, estratto dal database di partenza, composto da 5' 36" di parlato, corrispondenti a 626 parole, annotati separatamente da tutte e tre le annotatrici, senza condizionamenti reciproci.

Il livello di accordo è dato dal confronto delle diverse etichette di annotazione collocate dalle singole annotatrici su una determinata parola (o su un confine di parola). Per il calcolo di tale livello si fa uso di un indice che è dato dal prodotto tra il numero complessivo di parole e il numero di coppie possibili tra le annotatrici (Silverman et al., 1992; Pitrelli et al., 1994). Nel caso specifico, tre annotatrici (a, b, c) danno origine a tre possibili coppie (ab, ac, bc). Siccome il numero totale di parole è 626, ne consegue che l'indice è 1878 (626*3).
Il livello di accordo è dato quindi dal seguente rapporto:

```
agreement = 100 * exact_match / 1878
```

Quando si fa riferimento al livello di accordo tra annotazioni ci si riferisce in realtà a due indici, il livello di accordo a livello tonale (a sua volta suddiviso in due livelli) e il livello di accordo a livello di break. Per calcolare questi due indici è stata utilizzata la formula riportata qui sopra, variando opportunamente il valore di exact_match nei due casi.

5.1 Accordo a livello tonale

Si è calcolato separamente il livello di accordo su ciascun aspetto del livello tonale, tenendo presente che esso è caratterizzato da due tipi differenti di etichette prosodiche. Il primo tipo indica la presenza e il tipo del pitch accent, mentre il secondo tipo stabilisce i toni di confine.

5.1.1 Accordo a livello di pitch accent

Per il calcolo di questo livello di accordo sono stati utilizzati un criterio rigido e uno meno rigido. Il criterio rigido si basa sulla perfetta corrispondenza tra le etichette e le "non-etichette"[6] collocate

[5] Il confronto è stato realizzato con una frequenza maggiore nella prima fase del lavoro rispetto alle successive perché ovviamente nella prima fase del lavoro il grado di disomogeneità era più alto.

[6] L'accordo sussiste anche quando tutti gli annotatori ritengono che non ci sia nessun fenomeno prosodico da marcare.

dalle annotatrici, mentre quello meno rigido ammette delle lievi sfumature tra le etichette prosodiche (ad es. H* e (L+H)* sono stati considerati identici perché in entrambi il target sulla sillaba tonica è H*). La percentuale di accordo a livello di pitch accent, applicando il criterio rigido, è stata calcolata sulla base delle seguenti 7 etichette: **H***, **L***, **L+ +H*** (questa etichetta ingloba (L+ +H)* e (L+H)*), **H+ +L*** (questa etichetta ingloba (H+ +L)* e (+H+L)*), **X*?**, **!H***, **NO** (che indica l'assenza di pitch accent su una parola). Un'ulteriore statistica è stata fatta valutando semplicemente la presenza o meno di accenti e toni di confini, indipendentemente dal loro valore. Ad esempio:

Testo:	**Il**	**colombre**	**di**	**Dino**	**Buzzati**
Ann. 1:	NO	H*	NO	L*	L+ +H*
Ann. 2:	NO	L*	NO	L*	H*
Ann. 3:	NO	H*	NO	H*	H*

Per le parole "il" e "di" l'accordo è del 100% (infatti tutte le annotatrici hanno evidenziato l'assenza di pitch accent – etichetta NO) mentre per "colombre", "Dino" e "Buzzati" l'accordo è del 33.33% (solo una coppia è concorde). In totale, su 15 coppie annotatrici-parole, ci sono 9 concordanze; l'agreement complessivo è quindi del 60,0% (100*9/15).
Secondo il criterio meno rigido, più realizzazioni prosodiche sono state raggruppate in un'unica etichetta. In questo caso il livello di accordo è stato calcolato in base ad un numero inferiore di etichette[7], 4. Più precisamente: **HH***, raggruppante H*, (L+ +H)*, (L+H)* e !H*; **LL***, raggruppante L*, (H+ +L)* e (+H+L)*; **X*?** e **NO**. Applicando il criterio meno rigido all'esempio visto precedentemente otteniamo le seguenti percentuali di accordo: per le parole "il" e "di" la percentuale di accordo rimane del 100%; per "colombre" e "Dino" l'accordo è del 33.33%, mentre per "Buzzati" abbiamo un accordo del 100% (sia L+ +H* che H* finiscono nella classe comune HH*). L'agreement complessivo sale quindi al 73,33% (11 concordanze su 15 coppie).

5.1.2 Accordo a livello di confini prosodici
Come per l'accordo a livello di accenti intonativi, anche per l'annotazione del tono di confine si è deciso di applicare un criterio rigido e uno meno rigido. Il criterio rigido considera il livello di accordo sulla corrispondenza tra le etichette collocate dalle annotatrici. Quello meno rigido considera invece la presenza o meno di etichette di toni di confine.
Come per il criterio rigido a livello di pitch, la percentuale di accordo a livello di tono di confine, è stata calcolata tenendo conto delle classi corrispondenti alle 8 etichette riscontrate nel database. Piu` precisamente: **%H**, **H-**, **L-**, **L-L%**, **L-H%**, **H-L%**, **H-H%**, **NB** ("no boundary" ad indicare l'assenza di un tono di confine).
Esempio:

Testo:	**Il**	**colombre**	**di**	**Dino**	**Buzzati**
Ann. 1:	%H	L-	NB	NB	L-L%
Ann. 2:	NB	L-	NB	NB	L-H%
Ann. 3:	%H	H-	NB	NB	H-

Per le parole "il" e "colombre" l'accordo è del 33.33% (solo una coppia è concorde); per le parole "di" e "Dino" l'accordo è del 100% (tutte e tre le annotatrici concordano nel ritenere che su queste due parole non vi sono toni di confine), mentre per la parola "Buzzati" l'accordo è dello 0% in

[7] Tali etichette non sono proprie del formalismo ToBI ma sono state definite arbitrariamente.

quanto le annotatrici hanno collocato etichette differenti. Complessivamente si ha quindi un agreement pari al 53,33% (8 concordanze su 15 coppie).

Applicando il criterio meno rigido all'esempio visto prima otteniamo le seguenti percentuali di accordo: per la parola "il" l'accordo è del 33.33% (solo una coppia è concorde); per le parole "colombre" e "Buzzati" l'accordo è del 100% in quanto tutte le annotatrici sono concordi sulla presenza di un'etichetta prosodica; per le parole "di" e "Dino" l'accordo è del 100% in quanto tutte le annotatrici sono concordi sul fatto che non vada collocata alcuna etichetta. Complessivamente, 13 concordanze su 15 coppie portano l'agreement al 86,66%.

Le seguenti tabelle riassumono le percentuali relative all'accordo a livello tonale applicando i criteri precedentemente descritti:

	coppie	**accento presente**	**rigido (7 classi)**	**tollerante (4 classi)**
Pitch accent	1878	88.1%	73.0%	81,3%

Tabella 4: Accordo nell'assegnazione dei pitch accents

	coppie	**confine presente**	**rigido (8 classi)**
Confini Prosodici	1878	95.3%	89.4%

Tabella 5: Accordo nell'assegnazione dei confini prosodici

5.2 Accordo a livello di giuntura

Si è ritenuto opportuno calcolare l'accordo a livello di Break Indices tenendo presente che esso è caratterizzato da cinque tipi di etichette prosodiche, che servono ad individuare il livello di disgiuntura tra le parole. Come per l'accordo a livello tonale, sono stati applicati un criterio rigido e uno meno rigido. Il criterio rigido considera il livello di accordo sulla perfetta corrispondenza tra le etichette (0, 1, 2, 3, 4) collocate dalle annotatrici, mentre il criterio meno rigido ammette uno scarto di un punto, ovvero si ritiene comunque valido un intorno di ±1. Ad esempio nel caso di break a livello 2, l'accordo sussiste non solo tra etichette "2" ma anche con etichette "1" e "3".

Di seguito viene riportata la tabella in cui vengono illustrate le percentuali relative all'accordo a livello di break applicando i criteri precedentemente descritti.

	coppie	**rigido (5 classi)**	**tollerante (±1)**
Break index	1878	92.5%	99.2%

Tabella 6: Accordo nell'assegnazione dei break indices

Le figure seguenti (Fig. 3–4–5) riportano una statistica dei fenomeni prosodici presenti nell'intero database (7285 parole per 655 frasi).

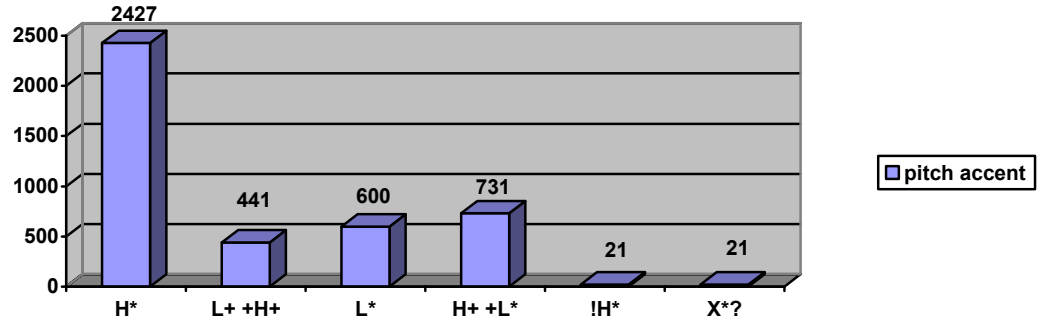

Figura 3: Distribuzione dei pitch accents nell'intero database

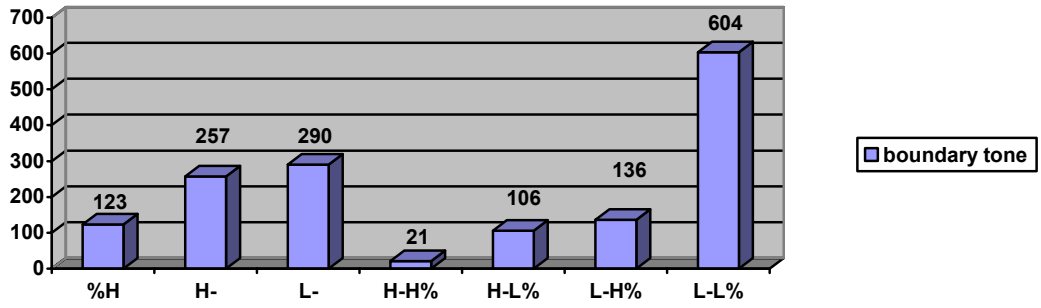

Figura 4: Distribuzione dei boundary tones nell'intero database

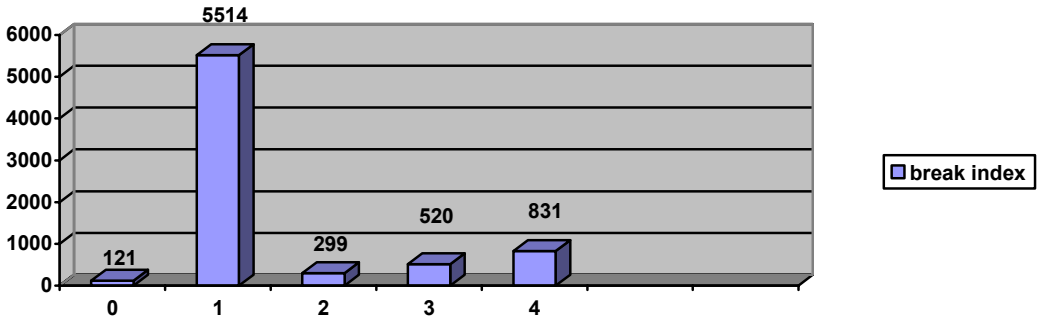

Figura 5: Distribuzione dei break indices nell'intero database

6. PROBLEMI NELL'ALLINEAMENTO

Per l'annotazione prosodica, il segnale audio è stato diviso in paragrafi, contenenti 7285 parole più i silenzi. Un paragrafo è stato definito come un elemento di lunghezza ragionevolmente prossima a 25 secondi, che ai bordi presentasse interruzioni prosodiche certe (normalmente pause lunghe).

Per l'annotazione linguistica, invece, i testi sono stati divisi in frasi (terminate da punto o equivalenti) e gli elementi presi in considerazione sono costituenti (7594) e foglie (8602). Queste ultime comprendono parole, punteggiatura, locuzioni e "multi-words" (di_quando_in_quando, su_e_giù, dino_buzzati), clitici separati dal verbo per permettere un'analisi sintattica (andiamoci *CI).

Non esiste nessuna relazione fissa tra paragrafi e frasi, se non che entrambi coprono l'intero racconto (Fig.6). Si è quindi posto il problema di effettuare un allineamento tra i due diversi tipi di annotazione. La procedura di allineamento ha operato su tutto il racconto (lista di paragrafi – lista di frasi) in questo modo:

- avanza di un elemento alla volta
  - salta silenzi nella catena prosodica
  - salta costituenti, clitici e punteggiatura nella catena linguistica
- se i due elementi attuali sono uguali
  - allinea ed avanza in entrambe le catene
  - altrimenti (caso multiword)
    - avanza nella catena prosodica, concatenando elementi fino ad eguagliare la multiword

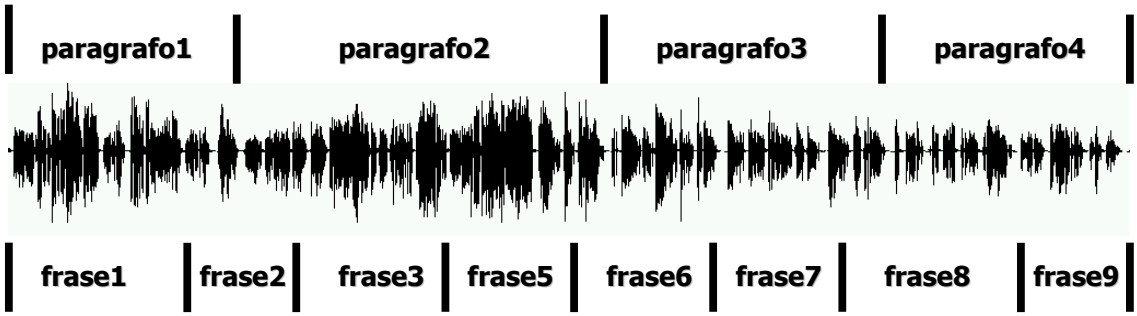

Figura 6: divisione di un racconto in paragrafi (blocchi prosodici) e frasi (elementi linguistici)

7. CONCLUSIONI

Il lavoro realizzato ha condotto alla costruzione di un database di parlato letto, annotato a livello sia linguistico che prosodico. Si ritiene che questo database sia significativo non solo in termini quantitativi (67' 02" di audio annotato a più livelli) ma anche in termini qualitativi, come comprovato dall'esito positivo della valutazione dell'accordo tra annotatrici.
In futuro contiamo di iniziare ad utilizzare il database per studiare le possibili relazioni tra strutture morfo-sintattiche e prosodiche. Inoltre, i modelli prosodici ricavati dai dati del database verranno applicati al sistema di sintesi per l'italiano, Festival (Cosi *et alii,* 2001). Prevediamo anche possibili ampliamenti del database (ad esempio del corpus di frasi interrogative con una maggiore copertura dei vari fenomeni linguistici). Infine, si sta ipotizzando la creazione di analoghi database "emotivi", ovvero di parlato con diverse emozioni (es. tono arrabbiato, felice, irritato, ecc.).

BIBLIOGRAFIA

Androutsopoulos I., Calder J., Not E., Pianesi F. , Roussou M.: "Multilingual Personalised Information Objects", *IPNMD*, Verona, Italy, 14-15 December 2001

Avesani C., ToBIt: un sistema di trascrizione per l'intonazione italiana, , in *Atti delle V giornate di Studio del Gruppo di Fonetica Sperimentale*, Povo, Trento, Novembre 1994, pp. 85-98

Avesani C. (1999). "Quantificatori, negazione e costituenza sintattica. Costruzioni potenzialmente ambigue e il ruolo della prosodia". In AAVV, *Fonologia e morfologia dell'Italiano e dei dialetti d'Italia*. *Atti del* XXXI *della Società di Linguistica Italiana,* Padova, ottobre 1997, pp. 153-200.

Beckman M., Gayle, Guidelines for ToBI Labelling (version 3, March 1997) copyright 1993, The Ohio State University Research Foundation X ; ToBI: http://ling.ohio-state.edu/~tobi/

Boersma P., Weenink D., Praat, (Version 4.0.49 - 2003) Institute of Phonetic Sciences, University of Amsterdam http://www.fon.hum.uva.nl/praat/

Cosi P., Tesser F., Gretter R., Avesani C., Macon M. : "Festival Speaks Italian!", *EUROSPEECH 2001*, Aalborg, Denmark, September 3-7, 2001, pp. 509-512.

Delmonte R, Pianta E. (1996), IMMORTALE –"Analizzatore Morfologico, Tagger e Lemmatizzatore per l'Italiano", *in Atti V Convegno AI*IA "Cibernetica e Machine Learning",* Napoli 1996, pp 19-22.

Silverman et al., "ToBI: A Standard for Labeling English Prosody", *Proceedings of the 1992International Conference on Spoken Language Processing (ICSLP*), Banff, Alberta, Canada, October 13-16, 1992, v. 2, pp. 867-870.

Pitrelli J.F., Beckman M., Hirschberg J., "Evaluation of Prosodic Transcription Labeling Reliability in the ToBI Framework". *Conference on Spoken Language Processing (ICSLP),* Yokohama, Japan, 1994, v. 1, pp. 123-126.