NUOVE FUNZIONALITÀ PER IL PORTALE CLIPS[1]

Aurelio De Rosa, Francesco Cutugno
Dipartimento di Scienze Fisiche Università di Napoli Federico II - LUSI Group
aurelioderosa@gmail.com , cutugno@na.infn.it

1. SOMMARIO

Il lavoro realizzato è il rifacimento del portale web CLIPS corredato da alcuni software per lo studio e l'analisi statistica di *corpora* linguistici. Come ampiamente noto CLIPS è un progetto per la raccolta di parlato italiano semi-spontaneo vastamente utilizzato negli ultimi anni da studiosi del settore della fonetica sperimentale in particolare e della linguistica in generale; per un approfondimento sul corpus si veda Savy & Cutugno (2009).

Il primo programma è un convertitore che opera su più livelli di annotazioni ed etichettatura (fonetico, sillabico, ed altri) codificati in formato TIMIT e dà origine ad un Annotation Graph (AG) in formato XML (Bird & Liberman, 2001).

La seconda funzionalità consente di interrogare gli AG. In esso è prevista la possibilità di compiere la stessa *query* contemporaneamente su più *corpora* e visualizzare i risultati insieme. L'elaborazione è effettuata secondo diversi criteri scelti dall'utente e su differenti livelli di etichettatura, ad esempio, sul piano fonetico o acustico o cercando delle stringhe all'interno dei vari livelli concatenando tra loro diversi livelli di selezione.

Un altro degli applicativi presenti genera statistiche generali inerenti i *corpora*. Dopo aver scelto il *corpus* od i *corpora* dialogici sui quali agire si creano statistiche, per ogni dialogo presente, circa il nome identificativo, il numero di parole pronunciate ed il tempo per il quale parla ogni interlocutore.

Un'altra funzionalità permette di analizzare il livello .WRD (contenente la trascrizione ortografica dei file audio) di un *corpus* prescelto e di produrre un file XML contenente le Part-of-Speech, il lemma ed altre informazioni inerenti le parole di quel dato *corpus*.

L'ultima funzionalità creata partendo dalla selezione di un *corpus*, origina statistiche sul numero di occorrenze dei lemmi presenti in esso e, per ognuno di essi, quante sono le occorrenze delle forme appartenenti a tale lemma.

2. INTRODUZIONE

Il lavoro realizzato si è svolto all'interno del gruppo LUSI (Language Understanding and Speech Inter/Action/Faces) ed è il rifacimento del portale web CLIPS (Savy & Cutugno, 2009) corredato da alcune nuove funzionalità per lo studio e l'analisi statistica di *corpora* linguistici. Il sito è stato sviluppato in due lingue, italiano ed inglese, così da soddisfare le esigenze di eventuali visitatori stranieri. L'intero portale è stato sviluppato usando le JSP di Java e seguendo gli standard W3C. Essendo validato W3C, il sito è definito "accessibile", ovvero è usufruibile da qualsiasi tipo di utente, quindi anche da persone con ridotta o impedita capacità sensoriale, motoria, o psichica. Ha lo scopo di essere punto di riferimento per lo scambio di dati ed informazioni tra studiosi della linguistica grazie ai documenti pubblicati circa l'andamento della ricerca e le novità che gli

[1]Il nuovo portale, disponibile all'indirizzo http://clips.unina.it, è liberamente accessibile a chiunque previa registrazione gratuita.

autori vorranno condividere con l'esterno. Tutto il materiale presente è libero per la consultazione ed i software sono utilizzabili gratuitamente previa registrazione al sito che ha scopi puramente statistici per gli autori. Anche il corpus CLIPS compreso di tutti i suoi file di etichettatura, è navigabile e scaricabile dopo essersi registrati.

Il nostro lavoro si prefigge di estrarre dati statistici che possano servire per lo studio della linguistica. Il nostro approccio, quindi, a differenza di quello strettamente teorico è guidato dai dati. I lavori sono strettamente collegati tra loro e si possono vedere come divisi in due blocchi. Il primo ed il secondo si occupano di creare, partendo dai file che costituiscono il *corpus*, dapprima una struttura facilmente usufruibile da sistemi automatici e poi danno la possibilità di effettuare interrogazioni incrociando più livelli dell'annotazione linguistica.

Per quanto riguarda gli altri tre strumenti proposti, essi possono essere utilizzati per estrarre informazioni adoperate per metodi come gli N-Grams per la previsione della prossima parola date le N precedenti oppure per creare probabilità applicati in modelli Markoviani utilizzati dai riconoscitori automatici di parlato.

3. I SOFTWARE

Come detto nell'introduzione, il portale è fornito di vari strumenti per l'analisi dei *corpora*. Il primo di essi è un convertitore che opera su più livelli di annotazioni ed etichettatura (fonetico, sillabico, ed altri) codificati in formato TIMIT, comprese le informazioni temporali, e dà origine ad un Annotation Graph (AG) in formato XML che rispetta appieno gli standard delineati da Bird e Liberman (Bird & Liberman, 2001). Gli AG permettono di gestire i *corpora* di parlato per i quali i metodi di rappresentazione tradizionali non sono sufficienti a causa del vasto numero di livelli e di rappresentare le annotazioni trascritte da un segnale contenente risorse linguistiche (Cecere, 2008). Di seguito è riportata un'immagine che illustra le modalità di selezione per la costruzione del documento XML (Figura 1) ed un estratto dello stesso (Tabella 1).

```
<AG id="DGtdA03D:p1#110" timeline="DGtdA03D:t6">
<Anchor id="DGtdA03D:p1#110:anchor1" offset="0.0" unit="sec" signals=""/>
<Anchor id="DGtdA03D:p1#110:anchor10" offset="80.8949966430664" unit="sec"
signals=""/>
<Anchor id="DGtdA03D:p1#110:anchor2" offset="0.8550000190734863" unit="sec"
signals=""/>
<Anchor id="DGtdA03D:p1#110:anchor3" offset="8.307999610900879" unit="sec"
signals=""/>
<Anchor id="DGtdA03D:p1#110:anchor4" offset="15.666000366210938" unit="sec"
signals=""/>
<Anchor id="DGtdA03D:p1#110:anchor5" offset="18.989999771118164" unit="sec"
signals=""/>
<Anchor id="DGtdA03D:p1#110:anchor6" offset="22.597999572753906" unit="sec"
signals=""/>
<Anchor id="DGtdA03D:p1#110:anchor7" offset="34.2760009765625" unit="sec"
signals=""/>
<Anchor id="DGtdA03D:p1#110:anchor8" offset="51.2239990234375" unit="sec"
```

```
signals=""/>
<Anchor id="DGtdA03D:p1#110:anchor9" offset="68.26699829101562" unit="sec"
signals=""/>
<Annotation id="DGtdA03D:p1#110:annotation1" type="wrd"
startAnchor="DGtdA03D:p1#110:anchor1"
endAnchor="DGtdA03D:p1#110:anchor2"><Feature name="valore">__</Feature>
</Annotation>
<Annotation id="DGtdA03D:p1#110:annotation2" type="wrd"
startAnchor="DGtdA03D:p1#110:anchor2"
endAnchor="DGtdA03D:p1#110:anchor3"><Feature name="valore">okay</Feature>
</Annotation>
<Annotation id="DGtdA03D:p1#110:annotation3" type="wrd"
startAnchor="DGtdA03D:p1#110:anchor3"
endAnchor="DGtdA03D:p1#110:anchor4"><Feature
name="valore">&lt;inspiration&gt;</Feature>
</Annotation>
<Annotation id="DGtdA03D:p1#110:annotation4" type="wrd"
startAnchor="DGtdA03D:p1#110:anchor4"
endAnchor="DGtdA03D:p1#110:anchor5"><Feature name="valore">e%</Feature>
</Annotation>
<Annotation id="DGtdA03D:p1#110:annotation5" type="wrd"
startAnchor="DGtdA03D:p1#110:anchor5"
endAnchor="DGtdA03D:p1#110:anchor6"><Feature name="valore">%il</Feature>
</Annotation>
<Annotation id="DGtdA03D:p1#110:annotation6" type="wrd"
startAnchor="DGtdA03D:p1#110:anchor6"
endAnchor="DGtdA03D:p1#110:anchor7"><Feature name="valore">sedile%</Feature>
</Annotation>
<Annotation id="DGtdA03D:p1#110:annotation7" type="wrd"
startAnchor="DGtdA03D:p1#110:anchor7"
endAnchor="DGtdA03D:p1#110:anchor8"><Feature name="valore">%invece</Feature>
</Annotation>
<Annotation id="DGtdA03D:p1#110:annotation8" type="wrd"
startAnchor="DGtdA03D:p1#110:anchor8"
endAnchor="DGtdA03D:p1#110:anchor9"><Feature
name="valore">anch’esso</Feature>
</Annotation>
<Annotation id="DGtdA03D:p1#110:annotation9" type="wrd"
startAnchor="DGtdA03D:p1#110:anchor9"
endAnchor="DGtdA03D:p1#110:anchor10"><Feature name="valore">bianco</Feature>
</Annotation>
</AG>
```

Tabella 1: Estratto del file XML generato.

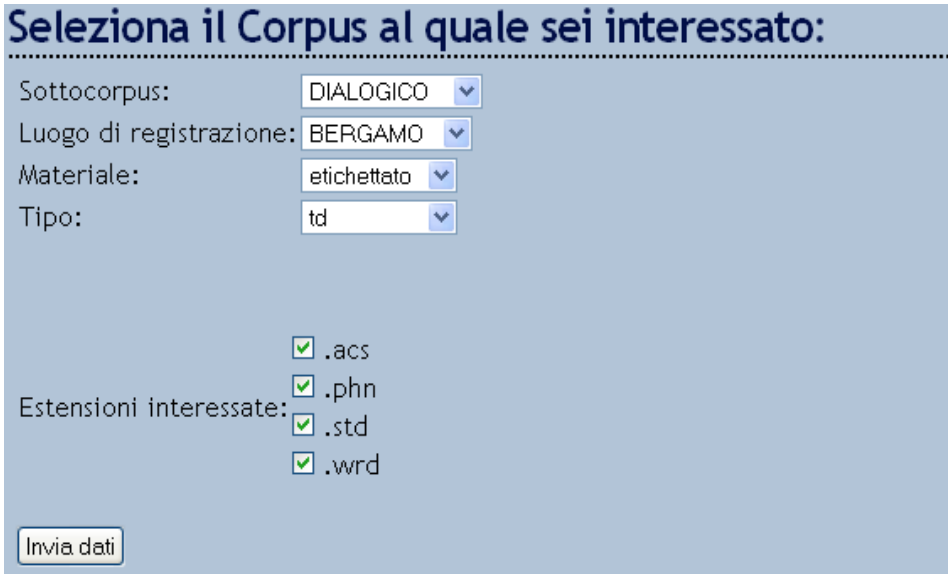


Figura 1: Metodo di selezione di un corpus per creare l'AG in XML.

La seconda funzionalità, attraverso l'implementazione grafica del linguaggio di *query* AGQL per scopi fonetici, si occupa di eseguire interrogazioni sui file XML creati tramite la prima funzionalità. In essa è prevista la possibilità di compiere la stessa interrogazione contemporaneamente su diversi *subcorpora* così da facilitare la comparazione dei risultati. Una volta selezionato il livello di etichettatura di proprio interesse, si può inserire una stringa e decidere se si vuole che il piano prescelto debba contenerla, iniziare o terminare con esso e la *query* si può effettuare concatenando tra loro due livelli di selezione. Il secondo livello ha le stesse opzioni del primo e trai due è possibile stabilire un legame trai seguenti: "coincidenza", "stesso inizio", "stessa fine", "sovrapposizione" e "strettamente incluso" i quali indicano rispettivamente che le due stringhe immesse coincidono a livello temporale, iniziano allo stesso tempo, terminano nello stesso istante, i loro tempi si accavallano, il secondo è totalmente incluso nel primo. Di seguito è mostrato un esempio di interrogazione (Figura 2) con i relativi risultati (Figura 3):

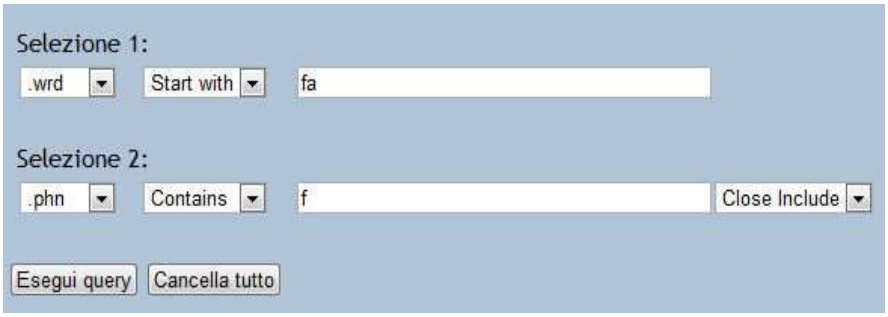


Figura 2: Esempio di interrogazione.

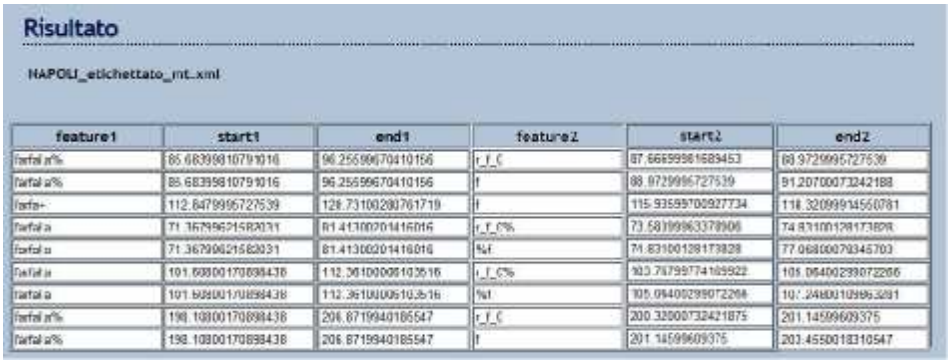


Figura 3: Risultati prodotti dall'interrogazione.

Un altro degli applicativi presenti genera statistiche generali inerenti i *corpora*. Una volta che è stato selezionato il *corpus* od i *corpora* dialogici sui quali agire (sugli altri non è possibile) si devono poter creare statistiche, per ogni dialogo presente, circa il nome identificativo, il numero di parole pronunciate, il tempo per il quale parla ogni interlocutore. La selezione è effettuabile dal livello più basso per un prefissato *corpus* crescendo fino alla scelta di tutti i *corpora* presenti sul server web. Ogni *corpus* analizzato mostra in aggiunta alle statistiche sui singoli parlanti, anche un quadro riassuntivo delle statistiche rilevate che nei conteggi include, nel caso ve ne siano, i dati dei suoi *subcorpora*. Ciò è stato fatto poiché se si esegue il programma sul livello più alto, che quindi racchiude tutti i *corpora* presenti, oltre alle singole statistiche si otterrà un quadro riassuntivo di tutto il materiale presente (De Rosa, 2009). Un esempio di risultato è riportato in Figura 4.

Un'altra funzionalità proposta consente, partendo da un script in PERL che effettua il parsing in parti del discorso (POS – morfosintassi) di un testo dato in ingresso, di analizzare il livello .WRD di un *corpus* prescelto e produrre un file XML contenente le sue POS, acronimo di Part-of-Speech. Lo script usato è il TreeTagger, uno strumento per annotare il testo con informazioni sul POS ed il lemma delle forme. Esso è stato sviluppato da H. Schmid (Schmid, 2009) nel progetto TC presso l'Istituto di Linguistica Computazionale dell'Università di Stoccarda per la lingua tedesca. Questo programma è stato esteso in varie lingue ed in particolare il suo uso per la lingua italiana è dovuta al lavoro di M. Baroni (Schmid et al., 2007). Il dialogo presente nel *corpus* è il nodo radice dell'albero del documento e ha un "id" che lo identifica univocamente. Il valore di questo campo deve essere lo stesso che è usato per individuare i file appartenenti al dialogo in questione. Ogni turno appartenente al dialogo è suo figlio diretto nell'albero del file ed è anch'esso corredato dell'attributo "id" che è un numero sequenziale rappresentante l'ordine di avvicendamento dei turni. Ogni parola presente nel turno è figlia diretta di quello di appartenenza e occorre dei seguenti attributi: "id" che è un numero indicante l'ordine in cui il termine è stato analizzato all'interno del dialogo; "start" identifica il tempo nel quale il dialogante ha iniziato a pronunciare il termine; "end" corrispondente al tempo in cui la pronuncia ha fine; "POS" indica la funzione della parola (in base all'analisi morfosintattica) all'interno della frase.

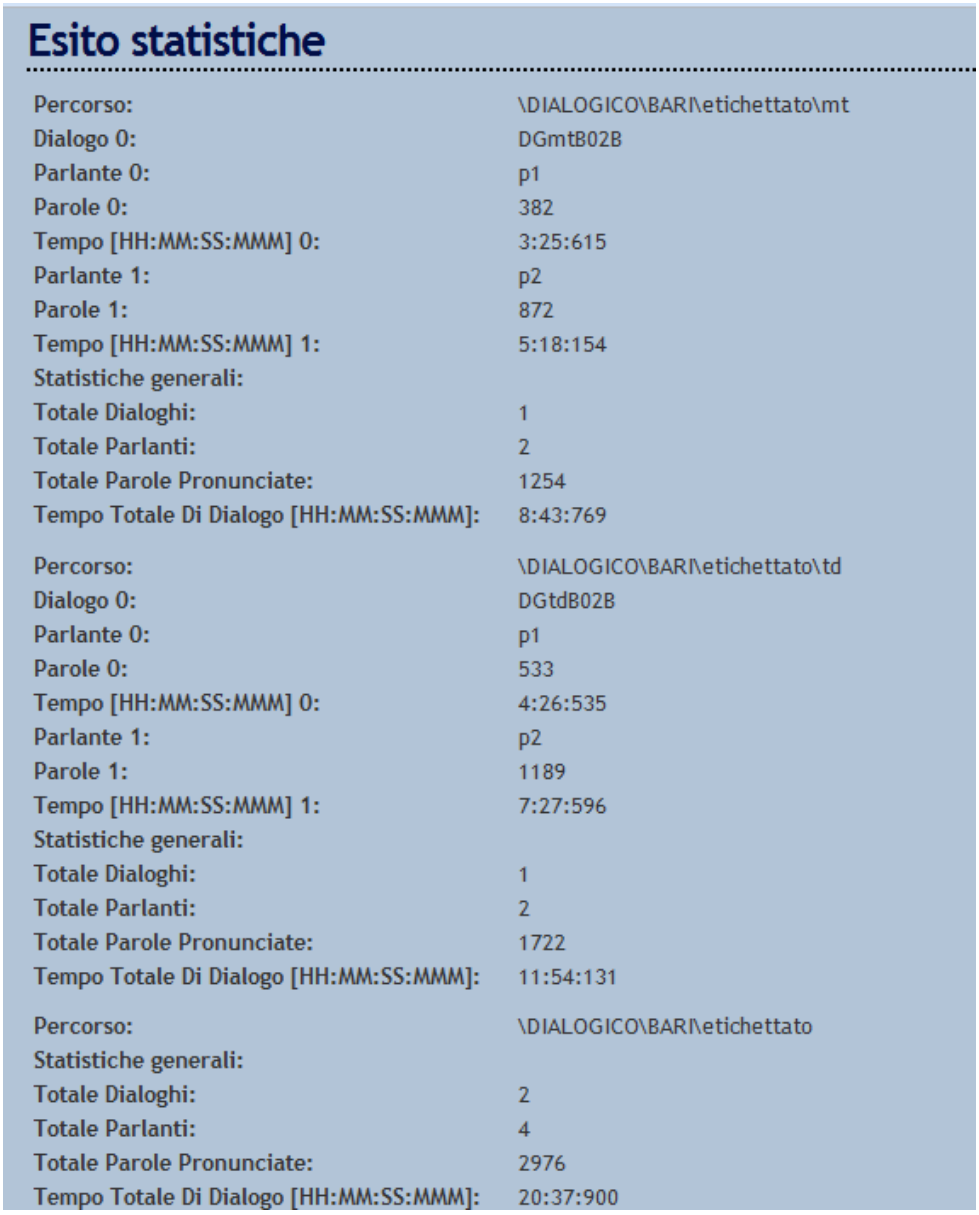


Figura 4: Esempio di risultato di statistiche. Il risultato mostrato è del corpus dialogico di Napoli.

Di seguito è riportato un estratto del file XML che viene generato (Tabella 2) ed un'immagine che mostra la struttura ad albero del documento (Figura 5).

```
    <DIALOGO id="DGmtB02B" parlanti="2" parole="1254" tempo="523769" >
      <TURNO id="1" >
        <WRD id="1" start="0" end="1493" POS="NOM" lemma="__" >__</WRD>
        <WRD id="2" start="1493" end="15978" POS="" lemma=""
><inspiration></WRD>
        <WRD id="3" start="15978" end="19549" POS="" lemma="" ><sp></WRD>
        <WRD id="4" start="19549" end="29593" POS="ADV" lemma="allora"
>allora</WRD>
        <WRD id="5" start="29593" end="39971" POS="" lemma="" ><sp></WRD>
        <WRD id="6" start="39971" end="50914" POS="" lemma="" ><ehm></WRD>
        <WRD id="7" start="50914" end="60860" POS="VER:pres" lemma="girare"
>gira%</WRD>
        <WRD id="8" start="60860" end="63270" POS="PRE" lemma="a" >%a</WRD>
        <WRD id="9" start="63270" end="73057" POS="NOM" lemma="destra"
>destra</WRD>
        <WRD id="10" start="73057" end="75472" POS="PRE:det" lemma="del"
>dei</WRD>
        <WRD id="11" start="75472" end="88206" POS="NOM" lemma="limone"
>limoni</WRD>
      </TURNO>
      <TURNO id="2" >
        <WRD id="12" start="0" end="3655" POS="NOM" lemma="__" >__</WRD>
        <WRD id="13" start="3655" end="6361" POS="PRE" lemma="a" >a</WRD>
        <WRD id="14" start="6361" end="17850" POS="NOM" lemma="destra"
>destra</WRD>
        <WRD id="15" start="17850" end="20651" POS="PRE:det" lemma="del"
>dei</WRD>
        <WRD id="16" start="20651" end="30193" POS="NOM" lemma="limone"
>limoni</WRD>
        <WRD id="17" start="30193" end="38074" POS="CON" lemma="sì"
>sì</WRD>
      </TURNO>

      [Altre righe omesse]

    </DIALOGO>
```

Tabella 2: Esempio di risultato di statistiche. Il risultato mostrato è del corpus dialogico di Napoli.

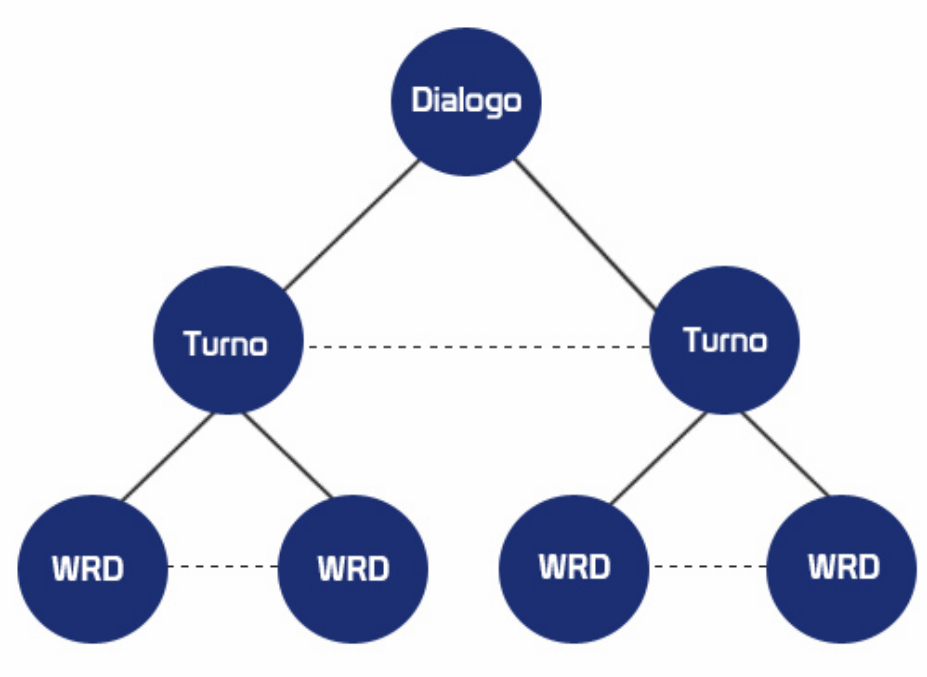
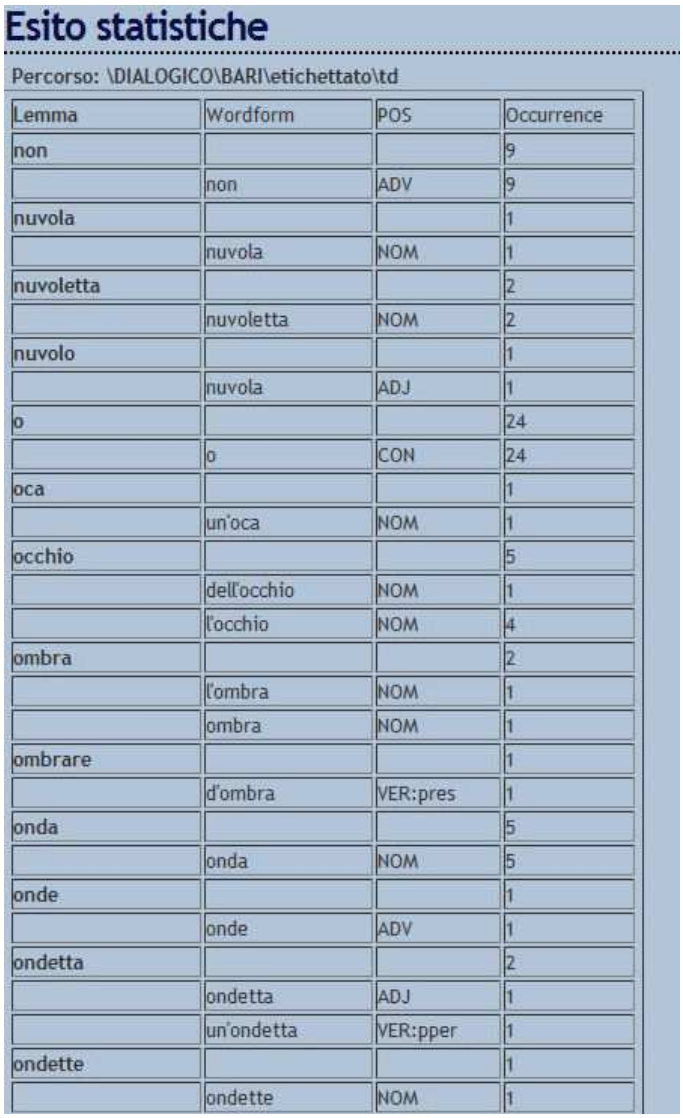


Figura 5: Albero XML dell'analisi morfologica di un corpus.

L'ultima funzionalità creata, partendo dalla selezione di un corpus, può dare origine a statistiche sul numero di occorrenze dei lemmi presenti in esso e, per ognuno di essi, quante sono le occorrenze delle forme appartenenti a tale lemma. L'applicativo cercherà il file XML che estrae i POS per i quali è indicato anche il lemma, così da analizzarlo e compierà la medesima azione per tutti i *subcorpora* presenti (nel caso ve ne fossero). Tutti i lemmi e le forme sono memorizzati in appositi alberi AVL (Adelson-Velsky & Landis, 1962). La configurazione dell'albero generato è degna di nota, infatti, il principale è un AVL contenente nodi che includono il lemma, il numero di occorrenze trovate fino a quel momento ed il riferimento alla radice dell'albero, anch'esso AVL, dei termini che lo hanno come lemma. I nodi di questo sottoalbero racchiudono i seguenti dati: il termine analizzato, il POS ed il numero di volte che ricorre. Dopo aver prelevato i dati dal documento XML, il primo passo è verificare se il lemma estratto esiste già nell'albero. In caso affermativo, viene solo aggiornato il contatore delle frequenze, altrimenti si inserisce un nuovo nodo. Anche se il lemma fosse già presente, non è detto che lo sia il termine individuato, quindi si effettuerà un'altra ricerca e nel caso non sia presente, verrà inserito. Come per gli ultimi due programmi esposti, anche questo è stato sviluppato di modo che esegua delle analisi ricorsive dei sottolivelli di ogni corpus. Di seguito è riportata un'immagine che mostra il tipo di statistiche generate (Figura 6):



Esito statistiche

Percorso: \DIALOGICO\BARI\etichettato\td

Lemma	Wordform	POS	Occurrence
non			9
	non	ADV	9
nuvola			1
	nuvola	NOM	1
nuvoletta			2
	nuvoletta	NOM	2
nuvolo			1
	nuvola	ADJ	1
o			24
	o	CON	24
oca			1
	un'oca	NOM	1
occhio			5
	dell'occhio	NOM	1
	l'occhio	NOM	4
ombra			2
	l'ombra	NOM	1
	ombra	NOM	1
ombrare			1
	d'ombra	VER:pres	1
onda			5
	onda	NOM	5
onde			1
	onde	ADV	1
ondetta			2
	ondetta	ADJ	1
	un'ondetta	VER:pper	1
ondette			1
	ondette	NOM	1

Figura 6: Esempio di visualizzazione delle statistiche sui lemmi e le forme.

4. UN PO' DI DATI

Grazie ad alcuni software di analisi usati nel sito siamo in grado di trarre qualche informazione circa gli accessi, le registrazioni ed, in generale, l'interesse verso il corpus CLIPS.

Dall'apertura del nuovo sito vi sono stati 258 iscritti con una media mensile di 34,29. I mesi con più iscrizioni sono stati Marzo ed Aprile 2010. Riteniamo tale incremento come conseguenza della "pubblicizzazione" del nuovo portale avvenuto a Febbraio a seguito del sesto convegno nazionale AISV. Il progetto non è recentissimo ed il numero medio di

iscrizioni ci porta a pensare che, nonostante sia passato qualche anno, il corpus CLIPS sia ancora attuale ed utile per la ricerca, la didattica e non solo.

Per quanto concerne la nazionalità dei visitatori, essi sono prettamente italiani (75% circa) ma spaziano in tutto il pianeta con 54 nazioni differenti che hanno visitato il portale e con USA e Germania tra le più interessate. Purtroppo i dati in nostro possesso non ci permettono di avere una visione globale delle persone che negli anni si sono interessate dato che, come evidenzia il grafico in basso (Figura 7), le statistiche partono da Novembre 2009 e, nel vecchio sito, non vi era una fase di registrazione.

Per quanto riguarda i downloads, essi sono perlopiù di file .wav (audio) e .txt (testi) contenenti le trascrizioni. Inoltre, le statistiche delineano il *subcorpus* "dialogico" come quello più interessante.

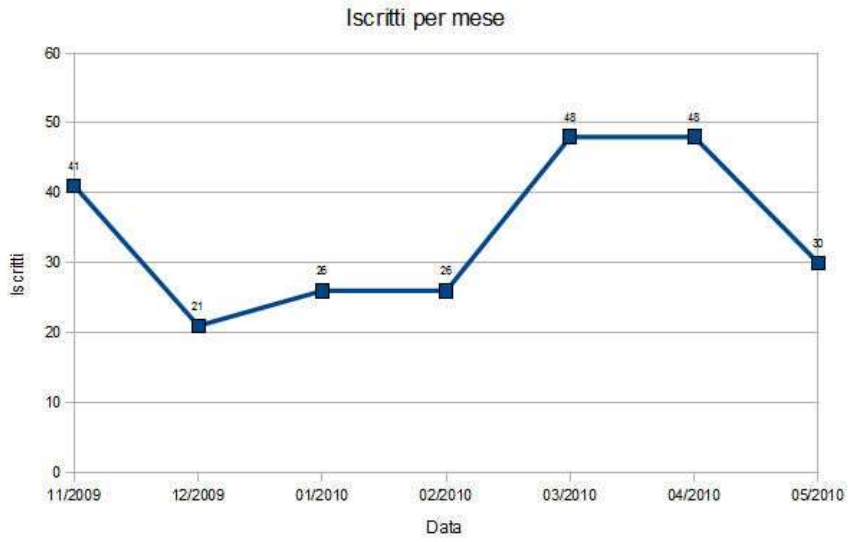


Figura 7: Grafico delle iscrizioni dall'apertura della ristrutturazione del portale.

5. CONCLUSIONI

Il nostro lavoro non è concluso ne conclusivo. Non è concluso perché i software, come ogni attività di ricerca, hanno margini di miglioramento per i dati prodotti e possono dar vita a nuove considerazioni. Inoltre, le prospettive future ci portano ad ipotizzare la creazione di un juke-box virtuale tramite il quale far ascoltare i file audio appartenenti al corpus direttamente online. Questo passo, potrebbe essere accompagnato dall'allineamento dell'audio con i file delle trascrizioni ortografiche ed ortofoniche. Tale progetto, da un lato renderebbe lo studio da parte dei ricercatori più agevole per quanto riguarda la ricerca di specifici fenomeni acustici e prosodici, e dall'altro consentirebbe ai visitatori stranieri una più facile fruizione dei contenuti e darebbe al portale CLIPS anche una veste didattica nell'apprendimento della lingua.

Un'ulteriore innovazione è quella di ampliare l'uso del sito CLIPS destinandolo non solo ai linguisti interessati alla comprensione dei fenomeni dell'italiano parlato, ma si vorrebbe arrivare a coinvolgere persone interessate dall'apprendimento della nostra lingua. Esso si integra con progetti come il FIRB 2009-2012, "Perdita, mantenimento e recupero

dello spazio linguistico e culturale nella II e III generazione di emigrati italiani nel mondo: lingua, lingue, identità. La lingua e cultura italiana come valore e patrimonio per nuove professionalità nelle comunità emigrate", ed il portale Parlaritaliano, il primo osservatorio dedicato allo studio dell'italiano, per l'analisi della lingua italiana parlata e l'elaborazione, nonché la promozione, di modelli di formazione linguistica.

BIBLIOGRAFIA

Adelson-Velskii, G. & Landis, E.M. (1962), An algorithm for the organization of information, *Doklady Akademii Nauk SSSR*, 146, 263-266.

Bird, S., & Liberman, M. (2001), A formal framework for linguistic annotation, *Speech Communication*, 33 (1-2), 23-60.

Cecere, E. (2008), *Un sistema di gestione e interrogazione di corpora linguistici multilivello: ambiente di sviluppo integrato per querying e gestione*, Tesi di laurea in Scienze e tecnologie informatiche, Università di Napoli 'Federico II'.

De Rosa, A. (2009) *Progetto CLIPS: un portale per lo studio e l'analisi di corpora linguistici*, Tesi di laurea in Scienze e tecnologie informatiche, Università di Napoli 'Federico II'.

Savy, R. & Cutugno, F. (2009), CLIPS: diatopic, diamesic and diaphasic variations of spoken Italian, in *Proceedings of Corpus Linguistics 2009*, Liverpool, July 20-23.

Schmid, H., Baroni, M., Zanchetta, E. & Stein, A. (2007) The enriched treetagger system, *Intelligenza Artificiale*, 4 (2), 22-23.

Schmid, H. (2009), http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/