NUOVE FUNZIONALITÀ PER IL PORTALE CLIPS

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

SOMMARIO

Il lavoro realizzato 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 italiano ed inglese e si propone come punto di riferimento per lo scambio di dati tra studiosi della linguistica grazie ai documenti pubblicati e le novità che gli autori vorranno condividere.

Il primo 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 AGset (Bird & Liberman, 2001) in formato XML. Gli AG permettono di gestire i *corpora* di parlato e di rappresentare le annotazioni trascritte da un segnale contenente risorse linguistiche (Cecere, 2008).

La seconda funzionalità, attraverso il linguaggio AGQL, consente di interrogare gli AGset. 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. È possibile agire, 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. L'operazione è eseguibile su vari livelli del *corpus*. 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*. 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 altra funzionalità permette di analizzare il livello di parola dell'anntoazione (corrispondente all'estensione dei file .WRD) di un *corpus* prescelto e produrre un file XML contenente le Part-of-Speach, il lemma ed altre informazioni inerenti le parole di quel dato *corpus*. Lo script usato su cui si basa la funzione è il TreeTagger di H. Schmid (Schmid, 2009).

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, restituisce quante sono le occorrenze delle forme appartenenti a tale lemma. Tutti i lemmi e le forme sono memorizzati in appositi alberi AVL (Adelson-Velskii & Landis, 1962).