La lessicografia per la descrizione e l'analisi dei corpus linguistici
Lexicography for the description and analysis of a linguistic corpus

Ritamaria Bucciarelli
Università Ca' Foscari, Venezia
r.bucciarelli@unive.it

ABSTRACT
The desire to build a corpus in education stems from the urgency to digitally make up for traditional paper-based tools. Fad or face-to-face classroom consultation for vocabulary or grammar activities can improve the educative action. A culture could be accessed through its linguistic uses, such as grammatical and lexical motives: for example, the frequency of a term. The willingness of researchers to give a useful tool for the description, analysis and comparison of written languages was implemented in the distant 1986 when, in collaboration with M. Gross, it was understood that it was required to play on the fixed aspect and non-compositionality of languages in order to effectively carry on professional communications, with the addition of lexicographic tools for the purpose of helping disability and providing sensory guidance. This is how WorkTool was born, that is, a database approach designed to handle lexicon building and to spot token forms in sign languages.

La volontà di costruire un corpora in ambito didattico nasce dall'urgenza di sopperire con strumento elettronico al tradizionale cartaceo. La consultazione diretta in aula o in fad per svolgere attività lessico-grammaticale può rendere efficace l'azione educativa. Le motivazioni grammaticali di azioni linguistiche di un corpus come: le forme lessicali e grammaticali, le molteplici occorrenze di un lemma possono permettere di accedere ad una cultura attraverso i suoi usi linguistici. La volontà dei ricercatori di dare uno strumento utile per la descrizione e analisi e comparazioni delle lingue scritte ha trovato attuazione nel lontanissimo 1986, quando in collaborazione con M. Gross si è capito che bisognava giocare sulla fissità e non composizionalità delle lingue per rendere efficace la comunicazione professionale e nel contempo utilizzare gli stessi strumenti lexicografici per la disabilità sensoriale ed orientativi. Nasce il WorkTool, un approccio su base di dati inteso per la gestione della costruzione del lessico e per l'individuazione di forme elementary nella lingua dei segni.

KEYWORDS
Linguistic corpus, Transduction, Decoding, Translation, Real time production.
Corpus linguistico, Transduzione, Decodifica, Traduzione, Produzione in tempi reali.

Formazione & Insegnamento XII – 4 – 2014
ISSN 1973-4778 print – 2279-7505 on line
doi: 107346/-fei-XII-04-14_27

Introduzione

Nel corso degli anni di studi e di ricerca sono stati messo a punto nuovi metodi per l'indagine linguistica basati essenzialmente sulla costruzione di lessici-sintattici che, giovandosi delle opportunità offerte dalla elaborazione informatica, mirano ad una descrizione, la più esaustiva e formalizzata possibile, di una data lingua. Le ricerche fanno parte del progetto Lessico-Grammatica della lingua italiana L.G.L.I. Il modello teorico di riferimento è rappresentato dalla grammatica "a operatori e argomenti" di (Harris 1957, 1963, 1970.) Ne è derivato un approccio rigorosamente analitico in cui, fermo restando la centralità della sintassi e la rigorosità delle regole trasformazionali, la grammatica di una lingua non va interpretata più come modello astratto, ma viene piuttosto indagata a partire da concreti enunciati. L'attività è stata incentrata sull'approfondimento dei metodi per l'indagine linguistica ed è stata finalizzata anche all'individuazione di modalità di applicazioni curriculari per una moderna glottodidattica.

Il rapporto tra linguistica e informatica ha avuto inizio, con (Gross 1975), come una relazione concernente un dominio fortemente transdisciplinare in cui la linguistica ha realizzato modelli, procedure di tipo informatico per raffinare, formalizzare i propri dati e i propri metodi. Le applicazioni della linguistica all'informatica sono state molteplici, basti ricordare l'analisi sintattica automatica e il trattamento automatico dei dati linguistici. La linguistica costituisce una strategia per la comunicazione per la trasmissione scritta dei nuovi linguaggi sintetici e per i professionisti della scrittura che utilizzano la comunicazione istituzionale. I nuovi strumenti per la didattica mirano a favorire la professionalizzazione del disabile sensoriale, che può superare quasi interamente nel campo della lettura e della scrittura, grazie alle sempre più numerose applicazioni informatiche. Nel progetto sono impegnati ricercatori con competenze specifiche nel settore linguistico, pedagogico ed informatico, ma nel contempo posseggono competenze plurime per la transazione.

1. Premessa

L'ipotesi operativa è nata dalla volontà di trovare metodi e strumenti per l'acquisizione della lingua, per il disabile sensoriale e per gli studenti stranieri, ma soprattutto per capire i meccanismi di manipolazioni e di trasformazione e riformulazione che permettono ad ogni umano di commutare le oralità in testi scritti. Il risultato finale è stato lungo e laborioso, ma ci ha permesso alla fine di creare un corpus didattico intelligente e cioè interattivo con il corpus di scrittura e di traduzione. La ricerca è partita dalla sperimentazione della diffusione delle tecniche generative- trasformazionali e del lessico grammatica nei progetti lingue per gli studenti L_2 e di supportarle con la grammatica prescrittiva. Quando nel 1999 abbiamo conosciuto le grammatiche tedesche (Leuniger 1992) la prima si è interessata alla fonetica descrittiva e più tardi T. Hanche e il secondo alla *il Lexicon a database approach to handle lexicon building and spotting token forms in sign languages,* ci ha permesso di intuire e di perfezionare le nostre ipotesi e cioè che il sordo deve comunicare in codice per poter comparare il linguaggio verbale in codice scritto. L'idea è stata supportata dal metodo Cuzzocrea "DFB" che ha ideato i cheremi per i digrammi cioè ha conferito al cherema il valore del codice linguistico dei grafemi. Il cherema trasdotto in codice conferisce alla lingua dei segni i valori dei grafici e ne permette la manipolazione.

2. Risultati attesi

Il risultato è la creazione del coprus linguistico e di traduzione in cui le entrate sono elencate in ordine alfabetico, corredate da informazioni di tipo morfogrammaticali e suddivise in base alla loro caratteristica di unità di significato autonome e contengono informazioni di carattere morfo-sintattico formalizzate in base alle proprieta distribuzionali e trasformazionali di ogni singolo elemento. Ogni singola tabella importata diventa una grammatica locale pronta per essere applicata durante l'analisi testuale automatica Le liste sono state compilate unicamente seguendo il dizionario della lingua le liste (AC) rilevate da (De Mauro, 2000)e Continuare nella ricerca e trasformare il corpus AC in LIS e cioè portare avanti la ricerca iniziata da colleghi in Germania, e cioè WorkTool, a database approach to handle lexicon building and spotting token forms in sign languages.

3. WorkTool

L'idea di un corpus per la didattica nasce dagli studi approfonditi di comparazione dei codici DGS:

The DGS Corpus project takes place at the *Institute for German Sign Language and Communication of the Deaf* at the University of Hamburg. The project is financed by funds of the academy programme which are borne by the federal government and the federal states (Länder).

WorkTool è composto da una base dati costruito in una prospettiva basata sulla possibilita di creare ontologie elaborate a partire dalla struttura sintattica in cui occorrono specifici predicati (A. ELIA 2013).I temi trattati sono i seguenti:

1. la formalizzazione delle proprieta sintattiche di un insieme specifico di predicati semantici, da effettuare in base alle regole di co-occorrenza e restrizione di selezione;
2. la struttura e l'applicazione di dizionari elettronici generici;
3. la costruzione di automi a stati finiti e trasduttori, da applicare durante l'analisi testualeautomatica.

Si otterrà dunque un dizionario di consultazione, analisi e lettura "quantitativa" dei dati e pertanto strumento didattico di consultazione che favorisce nuove modalità di fruizione delle risorse. Il nuovo digital computazionale è costituito da diversi momenti interagenti, perchè, potenziando gli ambiti delle conoscenze linguistiche, migliora qualitativamente l'*output* informatico (R. Bucciarelli 1986). Il criteri vettore guida che permette l'*input* è la linguistica strutturale. In fase di archiviazione dati l'analisi è diretta al sintagma e ancor più a sintesi di elementi morfemici e lessemici atti a produrre tecniche di apprendimento cognitivo per non complessità strutturale e semantica di testi specialistici. Esso prevede la realizzazione di un *parser* morfologico-sintattico-semantico, che sia in grado di operare scelte strutturali sulla base di codici personalizzati, quali la posizione di un certo costituente, nonché di strutture fisse, frasi sidomatiche, polirematiche ecc., formule utilizzate in varie parti del testo in analisi, allo scopo di intuire i sistemi di trascrizione e aggiunge l'icona della lingua dei segni.

Un file text di produzione, detto anche (*Human search converter*) perché l'operatore utilizza i codici inseriti nel DB mediante interrogazione, al fine di ridurre il testo in digitalizzazione. Inoltre compone il testo e gli attribuisce le pro-

prietà semantiche e d'uso linguistico. È questa segmentazione che manipola il linguaggio attraverso le tecniche di cancellazione, sostituzione, dislocazione, riconversione della tipologia testuale che va a riprodurre[5].

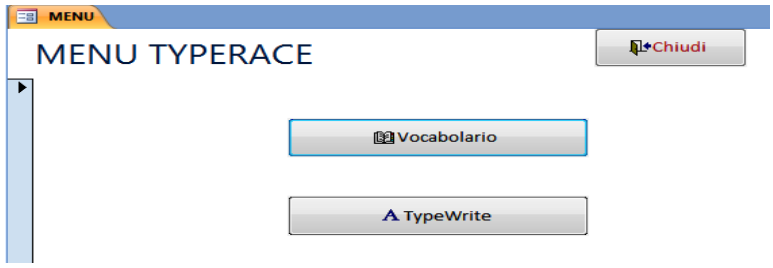
Un file text di riproduzione, trasduzione, traduzione della comunicazione in tempi reali.

Conclusioni

Le ricerche fin qui condotte hanno dato ottimi risultati per la didattica della lingua italiana in moodle di cui io sono docente e non solo: moodle.unive.it

Il prototipo informatico

II progetto affronta il dominio applicativo delle codifiche computazionali. Dovendosi prestare a definire un insieme di possibili vocaboli a rappresentazione del suo codice e, potendo essere espresso in differenti lingue, l'approccio applicativo è stato basato su Microsoft® ACCESS. Il risultato è l'implementazione di due moduli principali per coadiuvare le attività di generazione e sviluppo di sistemi di codifica (*DEFINER*) e per elaborare documenti in codice al fine di ottenere una traduzione veloce ed affidabile (*PARSER*). Lo sviluppo del progetto prevede uno stadio prototipale dei due moduli sopra indicati rilasciati in un unico DB, necessari alla messa a punto delle matrici di codifica e le caratteristiche basilari del *PARSER*. La versione prototipale contiene tutte le prestazioni fondamentali individuate nel progetto e relative al problema della codifica. Il prototipo consentirà la messa a punto dei codici e le verifiche tecniche e prestazionali del *PARSER*. [...]. Caratteristiche della prima versione definitiva, Sistemi di protezione, Aspetti di interfaccia e semplicità di uso, ecc. Fin dalla versione prototipale sono evidenti i due ambiti di lavoro, "Testo" per le attività di stesura e traduzione di documenti in codice sulla base del sistema di codifica (attività del modulo precedentemente denominato *PARSER*) e "Codifica" per consentire la definizione e la manutenzione del sistema di codifica (attività del modulo precedentemente denominato *DEFINER*).Nell'ambito delle attività di stesura e traduzione è disponibile la voce "Vocabolario" che consente la stesura di testo libero contenente codici. Dopo aver provveduto alla stesura del testo, può essere richiesta la traduzione immediata di quanto scritto sulla base del sistema di codifica specificato o disponibile, mediante "TypeWrite" [Fig.1].

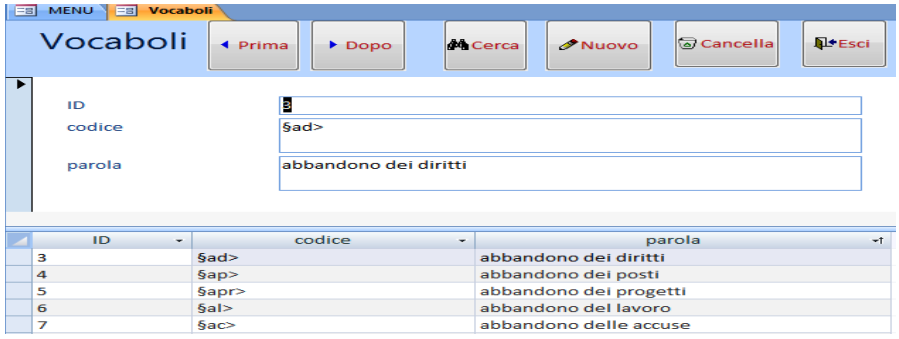


Il menu [fig.1]

Il controllo del *PARSER* consente di individuare esattamente i codici presenti nel testo scritto provvedendo alla traduzione, lasciando inalterato quant'altro ambiguo o inesatto. Il testo tradotto è ancora modificabile ed estendibile, sia nella versione in codice che in quella già tradotta [Fig.2].

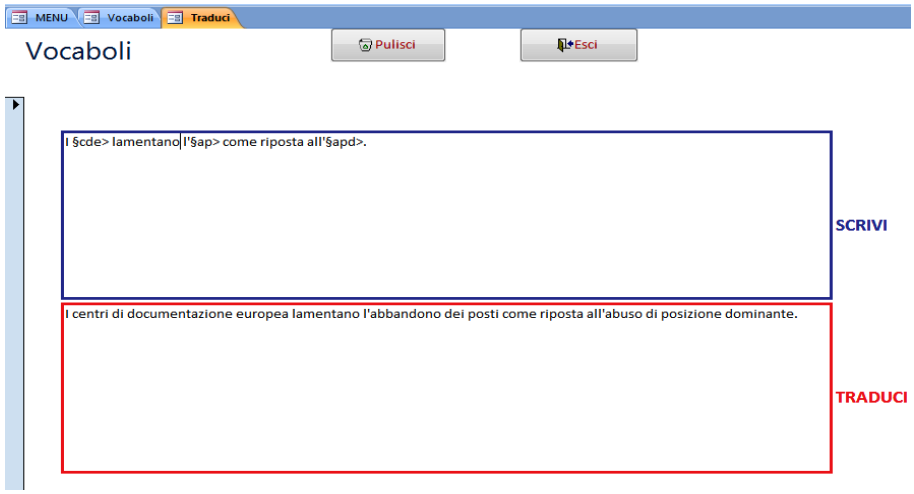
Liste di archiviazione [fig.2]

Nell'ambito dell'attività di definizione e sviluppo del sistema di codifica può essere richiesta una parola chiave dell'accesso alla consultazione e variazione del sistema di codifica, quindi consente la definizione delle strutture del sistema di codifica [Fig3].



Liste di struttura [fig.3]

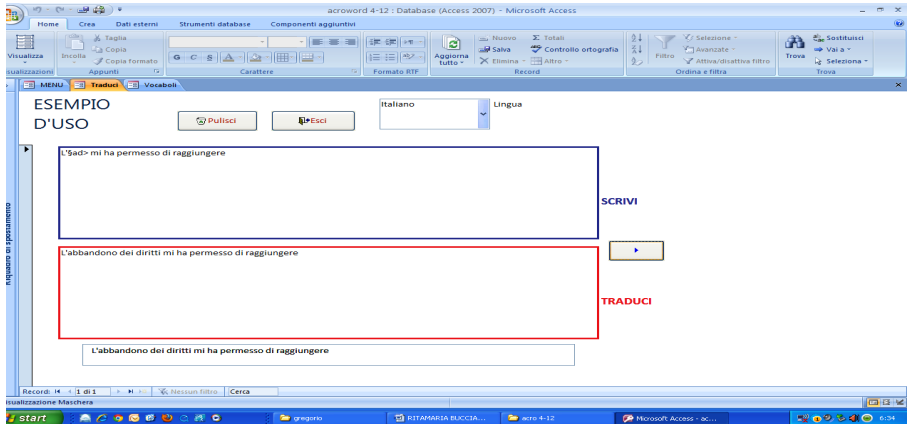
Criteri di proiezione dei Sistemi di Codifica, Rilascio di versioni Race usate solo per stesura e traduzione, ecc. Per ogni sistema di codifica potranno essere specificate le strutture, contenenti simboli caratteristici iniziali come "\", ".", "-", "*", "#" e "§" alle quali sono associate in una relazione "uno a molti" in una corrispondente traduzione. È disponibile fin dalla versione prototipale ogni possibilità di aggiornamento al sistema di codifica. È infatti possibile aggiungere una nuova relazione che associa una struttura ad una traduzione, modificare una relazione precedentemente inserita nel sistema di codifica o eliminare una relazione dopo averla selezionata. Il prototipo di Race richiede risorse elaborative standard per applicazioni monoutente in ambiente Windows ed è progettato per mantenere tali requisiti anche nelle versioni definitive [Fig. 4].



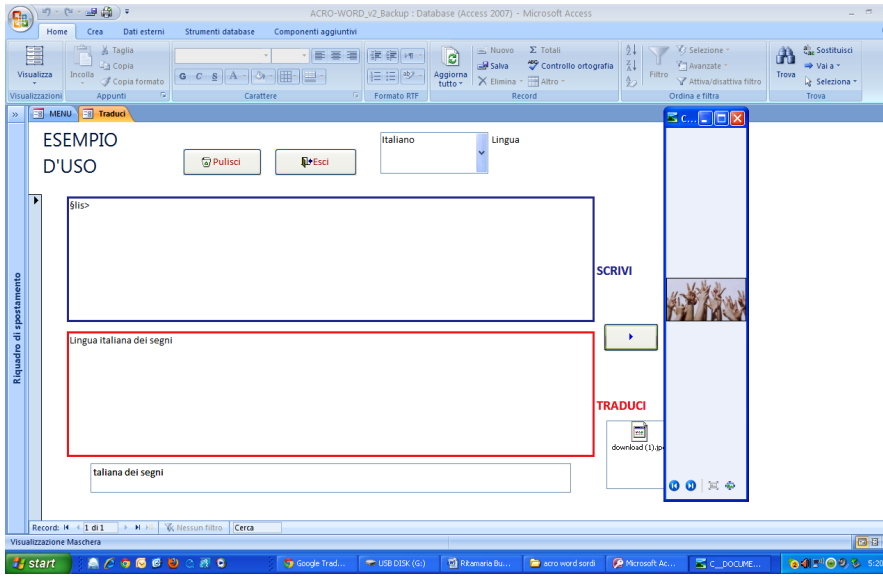
File text o corpus di produzione [fig.4]

Le caratteristiche hardware/software consegnate per il corretto funzionamento di Race sono:

PC 80386 con 4 MB o superiori; Windows XP o superiori; ACCESS 2007 e oltre. Essendo una applicazione progettata negli aspetti realizzativi sotto gli standard di interfaccia Windows, con menù altamente intuitivi, non richiede dettagliato manuale operativo [Fig. 5].



Esempio d'uso



Esempio d'uso [fig. 5]

Riferimenti bibliografici

Bucciarelli, R. (2013). *Didattica e tecnologie di rete*, risorsa educativa aperta elettronica, Virtual Library, Università Ca' Foscari di Venezia, reperibile presso il link: http://cird.unive.it/dspace/ della scheda dei metadati della risorsa; vi si accede cercando la risorsa nella library, e cliccando sul link proposto dal sistema: http://cird.unive.it/dspace/handle/123456789/1044.

Bucciarelli, R., © software Aco-Word: Copora language teaching and translation of the Italian language $DTELA_{CF}$.

Bucciarelli, R., Galdi, A. (a cura di) (2013). *Project work Immersion in the textual typologies of italian writing*. (collana di linguistica e glottologia). Salerno: I.R.I.S.

D'Agostino, E., Elia, A., Vietri, S. (2004). *Lexicon grammar, Eletronic dictionaires and local grammar in italian, in leclère, C. La porte, È Piot, M.Silberzten, Sintax lexis and lexicon –Grammar. Papers in hounour of M. Gross., J. Benjamins*. Amsterdam- Philadelphia.

De Mauro, T. (1994). I linguaggi scientifici. In De Mauro T. (a cura di). *Studi sul trattamento linguistico dell'informazione scientifica*. Roma: Bulzoni,

Elia, A., D'agostino, E., Martinelli, M., EMDA (2001). *Lexicon –grammar della lingua italiana*. Napoli: Liguori,

Hohenberger, A. (1996). *le categorie funzionali e acquisizione del linguaggio: auto-organizzazione di un sistema dinamico*. Tesi di dottorato, Johann Wolfgang Goethe-Universität di Francoforte, Germania.

Konradi, J. (2006). *La terapia Afasia nella prima fase acuta? Un gruppo di studio per confrontare gli effetti del trattamento e remissioni spontanee*. Milano-Lodi: IPAVSI.