Testi e linguaggi 4-2010 II bozza:Testi e linguaggi.qxd 09/12/10 17:26 Pagina 121

Computer Learner Corpora e sistemi di annotazione dell'errore

di *Linda Barone*

Introduzione

L'influenza delle tecnologie informatiche e l'utilizzo di banche di dati linguistiche nella didattica delle lingue straniere sono visibili nella progettazione di un curricolo, nei materiali didattici, nelle tecniche e nella metodologia dell'insegnamento. L'incidenza della linguistica dei *corpora* sull'insegnamento/apprendimento di una lingua straniera ha avuto un impatto notevole sui sillabi, soprattutto grazie all'acquisizione di descrizioni dettagliate della lingua in uso, mettendo in evidenza, per esempio, l'importanza del lessico e di frasi istituzionalizzate, semi-istituzionalizzate e collocazioni attraverso lo studio di fenomeni ricorrenti[1].

A partire dagli anni Novanta del secolo scorso si sono moltiplicate ricerche *data-based* e *data-driven* su campioni di lingua che hanno evidenziato quanto i dati a disposizione possano essere utili per mettere in luce schemi lessicali e sintattici fondamentali nella *progettazione* di materiali didattici, grammatiche, dizionari, libri di testo ecc. quanto più vicini alla lingua in uso. Per non parlare dell'uso che può essere fatto di *corpora* di nativi da parte di docenti e di studenti che si accostano alla lingua in modo originale e altamente motivante trovando nella ricerca, soprattutto di tipo induttivo, un valido alleato nella costruzione delle competenze linguistiche.

Un discorso diverso è rappresentato dall'utilizzo di banche dati di produzioni di apprendenti una lingua straniera per l'esplorazione delle difficoltà di acquisizione linguistica, grazie anche alla presenza sempre più massiccia di strumenti informatici utili all'analisi delle deviazioni dalla lingua target. Se un *corpus* di testi scritti e/o parlati di nativi evidenzia ciò che è tipico della lingua madre e di conseguenza si rende utile per quanto più su detto, i *corpora*, d'ora in avanti CLC (*Computer Learner Corpus/Corpora*) di apprendenti una lingua straniera evidenziano ciò che si discosta dal tipico uso nativo dal punto di vista morfosintattico e lessicale, non soltanto relativamente all'errore, ma anche per determinate scel-

Testi e linguaggi 4-2010 II bozza:Testi e linguaggi.qxd 09/12/10 17:26 Pagina 122

te linguistiche legate all'abuso o sottoutilizzo di alcuni pattern e all'aggiramento o elusione di altri.

Un CLC, come d'altronde un database di produzioni native, deve rispettare alcune regole ben precise relative alle tecniche per reperire i dati e all'uso di software per analizzarli. Le caratteristiche più rilevanti che un CLC deve avere sono la "lunghezza" del *corpus*, che dipende dall'analisi che si intende effettuare (i CLC oggi a disposizione vanno da diversi milioni a poche decine di migliaia di parole), la rappresentatività relativamente alla varietà o al genere linguistico e testuale di interesse, nonché alle caratteristiche che accomunano gli apprendenti (livello dell'interlingua, contesto di apprendimento, informazioni relative all'età e al sesso), alla possibilità di osservare lo stesso gruppo di apprendenti in senso sincronico e diacronico e al riferimento ad altri *corpora* per analisi contrastive.

Quest'ultimo punto è in realtà molto controverso:

> Is it right to study a learner *corpus* as an incomplete version of the target language rather as a self-contained system?… The choice of a native reference model is particularly complex for a widespread International language like English. Different norms may apply to learners of English as a foreign language in different areas of the world, which are more British – or American-oriented and to parts of Asia and Africa where second language norms have been standardized[2].

Tra i CLC più rappresentativi vanno ricordati il *Cambridge Learner Corpus* (CLC), il *Longman Learner's Corpus* (LLC), e l'*International Corpus of Learner English* (ICLE) che, a differenza dei primi due, che sono *corpora* "'commerciali", è un *corpus* cosiddetto "accademico". Per l'analisi dei dati contenuti nei *corpora* si può ricorrere a diversi software, tra i quali il più noto e citato nella ricerca scientifica è senz'altro il *WordSmith Tools*[3] e tale analisi può avvenire, come già detto, a diversi livelli. Il presente lavoro prende in esame alcuni strumenti utilizzati per la descrizione degli errori nei CLC e ha lo scopo di illustrare brevemente due progetti e software di annotazione dell'errore e di utilizzarne uno in particolare applicandolo a un database di produzioni scritte di apprendenti italiani di lingua inglese localizzati presso alcuni atenei nazionali tra i quali quelli di Catania, Roma e Napoli sedi con le quali il Dipartimento di Studi Linguistici e Letterari dell'Università di Salerno ha avviato tempo fa un progetto, coordinato dalla professoressa Bruna Di Sabato, per la creazione di un *learner corpus*, denominato UNISALC (*University of Salerno Learner Corpus*)[4].

Testi e linguaggi 4-2010 II bozza:Testi e linguaggi.qxd 09/12/10 17:26 Pagina 123

La *Computer-aided Error Analysis* (CEA), di cui si discute in questo lavoro, ha origine dall'analisi dell'errore che ha conosciuto una notevole fortuna, ma anche diverse critiche negative, negli anni Settanta dello scorso secolo. La metodologia alla base dell'analisi dell'errore era suscettibile di attacchi soprattutto in riferimento ai dati utilizzati che erano superficialmente associati a liste di errori comuni e questa pratica non teneva conto delle variabili linguistiche e della varietà degli apprendenti i cui risultati finivano con l'essere delle raccolte non naturali, approssimative e molto eterogenee di dati[5]. Un altro difetto dell'analisi dell'errore precedente alla CEA era da attribuirsi proprio ad una spiegazione non corretta degli errori per i quali venivano presentate delle tassonomie carenti e confuse basate su caratteristiche linguistiche a volte non osservabili o soggettive e contenenti categorie sovrapposte che rendevano l'identificazione dell'errore particolarmente difficile. Nell'era della CEA, la prima critica mossa all'analisi dell'errore sembra ormai superata grazie alle caratteristiche già enunciate con le quali i CLC sono oggi compilati, ma il secondo punto riguardante l'etichettatura degli errori è questione ancora aperta.

Diversi sono i progetti in atto sui sistemi di annotazione dell'errore e qui si cercherà di delinearne due in particolare e poi di applicare uno di questi al già citato database UNISALC. Il sistema scelto per l'analisi è quello dell'Università Cattolica di Louvain (UCLEE, *Université Catholique de Louvain Error Editor*) ma si introdurrà brevemente anche l'annotatore di errori del CLC, per capire in cosa differiscono. Entrambi sono basati sull'uso dell'inglese come lingua straniera e seconda (EFL e ESL), ma il primo è un annotatore di errori che proviene da un progetto accademico e il secondo è invece relativo al CLC che è di tipo commerciale, ossia nato allo scopo di fornire dati utili alla creazione di dizionari, grammatiche e di altri materiali didattici.

1
Sistema di annotazione del *Cambridge Learner Corpus*

Il *Cambridge Learner Corpus* (CLC) è un grande database di composizioni scritte di apprendenti di inglese come lingua straniera ed è un progetto della Cambridge University Press in collaborazione con Cambridge ESOL[6]. Il *corpus* contiene più di trenta milioni di occorrenze ed è stato quasi interamente etichettato con il software di annotazione dell'errore che illustreremo di seguito. La maggioranza delle etichette è basata su un

Testi e linguaggi 4-2010 II bozza:Testi e linguaggi.qxd 09/12/10 17:26 Pagina 124

sistema di annotazione di due lettere, di cui la prima rappresenta la categoria generale e la seconda identifica la classe di parole. Nella TAB. 1 vengono presentate le categorie e sottocategorie dell'*error editor* del CLC[7].

TABELLA 1
Categorie e sottocategorie del CLC *error editor*

General types of error (first letter)
F: wrong Form used
M: something Missing
R: word or phrase needs Replacing
U: word or phrase is Unnecessary
D: word is wrongly Derived
Le etichette M, R e U possono apparire da sole quando non si è in possesso di ulteriori informazioni

Word classes (second letter)
A: Pronoun (Anaphoric)
C: Conjunction
D: Determiner
J: Adjective
N: Noun
Q: Quantifier
T: Preposition
V: Verb
Y: Adverb (-Y)

Punctuation errors:
MP: punctuation Missing
RP: punctuation needs Replacing
UP: Unnecessary punctuation

Countability errors:
CN: countability of Noun error
CQ: wrong Quantifier because of noun countability
CD: wrong Determiner because of noun countability

False friend errors:
FF: False Friend error

Agreement errors:
AGA: Anaphoric (pronoun) agreement error
AGD: Determiner agreement error
AGN: Noun agreement error

(*segue*)

Testi e linguaggi 4-2010 II bozza:Testi e linguaggi.qxd 09/12/10 17:26 Pagina 125

TABELLA 1 (*segue*)

AGV: Verb agreement error

Additional error codes:
AS: Incorrect Argument Structure
CE: Compound Error
CL: Collocation error
ID: Idiom error
IN: Incorrect formation of Noun plural
IV: Incorrect Verb inflection
L: Inappropriate register (Label)
S: Spelling error
SA: American Spelling
SX: Spelling confusion error
TV: Wrong Tense of Verb
W: Incorrect Word order
X: Incorrect formation of negative

L'etichettatura del CLC, e di altri *corpora* "commerciali", a differenza di quella dei *corpora* "accademici", non ha il fine di creare una tassonomia dell'errore degli apprendenti, ma mira a individuare gli errori più frequenti creando delle categorie che, una volta analizzate, serviranno come base imprescindibile alla creazione di dizionari, grammatiche e libri di testo.

Un *corpus* annotato ha il vantaggio, rispetto ad un *corpus* non annotato, di sveltire il lavoro degli analisti che possono facilmente individuare le aree problematiche, deselezionando le frasi "senza errori" (etichettate con "NE", no error) per concentrarsi su "cosa gli studenti tendono a sbagliare". D'altro canto, la possibilità di avere liste separate di concordanze con e senza errori, aiuta a capire anche "cosa gli studenti sanno fare", ambito di indagine quasi sempre ignorato.

Un altro vantaggio dell'annotazione del *corpus* è la capacità del sistema di individuare velocemente gli errori cosiddetti di "omission" e "commission", ossia produzioni errate in cui qualcosa manca o qualcosa è usato in modo improprio:

Perhaps the greatest advantage over an uncoded corpus is that we can search for errors of omission as well as commission. After searching through a concordanced search on "at", for example, in an uncoded corpus, it is possible to locate errors such as the unnecessary use of the preposition (e.g. *watching at the television), and the erroneous use of the preposition (e.g. *she invited me at her birthday party). However, it is not possible easily to find:

Testi e linguaggi 4-2010 II bozza:Testi e linguaggi.qxd 09/12/10 17:26 Pagina 126

– Instances of failure to use the preposition (coded <#MT>) where it is required (e.g. *we looked each other);
– Instances of where "at" should have been the chosen preposition, but a wrong preposition was chosen instead (coded <#RT> (e.g. *we arrived to our destination). This would not be possible without the addition of and ability to search on a corrected version.
(Nicholls, http://ucrel.lancs.ac.uk/publications/CL2003/papers/nicholls.pdf)

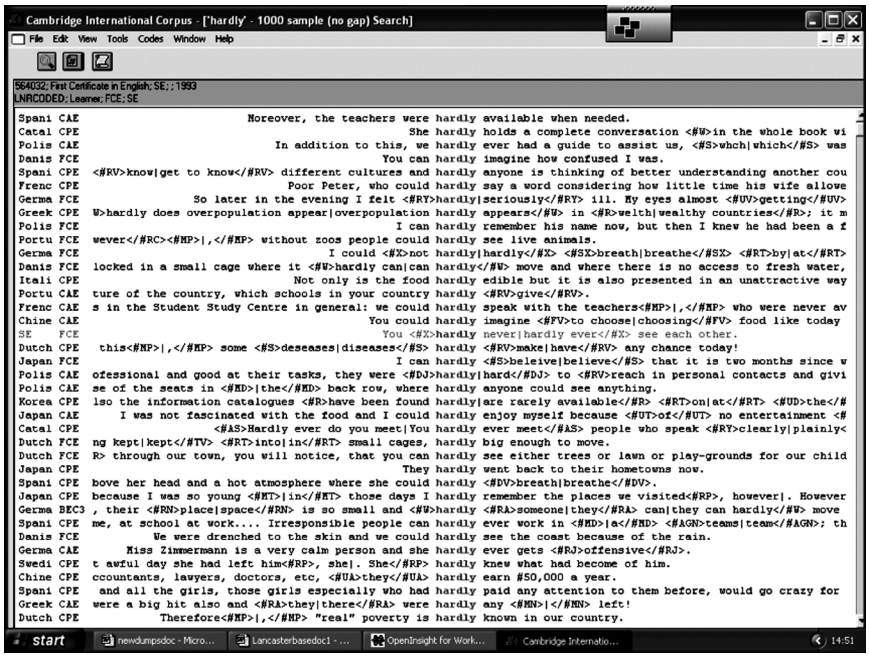
Un esempio di come funziona il software è riportato nelle figure seguenti.

Gli errori sono etichettati in questo modo:

<#ETICHETTA>*PAROLA SBAGLIATA*|*PAROLA CORRETTA* </#ETICHETTA>

La figura di seguito riportata contiene una ricerca su una singola parola, ossia la "keyword in context" di colore rosso è sempre la stessa, in questo caso l'avverbio "hardly".

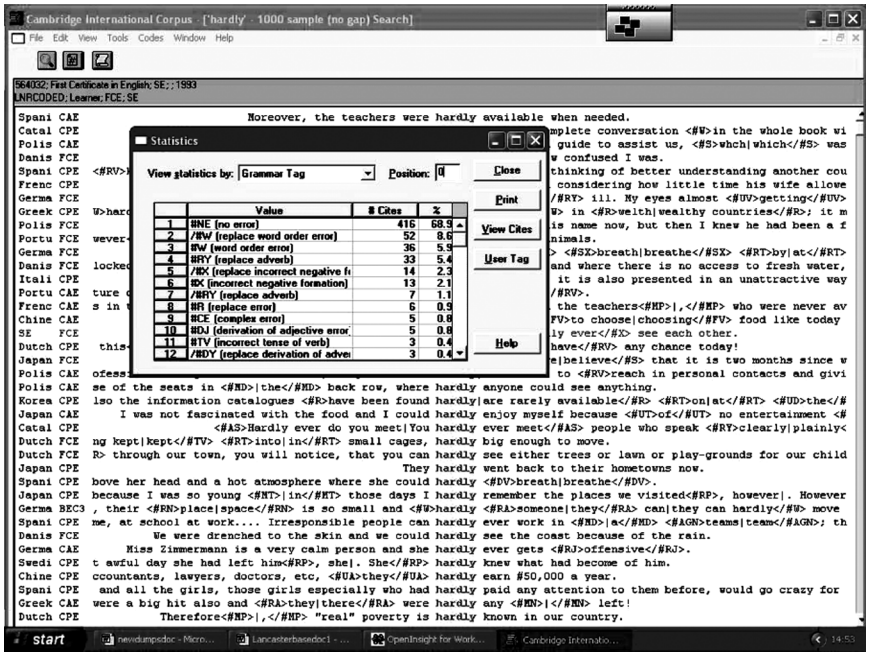
FIGURA 1



Testi e linguaggi 4-2010 II bozza:Testi e linguaggi.qxd 09/12/10 17:26 Pagina 127

La schermata contiene sulla sinistra la lingua madre dell'apprendente seguita dall'esame sostenuto (ad esempio, nel primo caso abbiamo uno studente spagnolo che sostiene l'esame per la certificazione CAE). Una volta ottenuta la schermata, attivando la funzione "statistics" si possono ottenere informazioni preziose sulla percentuale dell'uso corretto del termine e sul tipo di errore; in questo caso l'errore più frequente (con incidenza dell'8,6%) è di natura sintattica come si vede nella figura che segue.

FIGURA 2



Un altro vantaggio di questo sistema di annotazione sta nel fatto che si può individuare un'area problematica costituita non solo da parole singole, ma anche da gruppi di parole; in questo caso si propone una schermata con l'uso errato di sostantivi non numerabili, ossia sostantivi, quali "information", che non prendono la "s" del plurale.

Testi e linguaggi 4-2010 II bozza:Testi e linguaggi.qxd 09/12/10 17:26 Pagina 128

FIGURA 3

Anche qui, utilizzando il comando "statistics" si può facilmente verificare l'incidenza degli errori e si nota che gli errori più frequenti nell'uso degli "uncountable nouns" sono "informations" e "advices" con percentuali rispettivamente del 23,9 e del 7,9%. È anche possibile, sempre con la funzione "statistics", classificare gli errori per lingua madre e, nel caso di "informations", si nota che gli studenti che maggiormente incorrono nell'errore sono di madrelingua francese, seguiti dai tedeschi e dai portoghesi.

Prima di procedere alla descrizione e all'applicazione del sistema UCLEE, si vuole qui fare una considerazione che verrà ampiamente illustrata in seguito sui possibili limiti che tali software presentano. La lettura delle etichette sopra riportate può risultare di non facile interpretazione e può generare confusione perché l'errore può essere etichettato e corretto in modi diversi a seconda della sua natura. Molto spesso ci si trova di fronte a casi ambigui, nei quali riconosciamo immediatamente l'errore, ma non sappiamo quale sia il modo migliore per etichettarlo. Ad

Testi e linguaggi 4-2010 II bozza:Testi e linguaggi.qxd 09/12/10 17:26 Pagina 129

esempio, nel caso della frase «I've decided to use my third wish for my brother because I'm really proud of him and because we have always been very close each other as if we were the same person»[8], il testo sottolineato indica la zona problematica il cui errore è evidente, ma non è altrettanto evidente identificare l'etichetta appropriata e la conseguente correzione da apportare. In questo caso la decisione oscilla tra due possibilità: aggiungere la preposizione mancante dopo "close" oppure eliminare "each other". Questo, tra gli innumerevoli esempi possibili, suggerisce che la gerarchia dell'errore non esiste, o si potrebbe dire che la scelta da parte dell'etichettatore è libera e, di conseguenza, a nostro parere approssimativa.

2
Sistema di annotazione dell'Université Catholique de Louvain

Il software UCLEE (*Université Catholique de Louvain Error Editor*)[9] è, come il precedente, uno strumento utile al fine di etichettare errori contenuti in testi scritti in lingua inglese. Il programma è di tipo semi-automatico perché permette all'utente, dopo aver individuato l'errore e la sua natura, di etichettarlo scegliendo tra le categorie di errore che si espongono di seguito. La correzione dell'errore non è automatica, poiché la revisione va inserita manualmente. L'identificazione delle tipologie di errore da parte di UCLEE nasce dall'analisi di un *corpus* di saggi scritti da apprendenti di inglese come lingua straniera. A partire dal database e dagli errori selezionati sono state individuate sette macrocategorie: forma, grammatica, lessico-grammatica, lessico, parola, registro e stile. Per ciascuna categoria sono state determinate delle sotto-categorie e per alcune sotto-categorie esistono delle ulteriori suddivisioni, che tentano di illustrare al meglio il tipo di errore riscontrato.

Nella TAB. 2 vengono presentate le categorie e sottocategorie di UCLEE.

TABELLA 2
Categorie e sottocategorie di UCLEE

Form (F):
- Morphology
- Spelling

(*segue*)

Testi e linguaggi 4-2010 II bozza:Testi e linguaggi.qxd 09/12/10 17:26 Pagina 130

TABELLA 2 (*segue*)

Grammar (G):
– Articles
– Nouns (noun case, noun number)
– Pronouns
– Adjectives (adjective order, adjective number, comparative/superlative)
– Adverbs
– Verbs (verb number, verb morphology, non-finite/finite verb, verb voice, verb tense, auxiliaries)
– Word class

Lexico-Grammar (X):
– Complementation
– Dependent preposition
– Nouns (countable/uncountable)

Lexis (L):
– Lexical phrase
– Lexical single (false friends)
– Connectives [logical connectors (single/complex), coordinating conjunctions, subordinating conjunctions]

Word (W):
– Redundant
– Missing
– Order

Register (R)

Style (S):
– Incomplete
– Unclear

3
UCLEE in teoria e in pratica

Gli autori del software UCLEE sono consapevoli che le distinzioni non possono ritenersi assolute e che alcune categorie di errore possono essere sovrapposte, ma sostengono che un buon analista deve essere in grado di scegliere l'etichetta esatta tra le quaranta disponibili e quindi fugare ogni dubbio sulla natura dell'errore e la sua appartenenza ad una specifica categoria. I principi che regolano il corretto uso del software sono sei, verranno qui esposti in modo riassuntivo e accompagnati da alcuni

Testi e linguaggi 4-2010 II bozza:Testi e linguaggi.qxd 09/12/10 17:26 Pagina 131

esempi[10]. Su tali regole ritorneremo perché, come si vedrà, molte di esse presentano delle eccezioni che possono rendere più complessa la scelta dell'etichetta e il suo inserimento nel punto preciso della frase.

1. Non etichettare sulla base della forma corretta, ma soltanto su quella della forma errata.

Solo partendo dall'errore si potrà essere in grado di selezionare l'esatta etichetta. *Sheila's best quality is <u>the</u> optimism*: l'errore qui riscontrato è l'uso dell'articolo determinativo, per cui l'etichetta da scegliere è GA, ossia Grammar Articles e non GP, Grammar Pronouns, anche se la frase corretta dovrà essere *Sheila's best quality is <u>her</u> optimism* con l'aggiunta dell'aggettivo possessivo femminile.

2. L'etichetta va posta immediatamente prima dell'errore.
Si consideri l'esempio:

Anyway I'm sure that all the people who take time for (GP) <u>theirselves</u> $themselves$ to relax live in a better and happier way than people who live thinking only about what they have to do here and now.

Si nota che l'etichetta da apporre, in questo caso GP (*grammar pronoun*), precede l'errore.

3. La correzione va posta immediatamente dopo l'errore ed è compresa tra i simboli $.

4. Quando una stessa parola presenta due errori le etichette da inserire sono due. Ad esempio, se leggiamo una parola quale *happynesses*, essa presenta un errore di ortografia (FS, Form Spelling) e un errato uso del sostantivo non numerabile (XNUC, lexico-grammar nouns uncountable countable). L'etichettatura dovrà essere: (FS) happynesses $happiness$ (XNUC) happynesses $happiness$.

5. Per indicare la presenza di parole in eccesso o la mancanza di una parola va usato il simbolo "o".

Nel caso in cui la frase presenti parole che vanno cancellate l'etichettatura dovrà essere così effettuata:

Once one of my teachers asked (XVPR) to o me to play with the other children on the merry-go-round.

Testi e linguaggi 4-2010 II bozza:Testi e linguaggi.qxd 09/12/10 17:26 Pagina 132

In questo caso l'errore riscontrato è l'uso della preposizione "to" dopo il verbo "to ask" e l'etichetta corrispondente è XVPR ossia *verbs used with the wrong dependent preposition*.

Nel caso opposto, vale a dire quando una parola manca, la correzione avviene in questo modo:

I still remember all (GA) o the days I spent with my dog Fritz.

6. Gli errori che derivano da una correzione già effettuata non vanno etichettati e corretti. Il sesto e ultimo principio afferma che, se all'interno di una frase si riscontra un errore, questo va certamente etichettato e corretto, ma se, dopo la correzione, il resto della frase risulta conseguentemente errato, l'etichettatura non può essere effettuata. Per semplificare si riporta il seguente esempio: *he placed a handicap in my way*. L'errore che qui si rileva è di carattere lessicale; il sostantivo "handicap" non è un collocato del verbo "to place" e la correzione da apportare dovrebbe essere "obstacle". Tale correzione porterebbe di conseguenza alla frase: *he placed a obstacle in my way* che risulta comunque errata per l'uso dell'articolo indeterminativo "a" davanti a vocale. In questo caso, dunque, va etichettato l'errore lessicale, ma non il conseguente uso errato dell'articolo che è solo un effetto della correzione apportata.

Come più su affermato, i principi che regolano il corretto utilizzo del software UCLEE presentano delle eccezioni che potrebbero rendere il lavoro di etichettatura più complesso e insidioso.

Lo scopo di questo lavoro non è certo quello di mettere in discussione l'efficacia del software UCLEE, ma di sottolineare che in alcuni casi la scelta dell'etichetta diventa difficile non soltanto per le eccezioni ai principi, ma anche perché spesso si possono incontrare delle tipologie di errore che ricadono in più categorie o etichette o in nessuna di quelle a disposizione.

I settori problematici riguardano in particolare le macroaree *forma*, *grammatica*, *lessico-grammatica* e *lessico* la cui differenza non è sempre chiara e le cui categorie sembrano a volte sovrapporsi.

I concetti di *forma*, *grammatica*, *lessico* e *lessico-grammatica* sono molto generali e proprio questa vaghezza può causare confusione in chi utilizza un software come quello che si sta illustrando.

Per tentare di chiarire i concetti e cercare di distinguere in modo netto le quattro macroaree, si presentano di seguito alcune definizioni come appaiono nel *Dictionary of Linguistics and Phonetics*[11], in *The Handbook*

Testi e linguaggi 4-2010 II bozza:Testi e linguaggi.qxd 09/12/10 17:26 Pagina 133

of Discourse Analysis[12], nel *Dizionario di linguistica e di filologia, metrica, retorica*[13] e nell'*Oxford Dictionary of Linguistics*[14].

La categoria "forma" in UCLEE è suddivisa in due sezioni: morfologia e ortografia.

Nel dizionario di Crystal leggiamo:

> Morphology (n.) The branch of grammar which studies the structure or forms of words, primarily through the use of the morpheme construct. It is traditionally distinguished from syntax, which deals with the rules governing the combination of words in sentences. It is generally divided into two fields: the study of inflections (inflectional morphology) and of word-formation (lexical or derivational morphology) – a distinction which is sometimes accorded theoretical status (split morphology).

Appare immediatamente chiaro che la morfologia è parte integrante della grammatica, per cui distinguere errori di "forma" da errori di "grammatica" può provocare confusione in chi utilizza il software. Tra gli esempi di errori presentati nella categoria "forma, morfologia (FM)" troviamo la terza persona singolare del verbo "to watch" scritta "watchs" che però potrebbe appartenere anche alla categoria "spelling", oppure l'uso errato del comparativo irregolare di "bad" scritto "badder" che può ricadere anche nella categoria "lessico".

In Beccaria la definizione di "ortografia" è la seguente:

> Ortografia (gr. Orthographìa "scrittura corretta") Indica il modo di scrivere correttamente le parole e, come tale è quella parte della grammatica che si interessa dei fenomeni grafici, proponendo e codificando regole, mutevoli nel tempo, in base all'uso. […] L'o., dunque, regola la scrittura, il modo di dividere le parole, o per andare a capo, l'uso delle maiuscole e delle minuscole, i sistemi di punteggiatura.

Anche in questo caso la definizione non dà adito a dubbi, l'ortografia è parte della grammatica. In più, nelle categorizzazioni che troviamo nel dizionario di Beccaria non si fa menzione dell'ortografia relativamente alle varietà diatopiche di una lingua, ma tra gli esempi di errori ortografici che leggiamo nel manuale UCLEE, appaiono "center" e "color" che sono etichettati come errori perché viene usato lo spelling americano e non quello britannico.

Passando al concetto di grammatica non è facile fornire una definizione concisa data la vastità di implicazioni e fenomeni connessi ad essa e anche il disaccordo che regna tra gli studiosi sulla natura e sul signifi-

Testi e linguaggi 4-2010 II bozza:Testi e linguaggi.qxd 09/12/10 17:26 Pagina 134

cato del termine. E, dopo aver consultato i dizionari e le enciclopedie a disposizione, ci si rende conto che una definizione che in poche righe ci dia l'idea di cosa sia la grammatica in senso generale non esiste. L'unica "delimitazione" del termine – ma a nostro avviso non molto chiara proprio a causa della ricerca di sinteticità laddove impossibile – la troviamo nel dizionario di Matthews in cui si legge:

Grammar Any systematic account of the structure of a language; the patterns that it describes; the branch of linguistics concerned with such patterns. Often restricted to relations among units that have meaning. Hence opp. phonology: e.g. singing is a grammatical unit, as are sing and -ing, while [s] or the syllable [sj] are phonological. Also opposed, though again not always, to a dictionary or the lexicon. E.g. the meanings of sing belong to its entry in the lexicon: the role of -ing to grammar, where it is described for verbs in general. When limited in these ways the study of grammar reduces to that of morphology and syntax.

Tale definizione, seppur essenziale, rende l'idea che per *grammatica* in senso generale si intende lo studio della morfologia e della sintassi. Quindi, ancora una volta si sottolinea che i problemi morfologici andrebbero inclusi nella categoria "grammatica" o che almeno si dovrebbe scindere la categoria "grammatica" in "morfologia" e "sintassi", termine quest'ultimo che non compare affatto nelle macroaree UCLEE.

Nella categoria "grammatica" di UCLEE si inserisce l'etichetta GNC (Grammar, Noun Case) che prevede la correzione di usi impropri della "s" del possessivo (saxon genitive) con l'esempio *behind the Berlin's wall*, da correggere in *behind the Berlin wall*. Nella categoria "forma" troviamo invece l'etichetta FM (Form, morphology) e tra gli esempi denominati "inflectional errors" compare *l'uso errato del genitivo sassone* con l'esempio *girls's* da correggere con *girls'*. Seppure indubbio che i due errori presentati siano di natura diversa, la comparsa del genitivo sassone in entrambe le macroaree può causare confusione e conseguentemente un'errata identificazione dell'etichetta da apporre.

Spostandoci al concetto di *lessico*, la ricerca della definizione ideale appare subito ardua. Secondo le regole del manuale UCLEE il lessico comprende gli errori che riguardano le proprietà semantiche di parole o frasi. Nello specifico, l'area è suddivisa in tre sotto-categorie che citiamo di seguito.

Lexical Single (LS) is used for conceptual or collocational lexical errors in single words only. Included in this sub-category are solid and hyphenated compounds;

Testi e linguaggi 4-2010 II bozza:Testi e linguaggi.qxd 09/12/10 17:26 Pagina 135

Lexical Phrase (LP) includes most errors in (semi-)fixed multi-word expressions and idioms. Compound words separated by a blank are included in the sub-category;
Lexis, Connectives (LC) consists of errors involving connectives: coordinating conjunctions, subordinating conjunctions or logical connectors».

Qualche dubbio sorge immediatamente leggendo la terza sotto-categoria che comprende l'uso errato di congiunzioni (ma si dovrebbe aggiungere che i *connectives* possono essere anche avverbi o alcuni tipi di verbi, quali quelli con funzione sintattica predicativa). Leggendo la definizione che fornisce Crystal del termine *connective* si capisce immediatamente che il termine indica qualcosa di relativo alla grammatica, in particolare alla sintassi, ma non al lessico.

Connective (adj./n.) A term used in the grammatical classification of words to characterize words or morphemes whose function is primarily to link linguistic units at any level. Conjunctions are the most obvious types, but several types of adverb can be seen as connectives (*therefore*, *however*, *nevertheless* etc.), as can some verbs (the copulas *be*, *seem*, *feel* etc.).

Ma torniamo al concetto più generale di lessico e a come esso viene definito in Beccaria:

Lessico (gr. lexicón [biblión] "libro delle parole"). Insieme delle parole e delle locuzioni che compongono una lingua, sia essa la lingua intera di una comunità o una sua parte, ad es. un sottocodice o la lingua di uno scrittore o di un parlante: si parla quindi del l. dell'italiano trecentesco, del l. dell'economia, del l. leopardiano, del l. di un bambino di tre anni. Nell'accezione comune è visto come formato dalle unità lessicali dotate di significato non grammaticale ed è contrapposto alla grammatica intesa come insieme di regole che governano la combinazione di parole in frasi. […]

È evidente che il *lessico* in senso ampio si riferisce a tutte le parole che compongono una lingua, ma generalmente, come appena descritto, esso viene opposto alla *grammatica*; questo implica che per evitare confusione tra i termini, sarebbe il caso di considerare la *grammatica* e il *lessico* come due macroaree, la prima comprendente morfologia e sintassi, la seconda tutti gli aspetti relativi all'uso delle *unità lessicali dotate di significato non grammaticale*.

Per concludere questa panoramica sulle definizioni è importante soffermarsi sulla lessico-grammatica che unisce i due termini e che apre la stra-

Testi e linguaggi 4-2010 II bozza:Testi e linguaggi.qxd 09/12/10 17:26 Pagina 136

da a riflessioni linguistiche da prospettive del tutto diverse e strettamente legate alla linguistica dei *corpora*, in particolare al concetto di concordanza.

The introduction of concordancing as a common tool of linguistic research changed the traditional view of the internal structure of human language i.e. "the open-choice principle"[15]. […] Certain collocations are rather fixed, though they are not necessarily idioms (in the traditional meaning of the word) and certain vocabulary and grammar items often co-occur in an idiomatic manner. "Words do not occur at random in a text and the open-choice principle does not provide for substantial enough restraints on consecutive choices" [16]. The new schematic approach introduces the notion of lexico-grammar and combines the previously separate fields of grammar and vocabulary[17].

Concludendo, sulla lessico-grammatica non si possono non citare studiosi quali Halliday e Lewis che hanno fatto di questo concetto una base essenziale della loro ricerca scientifica. Halliday e Matthiessen[18], ci spiegano la loro idea di lessico-grammatica in questo modo:

We are interpreting lexicogrammar not as an isolated system, but rather as an integral subsystem of language-in-context. More specifically, lexicogrammar is the stratum that is internal to language, interfacing with both semantics and phonology/graphology; and together with semantics it forms the content system of language.

Lewis[19], partendo dal principio cardine dell'approccio lessicale per cui la dicotomia lessico/grammatica non ha ragione di esistere, ci illustra in modo pratico che cosa egli intenda per *grammaticalized lexis*:

Language consists not only of traditional grammar and vocabulary but of multi-word prefabricated chunks […] and these chunks occupy a crucial role in facilitating language production" […] If you learn *initial reaction* (one item) it is easy to split the chunk apart, and acquire *initial* and *reaction*, two more items. If you learn the words separately, you must also know a third item, the correct collocation. Separating collocations into their component words is easy; it is considerably more difficult to put words together to form natural collocations.

Questa lunga parentesi sulla lessico-grammatica serve a comprendere bene il significato del termine e soprattutto a confrontare le sopraindicate citazioni con quanto leggiamo nel manuale UCLEE a proposito della categoria *lexico-grammar*, la quale viene suddivisa in tre sotto-categorie nessuna delle quali sembra appartenere all'idea che di lessico-grammatica hanno gli studiosi citati.

Testi e linguaggi 4-2010 II bozza:Testi e linguaggi.qxd 09/12/10 17:26 Pagina 137

Nella prima suddivisione troviamo tutti gli errori definiti di *complementation*, ad esempio l'errato complementatore del sostantivo *possibility* (**students have the possibility to leave*), nella seconda parte della lessico-grammatica abbiamo l'uso errato di preposizioni dopo sostantivi, aggettivi e verbi e nella terza vengono inclusi tutti gli errori relativi ai sostantivi numerabili e non numerabili. Nessun accenno all'interno della classe lessico-grammatica di UCLEE viene fatto rispetto a concetti portanti quali la collocazione o l'uso errato di espressioni idiomatiche, di verbi sintagmatici, di espressioni istituzionalizzate o semi-istituzionalizzate.

Per concludere il discorso sull'analisi e commento del software UCLEE, qualcosa va detto anche in merito alla categoria chiamata "word" che non appare chiara in quanto innanzitutto non si capisce perché essa debba essere separata dalla categoria "lexis", ma anche perché la suddetta categoria presenta una suddivisione in tre sotto-categorie che sono "ridondanza", "omissione di parole" e "ordine delle parole" che non appaiono ben *comprese* nella macroarea, considerando che il concetto di ridondanza appartiene alla grammatica e/o alla semantica e soprattutto che l'ordine delle parole nelle frasi sembra riguardare problemi di tipo sintattico piuttosto che lessicale.

Conclusioni

I *learner corpora* e i sistemi di annotazione dell'errore rivestono un'importanza fondamentale nel dibattito scientifico e sono soprattutto utili ai fini della progettazione di materiali didattici che possano facilitare l'acquisizione della lingua straniera cercando di superare gli ostacoli che determinati apprendenti incontrano naturalmente nel proprio percorso.

Allo stesso tempo va detto che gli studi sull'annotazione dell'errore hanno molta strada da percorrere non soltanto perché il campo è di recente nascita, ma anche perché i software in uso oggi presentano delle lacune in merito all'uso e alla chiarezza nella scelta delle etichette.

Seppure le limitazioni maggiori siano ormai ricordi del passato e le tassonomie degli errori siano state modificate nel tempo e rese più attendibili, è ancora vero che non soltanto non c'è accordo universale da parte dei ricercatori sull'uso di uno schema unico di annotazione, ma anche che quelli presenti non sono del tutto affidabili proprio per le critiche che si muovevano alla vecchia analisi dell'errore, ossia accavallamento delle categorie e confusione nella gestione delle etichette. Tali limiti fanno ritenere a molti ricercatori che la *Computer-Aided Error Analysis* non sia del

Testi e linguaggi 4-2010 II bozza:Testi e linguaggi.qxd 09/12/10 17:26 Pagina 138

tutto convincente e di conseguenza utile nella comunità scientifica e nell'applicazione dei risultati per lo sviluppo di materiali didattici efficaci.

Note

1. M. Lewis, *The Lexical Approach: The State of* ELT *and a Way Forward*, Language Teaching Publications, Hove 1993; M. Lewis, *Pedagogical Implications of the Lexical Approach*, in S. Coady, T. Huckin (eds.), *Second Language Vocabulary Acquisition*, Cambridge University Press, Cambridge 1997; M. Lewis (ed.), *Teaching Collocation: Further Developments in the Lexical Approach*, Language Teaching Publications, Hove 2000.

2. M. T. Prat Zagrebelsky (ed.), *Computer Learner Corpora, Theoretical Issues and Empirical Case Studies of Italian Advanced* EFL *Learners' Interlanguage*, Edizioni dell'Orso, Alessandria 2004, pp. 45-6.

3. Il software *WordSmith Tools*, ideato da Mike Scott (Università di Liverpool), è composto da una serie di strumenti per l'analisi di *corpora* linguistici; tali strumenti, dapprima utilizzati dalla Oxford University Press per ricerche lessicografiche utili alla compilazione di dizionari e grammatiche, è ora usato in ambito scientifico per ricerche di diversa natura finalizzate alla comprensione dei meccanismi linguistici.

4. Questa operazione ha coinvolto diverse sedi universitarie dislocate in tutto il territorio nazionale. Si è provveduto a distribuire alle sedi coinvolte un pacchetto composto dai seguenti documenti: 1. Placement test, 2. Scheda dati rilevazione studenti, 3. Scheda illustrazione fasi operative (per docenti), 4. Lettera di istruzioni (per studenti), 5. Tracce per composizioni. Nel corso del primo semestre dell'a.a. 2006-2007, i colleghi che hanno aderito al progetto hanno provveduto alla raccolta dei dati e alla somministrazione ai loro studenti del test di piazzamento e delle prove di scrittura. Nel corso del secondo semestre di lavoro, i dati pervenuti hanno reso possibile la creazione di un *corpus* di circa 90.000 occorrenze. Gli obiettivi del progetto sono riassunti essenzialmente nello studio e nell'analisi del fenomeno Interlingua visto nel suo complesso. In primo luogo, si è proceduto a delineare le caratteristiche formali dell'interlingua attraverso l'analisi e l'etichettatura delle strutture sintagmatiche (SN, SV, SP) complesse impiegate dagli apprendenti. Si è poi proceduto ad identificare le principali aree di errore attraverso le metodologie della *Error Analysis* (EA). Lo scopo è di isolare tratti specifici caratterizzanti singoli gruppi in determinate aree geografiche ed evidenziare eventuali variabili diatopiche. Si è proceduto all'analisi computerizzata dell'errore attraverso l'impiego di specifici programmi di concordanze (*ConcApp* e *WordSmithTools*) e di etichettatura degli errori (UCLEE – *Université Catholique de Louvain Error Editor*). In questa fase del progetto ci si è prevalentemente soffermati su una ridefinizione di alcuni schemi di etichettatura dell'errore presenti nel programma UCLEE che risulta ad una prima indagine incompleto ai fini del progetto e non prevede una serie di possibili combinazioni di errori che gli studenti italiani sono portati a commettere nell'ambito dell'interlingua.

5. R. Ellis, *The Study of Second Language Acquisition*, Oxford University Press, Oxford 1994, pp. 49-50.

6. Gli esami ESOL (*English for Speakers of Other Languages*) sono organizzati dall'Università di Cambridge e si articolano in un'ampia tipologia di prove, aumentate progressivamente nel corso degli anni, miranti ad accertare nei candidati il livello di conoscenza della lingua inglese sia scritta che parlata nelle quattro abilità fondamentali.

7. Maggiori informazioni sul *Cambridge Learner Corpus* sono reperibili sul seguente sito: http://www.cambridge.org/elt/corpus/learner_corpus2.htm.

8. L'esempio riportato è parte del *corpus* UNISALC cui si è fatto riferimento nell'introduzione.

9. Il programma è disponibile al sito http://cecl.fltr.ucl.ac.be/.

Testi e linguaggi 4-2010 II bozza:Testi e linguaggi.qxd 09/12/10 17:26 Pagina 139

10. Tutti gli esempi di errori riportati fanno parte del database UNISALC cui si è fatto riferimento nell'introduzione.

11. D. Crystal, *Dictionary of Linguistics and Phonetics*, Blackwell Publishing, Oxford 2007.

12. D. Schiffrin, D. Tannen, H. E. Hamilton (eds.), *The Handbook of Discourse Analysis*, Blackwell Publishing, Oxford 2005.

13. G. L. Beccaria (a cura di), *Dizionario di linguistica e di filologia, metrica, retorica*, Piccola Biblioteca Einaudi, Torino 2004.

14. P. H. Matthews, *Oxford Dictionary of Linguistics*, Oxford University Press, Oxford 1997.

15. J. Sinclair, *Corpus, Concordance, Collocation*, Oxford University Press, Oxford 1991, p. 109.

16. *Ibid.*

17. J. Krajka, *Corpora and Language Teachers: From Ready-Made to Teacher-Made Collections*, in *CORELL*, *Computer Resources for Language Learning*, 1, 2007, p. 37.

18. M. A. K. Halliday, C. Matthiessen, *Systemic Functional Grammar: A First Step into the Theory*, Arnold, London 1997, p. 42.

19. Lewis, *Pedagogical Implications of the Lexical Approach*, cit., p. 7.

Testi e linguaggi 4-2010 II bozza:Testi e linguaggi.qxd 09/12/10 17:26 Pagina 140