Analisi dell'interlingua e sistemi di annotazione: implicazioni teoriche ed indagini empiriche

di *Rita Calabrese*

Introduzione

L'impiego di *corpora* elettronici nella ricerca linguistica si rivela particolarmente efficace quando ai dati raccolti (*raw data*) vengono aggiunte informazioni di carattere linguistico e funzionale che facilitano la loro interpretazione da un punto di vista morfosintattico. Il risultato di tali procedure di analisi è rappresentato da *Treebanks*[1] ovvero *corpora* annotati sintatticamente attraverso rappresentazioni ad albero.

Solitamente distinta dalla procedura di *mark-up*[2], l'annotazione di un *corpus* realizzata con l'applicazione di sistemi computerizzati (*parsers*) che eseguono un'etichettatura delle parti del discorso (*POS tagging*) costituisce una procedura particolarmente efficace nel chiarire, ad esempio, casi di ambiguità grammaticale in cui un lessema, se privo di annotazione, può essere interpretato sia come verbo (V) che come nome (N) (ad esempio *work*). I *parsers* generalmente propongono descrizioni delle produzioni linguistiche di una comunità di parlanti basate su regole grammaticali estratte da un campione di dati e tale caratteristica talvolta non permette di eseguire l'analisi automatica anche di frasi non grammaticali[3]. Si tratta, dunque, di sistemi di analisi basati su grammatiche di precisione che distinguono principalmente ciò che è grammaticale da ciò che non lo è basandosi su un *corpus* necessariamente limitato di dati, in quanto la descrizione fa esclusivamente riferimento a campioni di frasi grammaticalmente corrette di una lingua inseriti nel sistema. Tuttavia, la propensione dei parlanti a produrre costruzioni che presentano "deviazioni" dalla norma e il fatto che molti di essi usino una lingua straniera come strumento di comunicazione comporta in sé «that naturalistic ungrammatical sentences are of interest to linguists studying language production, language loss and language learning, and that the grammatical/ungrammatical distinction cannot therefore be completely dismissed»[4].

Con la distinzione tra *competence* e *performance*, Chomsky[5] intese sottolineare che la lingua d'uso prodotta dai parlanti può presentare im-

perfezioni ed errori (*slips*) rispetto alla competenza linguistica in sé, pertanto per ottenere una descrizione accurata di una lingua "reale", sarà necessario rappresentare anche tali imperfezioni all'interno dei sistemi descrittivi adottati nelle analisi linguistiche: «it is necessary to determine the ungrammatical sentences in a language in order to determine the grammatical ones»[6].

Partendo dunque dal presupposto che descrizioni "realistiche" della lingua d'uso devono tener conto anche dei fenomeni di deviazione e dalla considerazione che tale necessità possa estendersi anche all'analisi dell'interlingua, il presente contributo nasce nell'ambito di uno studio preliminare condotto su un campione di dati inclusi in un *corpus* (*University of Salerno Learner Corpus*; UNISALC) di produzioni scritte da studenti di inglese come lingua straniera provenienti da alcuni atenei del Centro-Sud d'Italia[7] con l'obiettivo di illustrare le caratteristiche tecniche e le capacità descrittive di alcuni dei più diffusi sistemi di annotazione.

In particolare, sono stati presi in esame il sistema di annotazione semi-automatica dell'*Université Catholique de Louvain Error Editor* (UCLEE) e il *Visual Interactive Syntax Learning* (VISL). I livelli di specificazione delle analisi che si ottengono con ciascun sistema saranno esposti con esempi tratti dal *corpus* UNISALC seguendo due criteri principali:

a) adeguatezza del repertorio di etichette del sistema quanto a *coverage* per estensione/qualità dell'analisi prodotta rispetto ai dati in esame/da analizzare;

b) adeguatezza del tipo di procedura semi-automatica e computerizzata rispetto agli obiettivi dell'analisi proposta sul campione di dati in esame.

L'ambito di indagine a cui fa riferimento lo studio comprende quindi l'etichettatura degli errori (*error tagging*), delle parti del discorso (*POS tagging*) e delle loro funzioni sintattiche (*parsing*) nell'interlingua.

In particolare, l'indagine condotta sul campione di dati è incentrata sull'interpretazione e conseguente annotazione dei sintagmi preposizionali che costituisce una delle aree della grammatica della lingua in cui risultano più evidenti le diverse strategie messe in atto dagli apprendenti a livello semantico-sintattico e la loro difficoltà ad integrare automaticamente struttura sintagmatica ed informazione lessico-semantica. Inoltre, poiché la nozione di argomento opera a livello sintattico e semantico nel determinare la valenza di un verbo e le sue strutture di sottocategorizzazione (*subcategorization frames*), lo studio tenderà a verificare come i sistemi di annotazione automatica interpretano

tali casi e se possono aiutarci a capire gli usi devianti dalla norma dei sintagmi preposizionali e della complementazione verbale da parte degli apprendenti di inglese L2.

Per verificare tale ipotesi il *corpus* di dati raccolti è stato prima annotato automaticamente utilizzando le applicazioni disponibili sull'interfaccia VISL (*Visual Intercative Syntax Learning*) che forniscono informazioni di tipo sintattico e semantico su una determinata struttura in costituenti, poi i sintagmi preposizionali presenti nel *corpus* sono stati analizzati in base alla testa (preposizione, verbo, nome) e alla funzione sintattica (argomento o aggiunzione) utilizzando il *concordancer* ConApp. Infine, i risultati dell'annotazione automatica sono stati confrontati con quelli derivanti dall'annotazione computerizzata dell'errore.

Nella sua natura di *work-in-progress*, il presente contributo presenta quindi riflessioni su procedure metodologiche volte a individuare non solo la reale natura delle produzioni interlinguistiche degli apprendenti di lingua straniera, ma anche le caratteristiche strutturali delle produzioni *non-nativelike*, di cui gli analizzatori sintattici dovrebbero tener conto. L'analisi statistica delle costruzioni sintattiche nell'interlingua dovrebbe, infatti, consentire di: 1. individuare i casi più frequenti di *non-nativelikelyhood* insieme alla natura e all'origine dell'errore; 2. approfondire la comprensione dei meccanismi di costruzione sintattica che operano a livello profondo nella L1 come nella L2[8].

1
Sistemi di annotazione: un'analisi su campione

Uno dei principali vantaggi dei *corpora* annotati consiste nella possibilità da essi offerta di essere utilizzati come risorsa disponibile a vari livelli di interrogazione qualitativa e quantitativa, in cui l'annotazione costituisce un "valore aggiunto" (*added value*)[9], in quanto rende esplicita l'analisi dei dati oggettivi che diversamente rimarrebbero semplici elementi registrati nel *corpus*: «annotation only means undertaking and making explicit a linguistic analysis. As such, it is something that linguists have been doing for centuries»[10].

Una questione ampiamente dibattuta riguarda l'accuratezza e l'affidabilità delle analisi proposte dai diversi sistemi di annotazione che possono essere suddivisi secondo tre diverse procedure: manuale, semi-automatica e automatica. Con l'annotazione automatica il computer interpreta i dati linguistici senza l'intervento manuale dell'analista, il quale

nella fase di programmazione del sistema ha stabilito le regole e gli algoritmi necessari ad interpretarli (*machine learning algorithms*).

I sistemi di annotazione semi-automatici sono costituiti da un'interfaccia tra il computer e l'analista che consente a quest'ultimo di intervenire per risolvere casi dubbi accrescendo di conseguenza il livello di accuratezza e l'attendibilità dei risultati rispetto alle analisi condotte attraverso procedure completamente automatizzate. L'annotazione può essere eseguita su più livelli: da quello fonologico, prosodico, morfologico e lessicale delle parti del discorso fino al livello di analisi sintattica e semantica.

L'efficacia dei sistemi manuali di annotazione è stata ampiamente discussa nella letteratura sull'argomento[11]. Da tali studi è emersa tra tutte la necessità di condurre l'analisi in base a parametri comuni prefissati secondo le procedure correnti del *Natural Language Processing* (vedi più avanti) che aumentano, di conseguenza, il grado di standardizzazione delle analisi.

1.1. Procedure manuali di annotazione degli errori

Nell'ambito delle ricerche sull'acquisizione delle lingue straniere, l'elaborazione di sistemi di annotazione degli errori si è rivelato uno strumento veramente efficace capace di fornire un'analisi dell'interlingua basata su dati statistici quantitativamente rilevanti. Diversamente dalle ricerche sull'analisi dell'errore ispirate ai principi dell'analisi contrastiva, più attenta a cogliere le differenze dell'interlingua rispetto alla lingua target piuttosto che a studiarne le caratteristiche peculiari e i tratti tipici più ricorrenti anche nei fenomeni di deviazione, la nuova metodologia di analisi computerizzata dell'errore denominata CEA (*Computerized Error Analysis*) si differenzia da quella di tipo tradizionale innanzitutto per la quantità di dati "elettronici" sottoposti ad analisi, per la definizione di chiari criteri oggettivi nella costruzione del *corpus* e per la creazione di una tassonomia di errori ispirata alle categorie grammaticali individuate già nella *Comprehensive Grammar of English*[12].

La sistematicità è dunque il tratto caratteristico della metodologia CEA che con l'individuazione di precisi criteri di raccolta e di analisi garantisce un livello maggiore di attendibilità e rappresentatività utili ai fini della formulazione di generalizzazioni. Gli studi condotti sino ad ora applicando la metodologia CEA sono stati incentrati su specifiche aree di errore, ad esempio sull'uso dei tempi verbali da parte de-

gli apprendenti[13] o sulla reggenza nominale e verbale[14] o di selezione della preposizione[15].

Nel campo delle ricerche sull'acquisizione delle lingue sono numerosi i sistemi di annotazione degli errori degli apprendenti finora creati[16]. La differenziazione dei sistemi di analisi costituisce al momento un elemento di criticità in quanto si riscontra una mancanza di uniformità delle tassonomie di errore che evidenzia la problematicità emergente dall'annotazione automatica degli errori[17].

Secondo Ellis[18], le procedure adottate nell'etichettatura dell'errore dovrebbero prevedere le seguenti fasi:

– identificazione dell'errore;
– descrizione dell'errore;
– spiegazione dell'errore;
– valutazione dell'errore.

Pertanto l'annotazione dovrebbe contenere la classificazione della categoria linguistica (grammatica, verbo, morfema, tempo) e la tassonomia della modificazione rispetto alla lingua target (omissione, aggiunta, errata formazione) laddove al contrario si manifesta la tendenza a ricostruire le deviazioni linguistiche presenti nei dati degli apprendenti attraverso le regole della lingua d'arrivo.

L'annotazione manuale degli errori riflette spesso specifici obiettivi di ricerca che corrispondono a determinate prospettive teoriche adottate ed assumono pertanto un valore relativo a seconda della validità dei principi teorici presi in esame che spesso non riescono a cogliere né le proprietà più generali dell'interlingua né la sua natura fortemente instabile e permeabile in continua trasformazione.

In definitiva, ciò che emerge chiaramente da tali studi sull'annotazione dell'errore è la prospettiva di analisi prevalentemente incentrata sullo scarto qualitativo tra interlingua e lingua d'arrivo con la conseguente inefficacia descrittiva relativamente all'interlingua quale sistema linguistico a se stante. Pertanto, al fine di raccogliere informazioni importanti che possano contribuire a definire meglio le caratteristiche dell'interlingua nella prospettiva teorica più generale delle ricerche sull'acquisizione della L2 (*SLA research*), è necessario compiere un'accurata annotazione dei dati degli apprendenti abbandonando la prospettiva *error-based* in favore di una descrizione per parti del discorso (POS) e funzioni sintattiche. Studi recenti condotti in questa direzione[19] hanno spesso, infatti, evidenziato le difficoltà inerenti l'annotazione dell'errore in casi dubbi quali ad esempio in (1):

(1) *he had became more and more conscious of his mind problem*

che mostra un certo disaccordo tra la morfologia del verbo (*past tense*) e la sua distribuzione (*past participle*), ma che potrebbe anche indicare un semplice errore di ortografia (*spelling*).

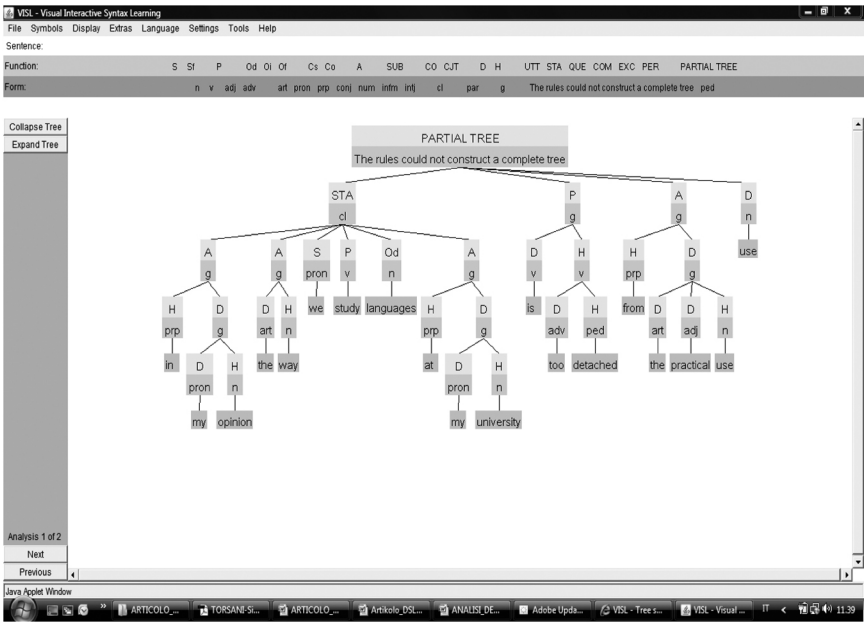
Alla luce di tali considerazioni, la prospettiva adottata nella ricerca empirica presentata nel par. 3 è stata esattamente opposta a quella generalmente impiegata nell'analisi dell'errore: si è partiti dall'annotazione dei dati senza presupporre l'occorrenza di errori, osservando quindi lo stesso schema di annotazione della dipendenza sintattica adottato per i dati di L1.

1.2. Il sistema di annotazione automatica della piattaforma VISL

Il *Visual Interactive Syntax Learning* (VISL)[20] è il risultato di un progetto, che prevede l'analisi (sintattica) di diverse lingue germaniche e romanze, realizzato presso l'*Institute of Language and Communication* dell'Università Southern Denmark – Odense Campus. VISL[21] consiste, in primo luogo, in un programma di analisi sintattica che interfacciandosi con un testo dato in una determinata lingua produce simultaneamente un diagramma ad albero del tipo illustrato nella FIG. 1.

Il sistema altamente *process-oriented* e interattivo è costruito su una complessa rete di pagine HTML, in cui banche dati di testi o frasi costruite su "grammatiche pedagogiche" e annotate manualmente si interfacciano con i programmi *Perl* e *Java* e altri strumenti basati sulla *Constraint Grammar* (CG) per l'analisi automatica del *corpus* di dati. La CG è un paradigma metodologico usato nell'analisi grammaticale delle lingue naturali costruito su regole di dipendenza contestuale in base alle quali vengono assegnate alle parole (*tokens*) del testo in esame delle etichette di lemmatizzazione, flessione, derivazione morfologica, funzione sintattica, dipendenza, valenza, ruolo dei casi, tipo semantico. Ogni regola aggiunge o rimuove, a seconda del contesto di occorrenza delle parole, determinate etichette. La "robustezza" del sistema deriva dal fatto che le regole vengono applicate automaticamente in maniera progressiva senza rimuovere l'interpretazione immediatamente precedente. In ogni rappresentazione ad albero prodotta dal sistema, ai dati da analizzare inseriti in un'apposita finestra della piattaforma vengono assegnate delle etichette riportanti informazioni relative alla forma o categoria del discorso (con-

FIGURA 1
Rappresentazione ad albero di un esempio tratto dal *corpus*



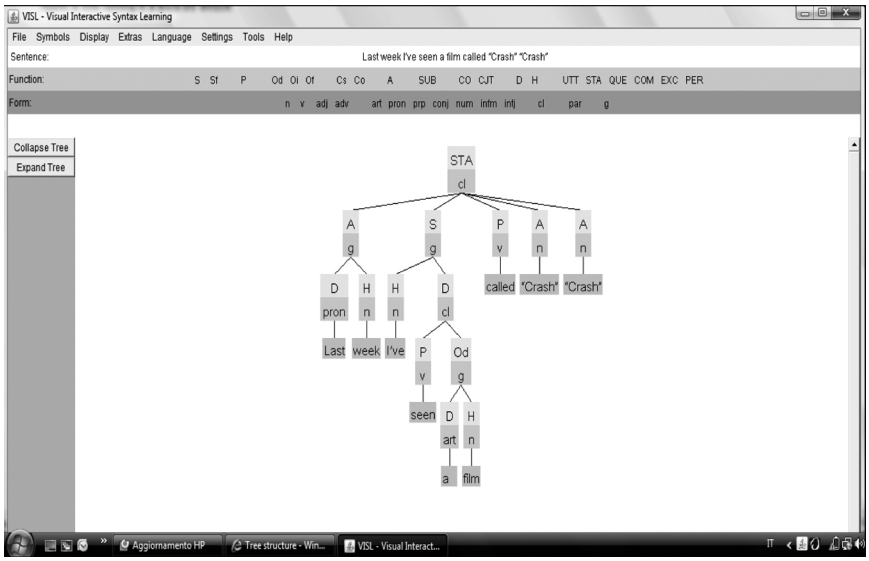
trassegnata in grigio scuro nel diagramma ad albero della FIG. 1) e alla funzione (in grigio chiaro) a livello di clausola, sintagma e singola entrata lessicale. La TAB. 1 illustra l'elenco completo di etichette per ciascun nodo del diagramma ad albero.

La completezza del sistema di annotazione unitamente alla sua duttilità e semplicità d'uso lo rendono uno strumento estremamente utile ed efficace nell'elaborazione di analisi sintattiche. Inoltre, il sistema svolge anche una funzione di "tutorial sintattico": selezionando nella finestra *Tools* la funzione *Build Tree*, il programma produce un albero sintattico per la frase che si vuole analizzare, ma privo di annotazione, che sarà completato a cura dell'analista. Tuttavia, l'interfaccia non presenta funzioni destinate all'immagazzinamento di informazioni utili a svolgere analisi statistiche e, nel caso di rappresentazioni ad albero, il sistema non è in grado di rilevare deviazioni dalle norme grammaticali che regolano, ad esempio, l'uso dei tempi verbali come mostrato nell'esempio tratto dal nostro *corpus* (FIG. 2).

TABELLA 1
Elenco delle etichette di VISL suddivise secondo forma e funzione

Function	Form
S = Subject	n = noun
Sf = Formal/Provisional Subject	v = verb
P = Predicator	adj = adjective
Od = Direct Object	adv = adverb
Of = Formal/Provisional Object	art = article
Cs = Subject Complement/Subject Predicative	pron = pronoun
Co = Object Complement/Object Predicative	prp = preposition
A = Adverbial	conj = conjunction
SUB = Subordinator	num = numeral
Co = Coordinator	infm = infinitive marker
CJT = Conjunct/Conjoint	intj = interjection
D = Dependent	cl = clause
H = Head	par = paratagma/compound unit
UTT = Utterance	g = group
STA = Statement	
QUE = Question	
COM = Command	
EXC = Exclamation	
PER = Performative	

FIGURA 2
Rappresentazione ad albero di un esempio con errore di selezione del tempo verbale



La piattaforma offre inoltre un altro importante strumento di annotazione automatica dei dati attraverso il quale le relazioni di dipendenza (*dependency relation structure*), che costituiscono la principale condizione di proiezione della testa (nominale, verbale o preposizionale), vengono rappresentate dal simbolo @ posto prima (>) o dopo (<) della testa stessa come mostrato in (2):

(2) characters [character] N P NOM **@P<** of [of] PRP **@N<** the [the] ART S/P **@>N** play [play] N S NOM

L'analizzatore sintattico quindi genera non solo strutture in costituenti e le etichette corrispondenti, ma anche le rappresentazioni delle relazioni di dipendenza tra i costituenti; pertanto, tale applicazione diviene particolarmente utile, ad esempio, nei casi riguardanti l'interpretazione dei legamenti preposizionali come argomenti o aggiunzioni rispetto ad una data testa.

2
Lo studio

Lo studio parte dall'assunto che gli apprendenti di L2 impiegano strategie di categorizzazione (*parsing*) qualitativamente diverse da quelle adottate dai parlanti nativi a livello semantico-sintattico e che quindi manifestano maggiori difficoltà ad integrare automaticamente struttura sintagmatica ed informazione di tipo lessico-semantico. Per verificare tale ipotesi ci si è soffermati sulla struttura argomentale e di complementazione dell'interlingua scritta di un gruppo di studenti di inglese come lingua straniera. Inoltre, poiché il concetto di argomento si colloca a livello di interfaccia lessico-semantica determinando la valenza di un verbo e la sua struttura di sottocategorizzazione, esso costituisce la principale ipotesi di ricerca del presente studio, vale a dire che le categorie potenzialmente indeterminabili (*fuzzy*) di argomento o aggiunzione possono condurre a frequenti errori di interpretazione e di uso dei legamenti preposizionali da parte degli apprendenti.

Per verificare tali ipotesi di ricerca sono stati utilizzati i dati raccolti in un *corpus* di composizioni scritte da studenti universitari di inglese come lingua straniera per determinare in primo luogo:

1. le strutture di sottocategorizzazione più frequenti nella loro IL;
2. il tipo più ricorrente di legamento preposizionale (argomento o aggiunzione);

3. la possibile interpretazione dei legamenti preposizionali evidenziata dall'uso o deviazioni da parte degli studenti in esame.

L'analisi si basa dunque sull'ipotesi che concetti puramente sintattici, quali quelli di argomento o aggiunzione, sono più problematici di altri e possono condurre a fenomeni di deviazione nell'uso delle preposizioni e della complementazione verbale da parte degli apprendenti di inglese L2. In particolare, essi tenderebbero ad usare in modo più diretto correlazioni (*mappings*) forma-funzione, piuttosto che strategie universali basate sulla struttura sintagmatica, dimostrando dunque un'evidente incapacità di integrare struttura sintagmatica e informazioni lessico-semantiche e di rielaborare automaticamente le strutture sintattiche.

2.1. Metodo

Partecipanti – Lo studio ha interessato un gruppo di 15 studenti di inglese dell'Università di Salerno. Il livello attestato di competenza linguistica dei partecipanti è elementare, anche se da un rapido esame delle composizioni degli studenti emergono alcune differenze che possono variare dal livello elementare al pre-intermedio. Al fine di confrontare i dati raccolti nelle prime fasi di acquisizione della L2 e quelli di fasi successive per osservare la possibile occorrenza degli stessi fenomeni, è stato costruito un secondo *corpus* di produzioni scritte da un gruppo eterogeneo di 15 studenti italiani provenienti da altre università (Napoli, Roma, Catania) ed impiegato nello studio come *corpus* di riferimento per il livello avanzato.

Materiali – I dati raccolti in entrambi i *corpora* sono simili in termini di mezzo impiegato (scrittura), genere (lettera, breve saggio argomentativo) e dominio (personale), anche se le composizioni mostrano differenze relativamente all'argomento[22], alla quantità[23] e, di conseguenza, all'occorrenza di sintagmi preposizionali.

Procedura – I partecipanti furono invitati a scrivere le loro composizioni a casa, ma senza il supporto di strumenti di riferimento (dizionari, grammatiche) e senza limitazione nel tempo di esecuzione del compito. Una volta raccolti, i dati sono stati annotati automaticamente applicando gli strumenti di analisi linguistica presenti sulla piattaforma VISL.

La TAB. 2 mostra le etichette che si riferiscono alle funzioni di argomento e aggiunzione riguardanti i costituenti preposizionali che in VISL vengono rappresentate attraverso la combinazione dell'etichetta relativa alla parte del discorso (POS) Preposizione (PRP) con altre categorie funzionali: nel caso specifico, l'etichetta PRP appare combinata con l'ag-

giunzione avverbiale ADVL (PRP@<ADVL), oppure accompagnata da un oggetto preposizionale con funzione di argomento PIV (PRP@<PIV) o ancora da un argomento riferito al soggetto di un verbo SA (PRP@<SA).

TABELLA 2
Annotazione delle funzioni di argomento e aggiunzione nel sistema VISL

POS	Categoria Funzionale	Definizione	Esempi
PRP	**@<ADVL**	Aggiunzione avverbiale (adjunct [free] adverbial)	you need to have jogging[you]PERS 2S/P NOM @SUBJ> need [need] <mv> V PR -3S @FS-<ACC to [to] INFM @INFM have [have] <mv> V INF @ICL-<ACC jogging [jog] <mv> V PCP1 @ICL-<ACC **for** [for] **PRP @<ADVL** half [half] DET S @>N an [a] ART S @>N hour [hour] N S NOM @P<
PRP	**@<PIV**	Oggetto preposizionale/ Argomento (prepositional object)	we [we] PERS 1P NOM @SUBJ> **asked** [ask] <mv> V IMPF @FS-STA **for** [for] **PRP @<PIV** a [a] ART S @>N new [new] ADJ POS @>N flight [flight] N S NOM @P< but [but] KC @CO
PRP	**@<SA**	Argomento (valency bound adverbial, referring to subject)	Andy [Andy] N S NOM @P< an [a] ART S @>N aspiring [aspiring] ADJ POS @>N journalist [journalist] N S NOM @N<PRED **who** [who] <rel> INDP S/P NOM @SUBJ> **move** [move] <mv> V PR -3S @FS-N< **to** [to] **PRP @<SA** New=York [New=York] N S NOM @P< to [to] INFM @INFM start [start] <mv> V INF @ICL-<ADVL her [she] PERS FEM 3S GEN @>N career [career] N S NOM @<ACC

Pertanto, i sintagmi preposizionali presenti nel *corpus* sono stati classificati (*mapped*) in argomenti e aggiunzioni nella maniera seguente:

a) Aggiunzione o "*free*" *PPs*: tutti i sintagmi preposizionali annotati con l'etichetta **@<ADVL**;
b) Argomento o "*bound*" *PPs:* tutti i sintagmi preposizionali annotati con le etichette **@<PIV, @<SA**.

Così, gli esempi (3) e (4) tratti dal *corpus* mostrano che le preposizioni che seguono il verbo *happen* sono state annotate rispettivamente come argomento e aggiunzione.

(3) <aux> V PR 3S @FS-N< happened [happen] <mv> **V** PCP2 AKT @ICL-AUX< to [to] **PRP** @<PIV you

(4) INDP S NOM @SUBJ> happens [happen] <mv> **V** PR 3S @FS-STA in [in] **PRP** @<ADVL just one [one] NUM S @>N day [day] N S NOM @P<.

I dati sono stati estratti dal *corpus* usando il programma di concordanze ConcApp fissando un determinato *frame* sintattico costituito da tutti i sintagmi preposizionali che nel *corpus* seguono i verbi (etichettati con *mv* = *main verb* dal *tagger*) transitivi e intransitivi e i sintagmi nominali (*n* = *noun*). Inoltre, al fine di classificare tutti i legamenti o sintagmi preposizionali presenti nel *corpus* annotato sono stati applicati due criteri diagnostici utilizzati nella ricerca linguistica per individuare gli argomenti di una determinata testa: la dipendenza dalla testa e l'opzionalità.

Dipendenza dalla testa – Gli argomenti dipendono dalle loro teste lessicali e sono parte integrante del sintagma, al contrario di quanto si verifica per le aggiunzioni. Queste ultime infatti, a differenza degli argomenti, possono occorrere liberamente con un'ampia gamma di teste diverse e contribuiscono alla corretta interpretazione di una determinata costruzione sintattica. Quindi il rapporto *token*/tipo, che è solitamente visto come indice di variazione lessicale, sarà presumibilmente più basso per i legamenti argomentali rispetto alle aggiunzioni che al contrario presenteranno un valore *token*/tipo più alto. La TAB. 3 mostra il numero di legamenti preposizionali con funzione di aggiunzione (**@<ADVL**) o di argomento (**@<PIV, @<SA**) su un totale di 1.230 sintagmi preposizionali presenti nel *corpus* (esclusa la preposizione infinitiva *to*), di cui n=628 legati a teste nominali e n=195 a teste verbali.

Le TABB. 4 e 5 mostrano alcuni esempi di fenomeni di deviazione nella selezione delle preposizioni o nella determinazione della struttura di sottocategorizzazione per un determinato lemma.

TABELLA 3
Numero di sintagmi preposizionali distribuiti per argomento e aggiunzione

POS	Functional Category	Frequency	% Token/Type Ratio
PRP	**@<ADVL**	925	2,4439
PRP	**@<PIV**	70	0,1629
PRP	**@<SA**	43	0,1001

TABELLA 4
Esempi di errore nell'uso di SP legati a verbi

[will] <aux> V IMPF @FS-<ACC ask [ask] <mv> **V** INF @ICL-AUX< to [to] **PRP** @<ADVL the [the] genius

@ICL-AUX< to [to] INFM @INFM ask [ask] <mv> **V** INF @ICL-<ACC to [to] **PRP** @<ADVL him [he]

@FS-N< @FS-STA benefit [benefit] <mv> **V** INF @ICL-AUX< of [of] <**prp**-stray> PRP @<ADVL

[scenery] N S NOM @P< composed [compose] <mv> **V** PCP2 PAS @ICL-N< by [by] **PRP** @<ADVL a [a] ART S sequence

[can] <aux> V PR @FS-N< derive [derive] <mv> **V** INF @ICL-AUX< by [by] **PRP** @<ADVL pride

V IMPF 1/3S @FS-STA realised [realise] <mv> **V** PCP2 PAS @ICL-AUX< from [from] **PRP** @<ADVL the [the] ART S/P @>N commune

PERS FEM 3S NOM @SUBJ> split [split] <mv> **V** IMPF @FS-STA with [with] **PRP** @<ADVL her [she] boyfriend

PERS FEM 3S NOM @SUBJ> suffers [suffer] <mv> **V** PR 3S @FS-ADVL> for [for] **PRP** @<ADVL a [a]

dreams [dream] N P NOM @<SC I [I] NUM @N< can [can] <aux> V PR @FS-STA think [think] <mv> **V** INF @ICL-AUX< about [about] <**prp**-stray> PRP

TABELLA 5
Esempi di errore nell'uso di SP legati al nome

my [I] PERS 1S GEN @>N balance [balance] **N** S NOM @<SC with [with] **PRP** @<ADVL what [what]<rel> INDP S/P @ACC> I [I] PERS 1S NOM @SUBJ> have

[and] KC @CO comprehension [comprehension] **N** S NOM @P< for [for] **PRP** @<ADVL a [a] ART S melting [melting] ADJ POS @>N pot

[terrorist] N S NOM @>N attacks [attack] **N** P NOM @P< at [at] **PRP** @N< the [the] ART S/P the [the] ART S/P @>N Twin=Towers

(*segue*)

TABELLA 5 (*segue*)

of [of] PRP @N< consequences [consequence] **N** P NOM @P< for [for] **PRP** @ADVL> this [this]

PERS 3P GEN @>N difficulties [difficulty] **N** P NOM @P< about [about] **PRP** @N< the [the]

popular [popular] ADJ POS @>N discos [disco] **N** P NOM @P< of [of] **PRP** @N< London [London] N

[important] ADJ POS @>N moment [moment] **N** S NOM @<SC of [of] **PRP** @N< people's life

had [have] <mv> V IMPF @FS-STA no [no] DET S @>**N** debt [debt] N S NOM @<ACC for [for] **PRP**

Opzionalità – I sintagmi preposizionali con funzione di argomento sono complementi obbligatori di una testa e la loro omissione genera costruzioni non grammaticali, mentre i sintagmi preposizionali con funzione di aggiunzione sono opzionali:

(5) happiness depends **on people's different characters** and on what they need to feel happy
(5') *happiness **depends**

(6) I have always lived **near the sea**
(6') *I have always lived

(7) Cristiano proposed to me to have jogging **with him at that time**
(7') Cristiano proposed to me to have jogging

Gli esempi (5)-(7) dimostrano che la differenza formale e funzionale tra aggiunzioni e argomenti è determinata dalla valenza e dal pattern semantico di un determinato verbo, pertanto si differenziano in base alla funzione che essi svolgono all'interno della frase, ma anche per il modo in cui vengono interpretati: per quanto riguarda la loro funzione, un argomento occupa la posizione assegnata dalla testa a cui è associato, mentre un'aggiunzione predica una proprietà separata dalla testa o dal sintagma a cui si riferisce; riguardo alla loro interpretazione, un complemento è un argomento quando la sua interpretazione rimane relativamente costante anche se associato a teste diverse[24].

3
Risultati e discussione

Uno degli elementi più discriminanti nel determinare la scelta e l'interpretazione di una preposizione è dunque il fattore semantico che dipende dalla classe di un verbo o di un nome e ciò emerge in maniera evidente nei dati degli apprendenti. Quando elabora una frase, l'apprendente ha bisogno di individuare le categorie sintattiche appropriate per i predicati semantici usati come input al fine di impostare correttamente i suoi parametri[25]. L'interpretazione semantica di una parola dipende da Principi Categoriali[26] che aiutano l'apprendente a determinare la forma sintattica e la struttura di categorizzazione di una data parola grazie/basandosi sul suo predicato semantico. Allo stesso modo nell'acquisizione di L2, la forma logica (o interpretazione semantica) associata ad un dato verbo si suppone possa favorire l'interpretazione e la produzione di verbi e le loro rispettive strutture di sottocategorizzazione con forme logiche simili. Così, ad esempio, nella frase:

(8) he talked to everyone directly in a straight way

il verbo *talk* (atto comunicativo) include nella sua struttura di sottocategorizzazione due argomenti, il SN soggetto *he*, e il SP *to everyone*, cioè (S\SN)/SP.

Il verbo *ask* appartiene allo stesso dominio semantico di *talk* e quindi nel processo di acquisizione della L2 gli viene assegnata la stessa struttura di sottocategorizzazione e la stessa forma logica.

(9) *I would like to ask to the genius

Questi dati indicano che la selezione delle preposizioni è in primo luogo operata in base a criteri semantici e quando non si verificano errori di selezione significa che la struttura di sottocategorizzazione di un verbo è stata interpretata correttamente e ciò spiega anche perché gli errori di selezione sono più frequenti nei sintagmi preposizionali legati ai nomi piuttosto che ai verbi. Pertanto, le strutture di sottocategorizzazione sono acquisite e sistematizzate nel sistema linguistico più velocemente di altre ed in particolare i verbi transitivi con due o più argomenti (SVO/SVO_iO_d) vengono acquisiti prima di quelli intransitivi (SV).

Per quanto riguarda i sostantivi, la principale area di errore nella selezione delle preposizioni riguarda il dominio semantico della "specifi-

cazione". Ad esempio, la preposizione it. *di* ricopre nell'interlingua di apprendenti di inglese come lingua straniera diverse aree di significato come quelle di agente (**a novel of the same author*) e tempo (**moment of people's life*). Al contrario la preposizione inglese *of* è in alcuni casi sostituita da *about* (**a difficulty about*) oppure da *for* (**a consequence for*, **comprehension for*). Ciò significa che partendo dalla forma logica associata ad un verbo, l'apprendente può decidere se un dato costituente è argomento del verbo e in tal caso dovrebbe essere inserito nella struttura di sottocategorizzazione del verbo oppure no.

Conclusioni

Le lingue "funzionano" in specifici contesti d'uso che a loro volta condizionano la costruzione degli enunciati di una lingua. Per valutare l'accettabilità di una determinata struttura nel suo contesto d'uso è necessario osservare il contesto linguistico in cui è stata prodotta ovvero l'intera struttura sintattica oltre che le sue singole componenti. La capacità di adattamento ad un determinato contesto è una caratteristica tipica del linguaggio umano, molto difficile da riprodurre con l'applicazione delle nuove tecnologie.

Tuttavia, al fine di consentire il corretto funzionamento dei programmi di analisi linguistica in un arco di tempo ragionevolmente breve, la maggior parte delle funzioni di controllo delle grammatiche di riferimento inserite nei database è *context-free*, vale a dire è stata concepita con scarsa considerazione del contesto linguistico, laddove sarebbe opportuno sviluppare software capaci di valutare la selezione di una parola o di una struttura in situazioni fortemente condizionate dal contesto sintattico. Infatti, in linea con l'ispirazione essenzialmente grammaticale e lessicale delle tassonomie di errore, la maggior parte dei sistemi di annotazione degli errori tende a basarsi sull'analisi dei singoli *tokens* e di conseguenza sulle specificazioni delle singole parti del discorso come principali unità di analisi linguistica.

Al contrario, l'impiego di sistemi di annotazione automatica progettati per l'analisi della L1 anche nell'annotazione di dati "interlinguistici" consente di formulare ipotesi sui processi di acquisizione di una lingua straniera, piuttosto che valutare ed eventualmente classificare come errore unicamente il "prodotto" di tale processo. Pertanto, i casi esaminati in questa rassegna possono offrire ulteriori sviluppi futuri nella ricerca sul NLP riguardante la progettazione di sistemi di annotazione *more com-*

prehensive e nell'ambito degli studi sul ruolo che la grammatica assume nel comportamento linguistico degli individui[27].

Note

1. I *Treebanks* vengono definiti come «linguistically interpreted *corpora* […] with structural information, centrally, the grammatical structure of the samples, though some resources include categories of information other than "grammar" sensu stricto». G. Sampson, A. Babarczy, *Limits to Annotation Precision*, in A. Copestake, J. Hajiç (eds.), *Proceedings of the 4th International Workshop on Linguistically Interpreted Corpora*, 2003, pp. 61-8.

2. Questo tipo di annotazione veicola un insieme di informazioni (metadati o *data about data*) sulle diverse componenti di un *corpus*, quali ad esempio tipo di testo (argomentativo, espositivo o altro per dati scritti o variabili sociolinguistiche per dati di parlato) e la sua struttura testuale, ad esempio paragrafo, frase, clausola. «The encoding, referred to as annotation or tagging, added to the texts that comprise a *corpus*, is a metalanguage that is generally done in some form of markup language. Two commonly used markup languages in the *corpora* surveyed in this review are XML and SGML. The Extensible Markup Language (XML) is the universal format for presenting structured documents and data on the World Wide Web. The functionality of the Web is improved through XML's design because it provides more flexible and adaptable information identification. It is called extensible because it is not a fixed format like HTML (hyper-text markup language), which is a single, pre-defined markup language». N. A. Pravec, *Survey of Learner Corpora*, in "ICAME Journal", 26, 2002, pp. 81-114.

3. J. Foster, *Real Bad Grammar: Realistic Grammatical Description with Grammaticality*, in "Corpus Linguistics and Linguistic Theory", 3/1, 2007, pp. 73-86.

4. Ivi, p. 74.

5. N. Chomsky, *Formal Discussion. The Development of Grammar in Child Language*, reprinted in J. P. B. Allen, P. van Buren (eds.), *Chomsky: Selected Readings*, Oxford University Press, Oxford 1971, pp. 129-34.

6. Ivi, p. 130.

7. La costruzione del *corpus* fa parte di un progetto in corso presso l'Università di Salerno coordinato da Bruna Di Sabato riguardante la raccolta di dati di interlingua scritta con lo scopo di verificare l'eventuale incidenza della variabile diatopica sulle performance degli studenti in esame e individuare eventuali somiglianze e/o differenze a parità di livello di competenza linguistica e task assegnato.

8. Ad eccezione del sistema di annotazione UCLEE, gli altri sistemi presi in esame sono stati progettati per l'analisi di produzioni linguistiche "grammaticali" di parlanti nativi di L1 inglese e non prevedono quindi in principio l'analisi computerizzata di testi "non-grammaticali" prodotti da parlanti nativi o non nativi. Tuttavia, come verrà dimostrato nel paragrafo relativo all'analisi su campione, il/i sistema/i che include restrizioni di selezione e/o dipendenza dalla struttura della testa (V, P, o N) sarà in grado di individuare le violazioni delle restrizioni suddette e consentirà quindi di individuare/recuperare l'informazione sull'errore presente nella frase.

9. T. McEnery, R. Xiao, Y. Tono, *Corpus-Based Language Studies. An Advanced Resource Book*, Routledge, London-New York 2006, p. 30.

10. Ivi, p. 32.

11. Per citarne alcuni: A. Babarczy, J. Carroll, G. R. Sampson, *Definitional, Personal, and Mechanical Constraints on Part of Speech Annotation Performance*, in "Journal of Natural Language Engineering", 11, 2006, pp. 11-4; G. R. Sampson, *A Proposal for Improving the Measurement of Parse Accuracy*, in "International Journal of Corpus Linguistics", 5, 2000, pp. 53-68; A.Voutilainen, *An Experiment on the Upper Bound of Interjudge Agreement: the Case of Tagging*, in "Proceedings of the 9th Conference of EACL", Bergen 1999, pp. 204-8.

12. R. Quirk, S. Greenbaum, G. Leech, J. Svartvik, *A Comprehensive Grammar of the English Language*, Longman, London 1985.

13. S. Granger, *Use of Tenses by Advanced EFL Learners: Evidence from an Error-tagged Computer Corpus*, in H. Hasselgard, S. Oksefjell (eds.), *Out of Corpora*, Rodopi, Amsterdam 1999, pp. 191-202.

14. N. Nesselhauf, *Collocations in a Learner Corpus. Studies in Corpus Linguistics*, J. Benjamins, Amsterdam 2005.

15. R. De Felice, S. G. Pulman, *Automatically Acquiring Models of Preposition Use*, in *Proceedings of the 4th ACL-SIGSEM*, Prague 2007, pp. 45-50. Questo studio si differenzia dai precedenti nel tipo di approccio adottato, completamente computerizzato: dato un determinato contesto semantico e sintattico è automaticamente possibile prevedere quale preposizione può occorrere in quel determinato contesto che è rappresentato da un vettore contenente un certo numero di tratti. I vettori sono rielaborati da un algoritmo che consente di stabilire delle associazioni tra determinati contesti e le corrispondenti preposizioni.

16. Si vedano, oltre al già citato UCLEE, il sistema elaborato nell'ambito del *Japanese Learner Corpus Project*, quello del *Cambridge Learner Corpus* e del progetto *Free Text* per il francese.

17. A. Diaz-Negrillo, J. Fernandez-Dominguez, *Error Tagging Systems for Learner Corpora*, in "RESLA", 19, 2006, pp. 83-102.

18. R. Ellis, *The Study of Second Language Acquisition*, Oxford University Press, Oxford 1994.

19. A. Díaz-Negrillo, D. Meurers, S. Valera, H. Wunsch, *Towards Interlanguage POS Annotation for Effective Learner Corpora in SLA and FLT*, in "Language Forum", Special Issue on New Trends in Language Teaching, C. Pérez Basanta, 2010; M. Dickinson, M. Ragheb, *Dependency Annotation for Learner Corpora*, in *Proceedings of the Eighth Workshop on Treebanks and Linguistic Theories (TLT-8)*, Milan 2009.

20. Disponibile all'indirizzo http://beta.visl.sdu.dk/http://beta.visl.sdu.dk/.

21. Il sistema prevede anche un'interfaccia didattica con diverse attività di tutoring.

22. I temi erano:
- *What is the best book/film you have recently read/seen? Describe its plot and tell what you liked about it.*
- *What makes you feel good and why.*
- *What is the strangest thing that has happened to you while travelling.*
- *What is your earliest childhood memory?*

23. I dati quantitativi non sono considerati rilevanti ai fini del presente studio, dal momento che lo scopo principale è analizzare il tipo/funzione di sintagma preposizionale usato dagli apprendenti.

24. Per la letteratura sull'argomento si veda R. Jackendoff, *X-Syntax: a Study of Phrase Structure*, The MIT Press, Cambridge (MA) 1977; A. Marantz, *On the Nature of Grammatical Relations*, The MIT Press, Cambridge (MA) 1984; C. Pollard, I. A. Sag, *Information-Based Syntax and Semantics*, 1, Center for the Study of Language and Information, Stanford (CA) 1987; J. Grimshaw, *Argument Structure*, The MIT Press, Cambridge 1990; P. Merlo, E. Esteve Ferrer, *The Notion of Argument in PP Attachment*, in "Computational Linguistics", 2006, XXXII, pp. 341-77.

25. A.Villavicencio, *Learning to Distinguish PP Arguments from Adjuncts*, in *Proceedings of the 6th Conference on Natural Language Learning*, Association for Computational Linguistics, Morristown 2002, pp. 1-7.

26. M. Steedman, *The Syntactic Process*, The MIT Press, Cambridge (MA) 2000.

27. G. Sampson, A. Babarczy, *Definitional and Human Constraints on Structural Annotation of English*, in "Natural Language Engineering", 14, 2008, pp. 471-94.