Parsing di corpora di apprendenti di italiano: un primo studio su VALICO

Elisa Corino[1]
Università di Torino
elisa.corino@unito.it

Claudio Russo
Università di Torino
clrusso@unito.it

Abstract

English. Modern learner corpora are now routinely PoS tagged, whereas syntactic parsing is much less frequent. This paper proposes a first attempt of parsing applied to a subcorpus of VALICO, in an effort to identify key elements to be further used to parse corpora of Italian as a foreign language in a proper way.

Italiano. *I moderni corpora di apprendenti sono ormai abitualmente annotati per PoS. Meno frequente è invece il parsing sintattico delle varietà di apprendimento. Questo contributo propone un primo tentativo di parsing applicato a un sottocorpus di VALICO, nel tentativo di individuare elementi chiave che possano servire a tracciare una rotta per l'etichettatura sintattica di corpora di italiano come lingua straniera/seconda (LS).*

1 Introduzione

È ormai prassi lemmatizzare e annotare i corpora di apprendenti per Parts of Speech (PoS) e molti sono gli esempi di etichettatura degli errori; non altrettanto diffuse sono invece le esperienze di parsing e annotazioni sintattiche di learner corpora.

Le ragioni sono molteplici, ma certamente riconducibili all'estrema imprevedibilità e marcatezza che caratterizza l'interlingua degli apprendenti. Nonostante la complessità dell'operazione, che richiede necessariamente una buona dose di intervento manuale, il parsing di un *learner corpus* può avere ricadute positive in più di un ambito: oltre al contributo per i contesti che si occupano in modo precipuo di linguistica computazionale, esso diventa estremamente utile per le ricerche su acquisizione e apprendimento, perché permette di individuare più facilmente errori, deviazioni dalla norma e distribuzione di categorie e strutture sintattiche altrimenti difficili da far emergere interrogando un corpus etichettato unicamente per parti del discorso.

Applicare processi di parsing alle varietà di apprendimento non è certo una novità: il trattamento sintattico delle interlingue degli apprendenti è da tempo il focus di sistemi di Computer Assisted Language Learning (CALL) e Intelligent Language Tutoring Systems (ILTS) (Vandeventer Faltin, 2003; Heift & Nicholson, 2001; Amaral & Meurers, 2011). In questi casi la violazione delle regole sintattiche, come ad esempio il mancato accordo tra soggetto e verbo, diventa spia della presenza di un errore da segnalare a chi produce il testo. Menzel & Schröder (1999) usano parametri che descrivono le dipendenze all'interno delle lingue degli apprendenti, alle quali vengono applicati vincoli pesati e procedimenti di *robust parsing* per definire la deviazione dalla norma.

La distanza da strutture canoniche e non marcate va tuttavia definita a partire da uno standard al quale le varietà di apprendimento vengono ricondotte. Ecco perché il parser del *learner corpus* deve essere preliminarmente applicato ad un corpus di nativi – possibilmente comparabile a quello degli apprendenti. Una prassi

[1] Il contributo è il risultato della collaborazione dei due autori; tuttavia a Elisa Corino vanno attribuiti i §§ 1, 2, 4 e 5; a Claudio Russo il § 3. Si ringraziano Cristina Bosco e Alessandro Mazzei per il prezioso aiuto e per aver messo a nostra disposizione gli strumenti elaborati dal loro gruppo di ricerca.

diffusa è poi quella di riportare le occorrenze errate ad una forma target, un procedimento manuale in genere usato da quei learner corpora annotati anche per errori. Alcuni tentativi sono stati fatti fino ad ora soprattutto per il tedesco (Nivre et al., 2007, Lüdeling 2008, Ott & Zai 2010), ma il terreno pare ancora inesplorato per quanto riguarda il panorama italiano.

Questo contributo si propone come un primo tentativo di applicare procedure di parsing a un corpus di apprendenti di italiano come lingua straniera, per definirne criticità e vantaggi sia dal punto di vista computazionale, sia rispetto allo studio delle varietà di apprendimento.

In particolare verranno presi in considerazione alcuni dati estratti dal corpus VALICO[2], selezionati a partire da sottocorpora definiti in base alla L1 degli apprendenti. Per questo studio pilota si è scelto di comparare i risultati del trattamenti di testi di apprendenti ispanofoni e germanofoni, con l'intento di individuare peculiarità sintattiche delle interlingue di discenti provenienti da lingue tipologicamente diverse: tipicamente romanza – e quindi vicina all'italiano l'una, rappresentativa dell'area germanica l'altra. Sono stati sottoposti al parser 12 testi (quattro per ciascuna delle prime tre annualità di studio di italiano) estratti in modo casuale dal sottocorpus germanofono e 12 testi derivati dal sottocorpus ispanofono, secondo gli stessi criteri di selezione. Per il gruppo tedesco sono state processate in totale 126 frasi, per il gruppo spagnolo 78.

Fine ultimo è stabilire quali deviazioni dalla norma dell'uno e dell'altro gruppo non sono etichettate correttamente, in modo da tracciare un percorso che possa portare alla definizione di regole di etichettatura per allenare il parser e migliorarne le prestazioni su questa varietà di lingua.

VINCA, il corpus di nativi appaiato a VALICO, servirà da corpus di riferimento per il parser a dipendenze sviluppato tra gli strumenti del Turin University Linguistic Environment (TULE).

2 Il corpus VALICO

VALICO è un corpus di scritti di apprendenti di italiano LS liberamente consultabile online (Marello et al. 2011, Marello & Corino i.s.), etichettato per PoS con il TreeTagger dell'IMS di Stoccarda e codificato in CQP (Heid 2007, 2009).

L'architettura del corpus permette di applicare le query a sottosezioni selezionate in base ai metadati contenuti nella *Header*, inseriti in una base dati (Colombo i.s.), tra questi la L1 degli apprendenti, l'annualità di studio di italiano, il luogo di produzione del testo, le altre LS conosciute.

Gli stimoli iconici dai quali è stato elicitato VALICO sono le stesse vignette somministrate agli autori di VINCA[3]; i due corpora sono quindi appaiati e comparabili per lessico, strutture sintattiche, organizzazione testuale (Corino & Marello 2009), si veda ad esempio la somiglianza tra (1a) e (1b) in relazione alla Fig. 1:

Fig. 1

(1a) […] la donna , arrabbiata , impreca contro di me dicendomi che quello che avevo appena picchiato è il suo ragazzo ! [VINCA]

(1b) Ha cominciato a sgridarme per farle una brutta passata e quello che riteneva molto grave era che avevo picchiato al suo ragazzo. [VALICO, L1 Spagnolo]

VINCA si presta quindi a fungere da corpus di controllo e modello per il parsing di VALICO.

In questa prima fase esplorativa delle modalità di etichettatura sintattica del corpus, è stato deciso di non ricorrere a ipotesi esplicite sul target delle produzioni degli apprendenti, così come avviene invece per altri learner corpora (ad es. FALKO, Lüdeling et al 2012). L'unica operazione di riconduzione a forme standard è stata fatta per identificare PoS e lemma di lessemi scorretti dal punto di vista ortografico, forme "creative" sia dal punto di vista lessicale che morfologico.

Già le forme segnalate dal TreeTagger come *unknnown* erano state corrette manualmente e ricondotte a PoS e lemma appropriato in relazione anche al co-testo nel quale esse occorrevano. Infatti è possibile interrogare VALICO, nella sua attuale forma, per lemmi e ottenere anche quelle occorrenze che morfologicamente

[2] Varietà di Apprendimento Lingua Italiana Corpus Online, liberamente consultabile all'indirizzo www.valico.org.

[3] www.valico.org/vignette

non rispondono alla norma linguistica standard (ad es. cercando il lemma *minacciare* otteniamo sì forme quali *minacciano*, *minacciato*, *minaccia*, ma anche *minaccava* e *minaciò*).

3 Dal PoS tagging al parsing

Il lavoro di disambiguazione e correzione degli *unknown* rilevati dal TreeTagger è tutt'ora in corso, anche se buona parte di essi sono già stati corretti e ad oggi tokenizzazione, PoS tagging e lemmatizzazione del corpus sono piuttosto affidabili.

Per questo primo tentativo di parsing del corpus si è tuttavia deciso di affidarsi – anche per la fase di PoS tagging - agli strumenti del TULE: un analizzatore morfologico, un tokenizzatore e un parser a dipendenze: tale scelta è giustificata dalla natura dei moduli del TULE, che, in quanto costituiti da regole di disambiguazione, non si avvalgono di addestramneti stocastici; in tale cornice, la preferenza del POS-tagger di TULE all'etichettatura già presente in VALICO ha permesso al paser di lavorare in condizioni ottimali, massimizzandone la precisione. Un'indagine a campione ha fatto emergere errori simili per entrambi gli strumenti, si veda ad esempio l'etichettatura della forma *dona* nella frase *è andato in freta ha ragiungere la sua dona è poi a preso questo uomo le ha fato male* (= donna, NOME), impropriamente riconosciuta come imperativo verbale da entrambi

dona (DONARE VERB MAIN IND PRES TRANS 3 SING) TULE
dona/VER:impe/donare è/VER:pres/essere TreeTagger

L'integrazione dei materiali già elaborati per il corpus e gli strumenti di parsing sarà necessariamente il prossimo passo nello sviluppo di questa ricerca ed eviterà di dover intervenire anche su forme che già sono state corrette nel processamento del corpus con il TreeTagger, si veda ad esempio

fidansato (FIDANSATO / NOUN COMMON M SING) VALICO TULE
fidansato (/NOM/fidanzato) VALICO 2016

Eliminerà inoltre a priori alcuni errori di annotazione sintattica e impedirà la generazione di errori dovuti al mancato riconoscimento delle forme "unknown" che bloccano la sequenza di processamento dei dati.

Ai fini del presente studio, è interessante sottolineare poi che il tokenizzatore qui utilizzato si basa su un automa deterministico a stati finiti che risulta robusto su sequenze di caratteri appartenenti a diverse lingue (tra cui rientrano inglese, spagnolo, hindi e italiano). Tale strumento compie una prima distinzione fra parole in generale (qui intese come sequenze di caratteri alfabetici), nomi propri, abbreviazioni/sigle, numeri in cifre, date in formati standard, segni di interpunzione, numerazioni di capitoli e paragrafi, anni (nelle forme contratte come '05 per 2005). Nella fase successiva di trattamento, il parsing si basa su una gerarchia precompilata di classi trasformate per generare gli alberi sintattici[4].

3.1 I dati di imprevedibilità morfosintattica

Le peculiarità dei dati linguistici sottoposti al parser hanno, come previsto, originato una serie di errori che hanno bloccato la sequenza di processamento in più momenti. I problemi non nascono soltanto in relazione alla flessione verbale o nominale, ma emergono anche laddove occorre un accumulo di clitici unitamente a forme ortografiche che deviano dalla norma. Si veda ad esempio la frase

(2) […], ma la ragazza bella è stata averecela.

Il verbo *avere* cliticizzato ha reso impossibile la disambiguazione e il tagging, anche a seguito della normalizzazione ortografica e di alcune progressive semplificazioni. È stato quindi necessario sostituire il token direttamente con il suo lemma, per permettere al TULE di procedere con il PoS tagging e, successivamente, con il parsing sintattico.

Fortunatamente, nei sottocorpora di VALICO qui considerati, tali errori si sono rivelati in quantità limitata: oltre al caso di *avercela*, sono state normalizzate in *È* tutte le istanze di *E'* e una istanza di *glielo* è stata normalizzata in *lui*[5], in un atto di correzione purtroppo invasivo e in futuro auspicabilmente evitabile.

[4] Si rimanda Lesmo (2009 e 2011) per la trattazione approfondita dell'implementazione del TULE.

[5] La frase originale recitava: "Il fratellino di Leo non capiva perchè e così lei ha spiegato glielo."

4 Prime osservazioni linguistiche: il gerundio e le preposizioni

Una prima verifica sull'etichettatura dei testi fa emergere due ordini ricorrenti di problemi che paiono essere indipendenti dalla L1 degli apprendenti: il primo è legato ad un uso peculiare del gerundio, assimilabile alle funzioni attributive delle participiali inglesi (fonte di interferenza per gli ispanofoni) e tedesche; l'altro dipende dalla presenza di preposizioni in posizione postverbale laddove invece il target richiede l'oggetto diretto.

L'uso anomalo del gerundio, come esemplificato da (2) e (3), pare ricalcato sulla struttura attributiva tipicamente inglese (participiale con la forma in *-ing*), ma possibile anche in tedesco, dove il participio specifica la condizione dell'antecedente dal quale dipende.

(2) Ieri al parco mio fratello stava leggendo il giornale quando ha visto un uomo portando una donna sopra i suoi ombrelli [L1 Spagnolo]

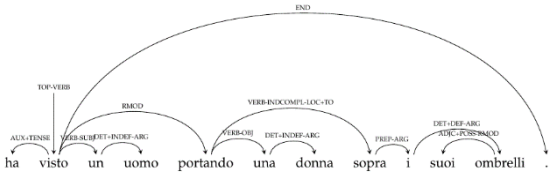

Fig. 2 Rappresentazione del segmento *visto un uomo portando una donna sopra i suoi ombrelli*

(3) Ieri al parco Giacomo è stato seduto sopra un banco nel parco e ha letto il giornale dello sport quando ha visto un uomo portando una donna che ha gridato molto forte. [L1 Tedesco]

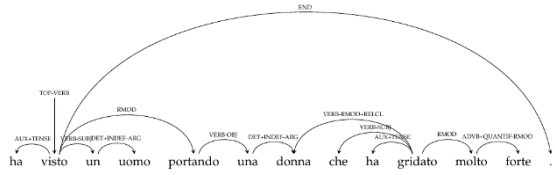

Fig. 3 Rappresentazione del segmento *quando ha visto un uomo portando una donna che ha gridato*

Come si osserva nelle Figg. 2-3[6], in entrambi casi il gerundio fa parte di una proposizione circostanziale (RMOD) al quale è attribuito un soggetto non espresso dalla struttura superficiale. Etichettare una serie di istanze verbali ravvicinate come entità circostanziali è un risultato che certo non sorprende; in casi come questi, tuttavia, il gerundio è il risultato di un'interferenza molto frequente nel corpus e non codifica la struttura rilevata dal parser.

La forma che meglio rispecchia in italiano la sfumatura attributiva dell'originale che ha causato il transfer è la frase relativa, pur considerando anche l'infinitiva come un'opzione possibile.

In VINCA effettivamente troviamo numerose occorrenze che descrivono la stessa situazione (si veda ad es. (4)), codificata in una relativa, e che come tali sono correttamente segnalate (VERBO-RMOD+RELCL, Fig. 3)

(4) Ieri al parco stavo leggendo il giornale, seduto su una panchina quando vedo passare davanti a me un energumeno che trascina a forza una donna sulle spalle. [VINCA]

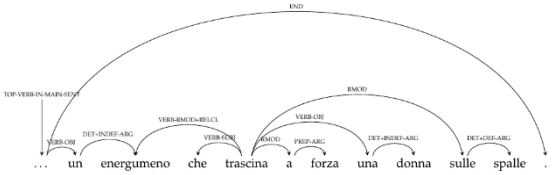

Fig. 4 Rappresentazione della frase relativa *un energumeno che trascina a forza una donna sulle spalle*

Una seconda importante causa di errore di etichettatura sintattica è la sovraestensione dell'uso delle preposizioni anche a quei casi in cui il verbo richiede l'oggetto diretto. (5) ne è un esempio emblematico:

(5) Per quello, il ragazzo di occhiali (che stava a un banco del parco) va golpeare al brutto ragazzo per liberare a la ragazza di bricci dil ragazzo che la portava. [L1 Spagnolo]

In casi come *va a golpeare al brutto ragazzo*, il parser interpreta come VERB-INDCOMP-LOC+TO il sintagma preposizionale, che invece dovrebbe essere ricondotto a un complemento diretto

[6] Le rappresentazioni degli alberi sintattici sono state generate dal software viewerTULE, implementato da L. Robaldo.

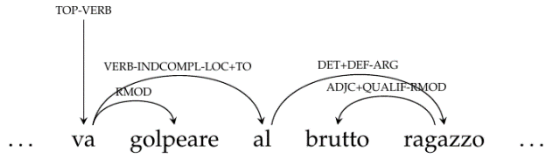

Fig. 5 Rappresentazione del segmento *va a golpeare al brutto ragazzo*

Tuttavia non si tratta di un comportamento consistente, poiché emergono anche casi come (6), in cui invece l'etichettatura è corretta (Fig. 6), nonostante l'agrammaticalità dell'enunciato.

(6) Lui stava contentissimo perche pensava che cossì aveva aiutato a la povera donna. [L1 Spagnolo]

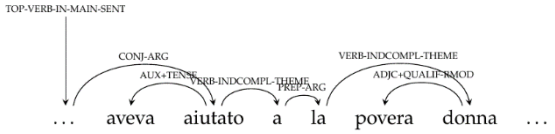

Fig. 6 Rappresentazione del segmento *aveva aiutato a la povera donna*

L'identificazione corretta della relazione sintattica sembra essere dovuta alla semantica del verbo e alla capacità del TULE di lemmatizzare correttamente la forma.

5 Conclusioni provvisorie e futuri sviluppi

L'esperienza maturata in questa prima fase del percorso di ricerca e implementazione del parsing sul corpus di apprendenti VALICO ha permesso di mettere in luce vantaggi e criticità degli strumenti utilizzati.

Dal punto di vista computazionale è emerso come sia necessario integrare risorse e definire "protocolli di allenamento" del parser orientati al trattamento di quella varietà particolare di lingua che è l'interlingua di apprendenti.

Una valutazione qualitativa ha mostrato che errori causati dall'omissione di elementi sintattici chiave per la definizione delle dipendenze porta a errori nella stessa etichettatura, così come avviene in occasione della sovraestensione di alcune forme a funzioni che esse non rivestono, o ancora della sinergia negativa tra forme morfologicamente complesse ed errori ortografici; altri errori invece sono trattati in modo aproblematico.

Le difficoltà di gestione dei sintagmi preposizionali e fenomeni diffusi di transfer linguistico soprattutto a livello di sistema verbale rivelano come l'individuazione di tendenze e la definizione di alcune indicazioni sul trattamento di tali costruzioni potrebbero incrementare notevolmente la precisione dell'etichettatura automatica.

Rispetto alla definizione di un'architettura che integri il parsing sintattico al PoS tagging, è necessario elaborare una chiara proposta di definizione di regole di etichettatura; inoltre resta da stabilire se sia veramente necessario allinearsi alle scelte operate per altri *learner corpora* in cui viene definita un'ipotesi target che segna la struttura obiettivo della produzione dell'apprendente e ne mette in luce la distanza dalla versione effettivamente riportata nel corpus.

Bibliografia

Amaral, L.; Meurers, D.(2011). On Using Intelligent Computer-Assisted Language Learning in Real-Life Foreign Language Teaching and Learning. ReCALL 23(1), 4–24. URL http://purl.org/dm/papers/amaral-meurers-10.html.

Colombo, S. (in stampa). Storia dell'architettura di VALICO. In E. Corino, C. Marello, VALICO e VINCA: corpora di apprendenti di italiano, Guerra, Perugia.

Corino, E.; Marello, C. (2009). Elicitare scritti a Partire da storie disegnate: il corpus di apprendenti Valico. In C. Andorno, S. Rastelli (a cura di), Corpora di Italiano L2: Tecnologie, metodi, spunti teorici, Perugia: Guerra.

Dickinson, M.; Ragheb, M. (2009). Dependency Annotation for Learner Corpora. In Proceedings of the Eighth Workshop on Treebanks and Linguistic Theories (TLT-8).

Heid, U. (2007). Il corpus WorkBench come strumento per la linguistica dei corpora. Principi ed applicazioni' in M. Barbera, E. Corino, C. Onesti (a cura di) (2007), *Corpora e linguistica in rete*. Guerra, Perugia, pp. 89-108.

Heid, U. (2009). Metadata for learner corpora. A case study on VALICO. In E. Corino, C. Marello (a cura di), VALICO. Studi di linguistica e didadttica. Perugia: Guerra, 151-165.

Heift, T; Nicholson, D. (2001). Web Delivery of Adaptive and Interactive Language Tutoring. *International Journal of Artificial Intelligence in Education* 12(4), 310–325

Lesmo, L. (2009). The Turin University Parser at Evalita 2009. In: Proceedings of Evalita '09, Reggio Emilia, Italy.

Lesmo, L. (2011). Use of semantic information in a syntactic dependency parser. In Magnini B., Cutugno F., Falcone M., Pianta E. (eds) "Evaluation of Natural Language and Speech Tools for Italian - Proceedings of Evalita 2011". LNCS/LNAI, Springer-Verlag

Lüdeling, A. et al. (2008). Syntactic Misuse, Overuse and Underuse: A Study of a Parsed Learner Corpus and its Target Hypothesis

Lüdeling, A. et al. (2012). Das Falko-Handbuch Korpusaufbau und Annotationen Version 2.01

Marello, C. et al (2011). I corpora VALICO e VINCA: stranieri e italiani alle prese con le stesse attività scritte. In La Piazza delle lingue L'italiano degli altri. Firenze, 27-31 maggio 2010. Atti, a cura di Nicoletta Maraschio e Domenico De Martino, Firenze, Accademia della Crusca, 2011 ("La Piazza delle lingue", 2). pp.49-61

Menzel, W.; Schröder, I. (1999). Error Diagnosis for Language Learning Systems [http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.34.4723&rep=rep1&type=pdf]

Nivre et al., (2007). MaltParser: A Language-Independent System for Data-Driven Dependency Parsing. *Natural Language Engineering* 13(1), 1–41.

Ott, N.; Ziai, R. (2010). Evaluating Dependency Parsing Performance on German Learner Language. In Proceedings of the Ninth Workshop on Treebanks and Linguistic Theories (TLT-9), Tartu.

Vandeventer Faltin, A. (2003). Natural language processing tools for computer assisted language learning. In Linguistik online 17, 15/03 [http://www.linguistik-online.de/17_03/vandeventer.html]