L'annotazione morfosintattica del *Padua Corpus*: strategie adottate e problemi di acquisizione

di

Manuel Barbera e Carla Marello

1. Che cosa si annota?

Il *Padua Corpus* è un sottoinsieme della raccolta di testi del *TLIO, Tesoro della lingua italiana delle origini* messo a disposizione da Pietro Beltrami e selezionato da Lorenzo Renzi e Giampaolo Salvi (cfr. Renzi 1998a, p. 29) come base per la compilazione di *ITALANT, Grammatica dell'Italiano Antico.*

I criteri di selezione dei testi sono spiegati in Renzi (1998a, pp. 29-30) e li riassumiamo qui per sommi capi. Una prima limitazione restringe il corpus diacronicamente alla sola seconda metà del Duecento e diatopicamente alla sola varietà fiorentina; all'interno, poi, del macrocorpus dei testi fiorentini datati tra il 1251 ed il 1300 del *TLIO*, la scelta è stata ulteriormente ristretta per ragioni di maneggevolezza e praticità, «eliminando i frammenti troppo brevi per poter essere utili, certi testi ripetitivi (conti di banchieri), altri più difficili e ardui (specialmente nella lirica), (...) le versioni rispettivamente dal latino e dal francese» (Renzi 1998a, p. 29)[1].

La decisione di prendere il fiorentino antico come unica base per una grammatica dell'italiano antico[2] è stata motivata da Lorenzo Renzi, coordinatore del progetto, con due tipi di considerazioni.

Una è d'ordine sincronico e di grammatica descrittiva: «non si possono descrivere contemporaneamente diverse lingue senza violare il presupposto stesso della descrizione, che è che questa si eserciti su uno stato sincronico di lingua. Ogni sistema va descritto in sé e per sé. Questa discri-

minante teorica ha un'ovvia appendice pratica: un corpus che comprendesse tutte le varietà antiche italiane sarebbe così vasto da essere ingovernabile» (Renzi 1998a, pp. 22-23).

L'altra è invece d'ordine diacronico e di grammatica storica: «l'italiano antico, cioè la fase antica della lingua che parliamo oggi in Italia come lingua comune, è il fiorentino antico» (Renzi 1998a, p. 23), come ha ampiamente dimostrato la tradizione tutta della grammatica storica romanza ed italiana, a partire dal profilo ascoliano dell'*Italia dialettale* (Ascoli 1882-1885), per arrivare a Castellani e Tagliavini;[3] la dimostrazione, dapprima condotta principalmente in base a tratti fonetici e morfologici, si è col tempo arricchita, ad opera soprattutto di Renzi, anche di considerazioni sintattiche[4].

Certo non vi possono esser dubbi che dal punto di vista di una tassonomia puramente genealogica delle lingue romanze, (a) l'italiano contemporaneo sia una varietà toscana a base fiorentina, alla stessa maniera che il francese moderno è una varietà oitanica a base franciana (parigina); e (b) che il sistema linguistico dei dialetti toscani (considerato ed individuato nel suo complesso, ancora una volta ascolianamente, alla stessa stregua, ad esempio, del francoprovenzale) sia una varietà romanza autonoma rispetto ad esempio al siciliano, al lombardo («galloitalico») od al veneto, per non menzionare che le varietà «italoromanze» medievalmente più importanti, e che pertanto esiga una descrizione indipendente.

Pur ciò stabilito, la realtà dei fatti è, come spesso accade, più sfumata e complessa. Non si possono, infatti, trascurare i rapporti tra il fiorentino e le altre varietà toscane, né quelli tra le varietà letterarie toscane[5] e le altre varietà letterarie, all'epoca più antiche ed illustri, del medioevo romanzo: l'oitanico, l'occitanico ed il siciliano.

Crediamo infatti che astrarre il fiorentino dugentesco (ossia, nella nostra prospettiva, i *testi* selezionati di fidata origine e datazione) dalla rete di ricchi rapporti che esso intrattiene con le altre varietà extramunicipali sia in qualche modo impoverirlo e rischiare di renderlo, almeno in singoli punti, incomprensibile. Questa istanza non deve necessariamente intendersi come un mero richiamo filologico alla restituzione ai testi del concerto intertestuale cui appartennero[6], ma anche, e più linguisticamente, come un invito a considerare il toscano come un diasistema al cui interno il fiorentino sarebbe, inizialmente, una mera varietà diacorica che va progressivamente smunicipalizzandosi ed assumendo il ruolo ed il prestigio di forma letteraria[7]. Analoghe considerazioni (solo, questa volta, contrastive anziché diasistemiche) valgono per i rapporti tra l'italiano (toscano e fiorentino) e le altre lingue romanze medievali che lo hanno

intimamente sostanziato (siciliano) o che ne sono state i grandi modelli culturali (occitanico e soprattutto oitanico): poche grammatiche, per fare un esempio illustre, hanno illuminato l'orizzonte linguistico e culturale greco-romano con altrettanta penetrazione della *Grammaire comparée des langues classiques* del Meillet. Una parziale riduzione di questi inconvenienti[8] potrebbe forse già ottenersi anche solo con una maggiore attenzione ai volgarizzamenti, specie dal francese e dal latino, ed ai testi fiorentini «ibridi» (fiorentini in tradizione allodialettale o viceversa).

Naturalmente le osservazioni sopra accennate non vanno intese come una critica al progetto di *ITALANT*, o come un tentativo di sminuirne la portata storica, ma solo come una ulteriore messa a punto, in funzione del nostro progetto coordinato, ma anche in vista di sviluppi futuri. Ci rendiamo ben conto che, secondo recita il proverbio, «il meglio è nemico del bene» e che, visto che da qualche parte bisognava pur incominciare, la strategia adottata da Renzi è forse l'unica teoricamente ragionevole e praticamente realizzabile in tempi accettabili.

Ai fini di chi deve annotare, inoltre, la relativa omogeneità del corpus rappresenta un indubitabile vantaggio: il fatto che la poesia delle origini inclusa nel *Padua Corpus* «non presenti differenze sistematiche dalla prosa» (Renzi 1998a, p. 29) viene assunto come presupposto per non separare almeno inizialmente le parti poetiche da quelle in prosa. Tuttavia, come si potrà meglio constatare dalla successiva descrizione delle procedure, una volta che il corpus sarà annotato, gli studiosi potranno servirsene efficacemente per verificare e validare o modificare tale assunto[9].

Non solo: poiché fra i lavori preliminari all'annotazione morfosintattica c'è la preparazione di liste di forme ricondotte a lemmi, chi volesse in séguito fare grammatiche di piemontese o di siciliano o di altre varietà italoromanze antiche potrà comunque servirsene come punto di partenza, così come potrà anche servirsi di *ITALANT* e delle sequenze sintattiche più frequenti emerse dalle procedure intermedie e finali di annotazione del *Padua Corpus* come base per la formulazione di regole diverse.

2. Perché annotare morfosintatticamente il *Padua Corpus*?

L'utilità per il linguista d'avere a disposizione corpora elettronici annotati morfosintatticamente è ovvia: poter interrogare per sequenze di categorie grammaticali (ad es. estrarre tutte le sequenze Articolo + Nome + Aggettivo o tutte le sequenze Articolo + Nome + Congiunzione + Articolo + Nome + Preposizione + Articolo + Nome) è molto più interessante per chi si occupa di sintassi che non l'interrogare per sequenze di forme, sia pure ricondotte a lemmi.

Dal momento che *ITALANT* si propone soprattutto come studio sintattico dell'italiano antico[10], non sembrerà ingiustificata l'idea, originatasi nel gruppo torinese del progetto «Ricerche linguistiche sull'italiano antico», di annotare morfosintatticamente il *Padua Corpus*.

Alcuni degli appartenenti a tale gruppo, in particolare Bice Mortara Garavelli e Carla Marello, si interessano di aspetti come il discorso riportato e l'ellissi che si studiano molto più agevolmente su un corpus annotato con informazioni di parte del discorso e di morfologia flessionale. Tuttavia l'intero gruppo dei partecipanti ad *ITALANT* trarrà beneficio dall'uso del *Padua Corpus* annotato.

D'altra parte il gruppo torinese di ricerca[11] potrà affinare il proprio lavoro alla luce dei suggerimenti che gli verranno dai partecipanti ad *ITALANT* soprattutto in due fasi della loro attività di annotazione: la revisione delle etichette grammaticali e la scrittura di regole di disambiguazione.

La linguistica dei corpora è il terreno sul quale da anni si attua quel «balancing act», quella combinazione di approcci simbolici e statistici allo studio delle lingue che di recente, da più parti e sempre più esplicitamente, è auspicato[12].

3. Quale strategia di annotazione? Stocastica o basata su regole?

L'annotazione morfosintattica (*part-of-speech tagging*) di un corpus è una delle tecniche che sono raggruppate sotto il termine ombrello di analisi parziale (*partial parsing*). Infatti permette di recuperare una parte delle informazioni che si ottengono dalla tradizionale analisi sintattica, ed è pure una temporanea rinuncia alla completezza dell'analisi in favore di una ragionevole approssimazione, attraverso il raggiungimento abbastanza agevole di discreti risultati.

Fra gli addetti ai lavori sono diffuse due diciture abbreviate per distinguere i principali procedimenti di annotazione di corpora: annotatori (*taggers*) stocastici e annotatori basati su regole.

I primi si basano su modelli marcoviani nascosti (*HMM* «Hidden Markov Models») importati dalle tecniche di riconoscimento vocale. In tali modelli gli stati sono annotazioni (*tags*) o *n*-tuple di annotazioni. Le probabilità di transizione sono le probabilità di comparire che un'annotazione ha in relazione all'annotazione precedente; le probabilità di emissione sono le probabilità di comparire di una parola in relazione ad un'annotazione; la probabilità di una particolare sequenza di parti del discorso in una frase è il prodotto delle probabilità di transizione ed emissione che contiene[13].

I vantaggi di questo tipo di annotatori stocastici consistono nella loro accuratezza e nel fatto che si possono applicare anche a testi non annotati. Tuttavia è necessario chiarire cosa s'intende per accuratezza e le conseguenze a cui si va incontro quando li si applica a testi non annotati in alcun modo: se si assume la frase come unità rilevante per il proprio studio, il tasso di errore caratteristico degli annotatori stocastici, il 4% (vale a dire che quattro parole su cento sono annotate erroneamente), si trasforma in un 56% di possibilità di errore per frase[14]. È stato fatto notare come senza annotatori stocastici, solo attribuendo alla parola la sua più probabile parte del discorso, indipendentemente dal contesto, si ottiene un tasso di errore del 10% (Church & Mercer 1993). Applicare un annotatore stocastico a un testo non annotato comporta un grande lavoro successivo di perfezionamento manuale o semiautomatico: è perciò vantaggioso partire da un corpus non annotato con l'aggiunta di una porzione di testo (detto *training corpus*) annotato manualmente con cui «confrontare e correggere» i risultati dell'annotazione stocastica.

Gli annotatori basati su regole assegnano le annotazioni sulla base di un lessico annotato acquisito, sulla base dell'analisi morfologica e sul successivo intervento di regole contestuali del tipo «la parola in esame non è un verbo se la parola precedente è un determinante». Questo tipo di annotatori richiede molto tempo per l'individuazione delle regole di disambiguazione contestuale, ma generalmente è più rapido e preciso del precedente; negli ultimi anni si sono studiate tecniche per ricavare automaticamente regole di disambiguazione da testi annotati manualmente e per poi reinserire queste regole in annotatori.

Gli annotatori stocastici basati su *HMM* sono modelli generativi: la sequenza di annotazioni che assegnano come più probabile è quella associata alla sequenza di passi secondo la quale la frase è stata più probabilmente generata; gli annotatori basati su regole seguono invece modelli di regressione/classificazione[15]. Gli annotatori basati su regole, pertanto, sono generalmente più intuitivi, ma richiedono moltissimo lavoro preparatorio: è praticamente indispensabile avere a disposizione una parte del corpus annotata manualmente.

Le due strategie di annotazione non sono in realtà in opposizione: ultimamente, anzi, si stanno diffondendo annotatori che usano sia procedimenti stocastici sia modelli a classificazione/regressione[16].

Nel nostro caso la scelta di un annotatore basato su regole è stata guidata da una serie di caratteristiche del *Padua corpus* e dal fatto di poter acquisire una parte di informazioni utili per l'annotazione da una preesistente base di dati.

3.1 Peculiarità del Padua Corpus.

Il *Padua Corpus* è di dimensioni piuttosto limitate (circa ventimila forme per duecentomila occorrenze), almeno secondo gli attuali parametri della linguistica dei corpora, e perciò i procedimenti di annotazione stocastica sono meno consigliabili di quelli basati su regole[17]. Un *training corpus* annotato a mano per «aiutare» l'annotazione stocastica avrebbe comportato altrettanto lavoro quanto la confezione di regole di disambiguazione: mediamente[18], infatti, si può affermare che una accuratezza del 92% è raggiunta solo con *training corpora* di circa centomila parole.

D'altra parte l'italiano del Duecento è caratterizzato da una sintassi diversa da quella dell'italiano d'oggi in misura tale da sconsigliare l'uso di un *training corpus* annotato manualmente ma non coevo[19]. Si pensi che Ulrich Heid (c.p.) ha riferito di cattive *performances* di annotatori stocastici su corpora tedeschi per i quali si era provato ad usare un *training corpus* fatto di testi coevi, ma di tipo testuale diverso: un *training corpus* costituito da favole aveva avuto risultati molto insoddisfacenti su un corpus di manuali e testi di riferimento, un po' meno insoddisfacenti, ma sempre mediocri, su un corpus di articoli di giornale.

Tuttavia la spinta più forte a scegliere un annotatore basato su regole è venuta dall'avere a disposizione oltre al *Padua Corpus* un insieme di dati sul materiale lessicale che lo compone.

3.2 L'eredità condizionante del GATTO.

La procedura GATTO («**G**estione degli **A**rchivi **T**estuali del **T**esoro delle **O**rigini») è nata come strumento «finalizzato alla costruzione, gestione ed interrogazione del corpus di testi che è alla base del *Vocabolario Storico della Lingua Italiana* in corso di realizzazione presso l'OVI».[20]

Il fatto che GATTO sia una procedura di costruzione, interrogazione e gestione del corpus più ampio da cui il *Padua Corpus* è stato tratto, assume per noi, allo stato attuale del nostro progetto di annotazione morfosintattica, uno speciale valore in quanto, per svolgere le sue triplici summenzionate funzioni, GATTO è dotato di una serie di informazioni per noi assai utili. In particolare contiene, sia pure distribuita in modo non omogeneo su tutti i testi, una prima parziale lemmatizzazione: più di un quinto delle occorrenze del corpus è già stato assegnato ad un lemma associato ad una categoria grammaticale[21].

Alla luce di questo fatto sembrerebbe che il nostro lavoro consista in poco più di un mero problema di estrazione di dati e che di fatto il gruppo di lavoro torinese disponga anche di un *training corpus*.

In realtà non è proprio così, perché nel *Padua Corpus* non ci sono testi o pezzi significanti di testi annotati per intero, tali cioè da fornire (oltre a

forme riconosciute nella loro flessione e ricondotte a lemmi associati ad annotazioni di parti del discorso) *sequenze di frasi intere annotate*. L'annotazione è un'annotazione fatta a fini lessicografici: è il frutto della ricerca di luoghi interessanti per chi deve costruire la voce del *Vocabolario Storico della Lingua Italiana*.

È perciò un'annotazione «a placche»: poiché fine della procedura non è produrre un corpus annotato morfosintatticamente, non tutte le forme di un lemma presenti nel corpus sono necessariamente state ricondotte a quel lemma; non tutte le occorrenze di quella forma sono ricondotte al lemma[22]; non è detto che tutti i lemmi presenti nel corpus siano nel lemmario; non è detto che di una forma o di un lemma si diano tutte le parti del discorso che riveste nel corpus.

Nostro primo compito, pertanto, è acquisire le informazioni presenti nei *files* di GATTO (e GATTO, per fortuna, contempla una propria procedura per estrarre i testi ed i loro lemmari: cfr. Iorio-Fili 1998a, 21-22 e 28-29, e qui *infra* § 4.1), «spalmarle» su tutte le occorrenze, e colmare le lacune in modo che non vi sia discrepanza, ma anzi completa omogeneità tra le informazioni acquisite e quelle integrate in un secondo momento.

Vorremmo inoltre rendere l'annotazione meno lessicografica di quanto non sia attualmente e più orientata a soddisfare gli standard richiesti dalle indicazioni di EAGLES[23] (cfr. *infra* § 5.1 a proposito della scelta della batteria di etichette), in modo da entrare a far parte del virtuale insieme di corpora della ricerca internazionale annotati in modo «compatibile»: in tal senso andranno lette alcune delle proposte del § 5.1. Poi comincerà davvero il lavoro sulle regole di disambiguazione, ed in quella fase faremo uso anche di processi probabilistici.

4. «Acquisizione (di informazione) lessicale»: un'espressione polisemica

Fino a non molti anni fa quando si parlava di *lexical acquisition*[24] la mente degli addetti ai lavori andava subito a studi sull'apprendimento spontaneo da parte del bambino del lessico di una lingua naturale ed eventualmente di più d'una se era allevato in ambiente multilingue[25].

Già negli anni '80, però, il termine estrazione automatica di informazioni da banche di dati elettroniche comincia ad essere affiancato da *acquisizione,* complice il fatto che si parla di *knowledge acquisition*. Dal 1990 in poi i titoli di contributi in cui *acquisizione di informazioni lessicali* o *lexical acquisition* compaiono in relazione a *corpus* od a *dictionary* aumentano esponenzialmente fino ad uguagliare e sorpassare quelli in cui *lexical acquisition* è intesa come apprendimento spontaneo da parte di umani.

Nel 1996 Boguraev & Pustejovsky curano un libro dal titolo *Corpus Processing for Lexical Acquisition* la cui copertina riproduce una definizione di *bread* così come apparirebbe in un dizionario elettronico. È appunto nell'ottica dell'elaborazione di testi di un corpus in vista di acquisire (estrarne, se si preferisce) informazioni lessicali che si è mossa una procedura come GATTO. Anzi GATTO[26] stesso è il risultato di precedenti elaborazioni informatiche del corpus dell'OVI[27], e nel contempo uno strumento per la costruzione del *Vocabolario Storico della Lingua Italiana*, almeno nella sua forma *on-line*.

Ed è in questa prospettiva che ci muoviamo anche noi: per il nostro progetto di annotazione morfosintattica, infatti, l'acquisizione lessicale ha un duplice significato:

(a) acquisire (estrarre a prezzo di un certo sforzo, in verità) i *files* già presenti in GATTO;
(b) acquisire dal *Padua Corpus* attraverso tecniche di elaborazione del corpus (in parte simili e in parte più sofisticate di quelle presenti in GATTO) le informazioni necessarie per arrivare ad una annotazione morfosintattica semiautomatica.

Nei paragrafi successivi parleremo in dettaglio dei problemi connessi all'acquisizione e all'integrazione dei dati presenti in GATTO.

4.1 Acquisizione dei dati di GATTO.

Come abbiamo già accennato, GATTO contempla una propria procedura (ufficialmente descritta nel manuale di riferimento Iorio-Fili 1998a alle pagine 21-22 e 28-29) per estrarre i testi, completi dei loro codici di lemma[28] e di varie altre codifiche[29], ed i loro lemmari, due *outputs* per noi particolarmente preziosi. I formari, direttamente ottenibili in GATTO stampando su singoli *files* TXT i testi selezionati nel menu *Ricerche* | *Formario singolo testo* non risultano, invece, purtroppo utilizzabili ai nostri fini, dato che presentano due sole colonne di dati, una con le forme ed una con il numero di occorrenze presenti, e mancano completamente dei riferimenti al lemmario ed alle sue associazioni[30].

Comunque con questo procedimento ci si trova almeno a disporre di una versione dei ventun testi con tutte le codifiche ed i riferimenti («codici di lemma») ai rispettivi lemmari etichettati. Per meglio valutare pregi e limiti dei materiali ottenuti esaminiamo prima un piccolo campione (il celebre incipit della *Vita Nuova*) che presenti un buon rapporto tra lemmatizzato e non lemmatizzato (18 su 54), ed in cui non si verifichino particolari problemi di errate transcodifiche od altri minori inconvenienti[31].

extract.fr

@@AS
% cap. 01
0003 In quella parte del =1=libro de la mia =2=memoria =3=dinanzi a la quale =4=poco si =5=potrebbe =6=_*_leggere, si =7=trova una =8=rubrica la quale =9=dice: &CIncipit vita nova&c. =10=Sotto la quale =8=rubrica io =11=_*_trovo =12=scritte
0004 le parole le =13=_*_quali =14=_*_è mio =15=intendimento d' =16=assemplare in questo =17=libello; e se non tutte, =18=almeno la loro sentenzia.

legenda

@@...	codifica di *character set* (ASCII)
%...	riferimento organico (capitolo)
$...$	riferimento di pagina
&C... &c	campo speciale (corsivo)
_*_	carattere di evidenziazione
= *n* =	codice di lemma (della parola seguente)

lemmitxt.fr

1	libro	s.m.
2	memoria	s.f.
3	dinanzi a	prep.
4	poco	indef.
5	potere	v.
6	leggere	v.
7	trovare(-si)	v.
8	rubrica	s.f.
9	dire	v.
10	sotto	prep.
11	trovare	v.
12	scrivere	v.
13	quale	rel.
14	essere	v.
15	intendimento	s.m.
16	esemplare	v.
17	libello	s.m.
18	almeno	avv.

Tav. 1

4.2 Prima elaborazione dei dati di GATTO.
A parte una preliminare e provvisoria ripulitura dei dati (cfr. quanto detto in n. 31), si imponevano alcuni ulteriori trattamenti dei dati già presenti in GATTO.

In primo luogo bisognava attuare una unificazione dei lemmari (la cui gestione come monadi separate sarebbe risultata troppo complessa da gestire ai nostri scopi) ed un aggiornamento dei riferimenti presenti nei testi.

Questi ultimi sono per il momento stati risolti introducendo direttamente il lemma e le categorie grammaticali corrispondenti accanto al vecchio codice di lemma. Il nostro incipit della *Vita Nuova* ora si presenta[32] così:

```
@@AS
  % cap. 01
$0003$ In quella parte del =s.m._libro=1=libro de la mia=s.f._me-
moria=2=memoria =prep._dinanzi a=3=dinanzi a la quale=indef._
poco=4=poco si =v._potere=5=potrebbe =v._leggere=6=_*_leggere, si
=v._trovare(-si)=7=trova una =s.f._rubrica=8=rubrica la quale =v.
_dire=9=dice: &CIncipit vita nova&c. =prep._sotto=10=Sotto la quale
=s.f._rubrica=8=rubrica io =v._trovare=11=_*_trovo =v._scrivere=
12=scritte
$0004$ le parole le =rel._quale=13=_*_quali =v._essere=14=_*_è mio
=s.m._intendimento=15=intendimento d' =v._esemplare=16= assemplare in
questo =s.m._libello=17=libello; e se non tutte, =avv._ almeno= 18=almeno
la loro sentenzia.
```

Tav. 2

Un lemmario unificato, invece, è stato ottenuto dalla somma dei ventun lemmari parziali e dal riaggiornamento delle occorrenze sul totale dei testi (cfr. Tav. 3a).

In secondo luogo, poi, oltre a testi e lemmario, urgeva anche l'allestimento di un formario unificato, in cui a fianco di ogni forma fossero presentati il numero di occorrenze e (quando già presenti) i lemmi \ categoria grammaticale associati. Tale risultato è stato ottenuto dal gruppo DIMA a partire dai *files* «preripuliti» di testo e lemmario. Nella tavola seguente presentiamo un campione scelto casualmente del formario unificato (Tav. 3b) affiancato dalla porzione pressapoco corrispondente del lemmario unificato (Tav. 3a):

lemmario		
prologo	6	s.m.
promessa	1	s.f.
promettere	31	v.
promissione	6	s.f.
promuovere	1	v.
pronto	2	agg.
pronunciagione	1	s.f.
pronuntiatio	3	s.f.
propensare	1	v.
propinquità	1	s.f.
propinquo	3	agg.
proponimento	8	s.m.
proporre	27	v.

Tav. 3a

formario		
promisero	1	v._promettere
promisi	2	v._promettere
promosso	1	v._promuovere
promover	1	
pronta	1	agg._pronto
prontezza	1	
pronto	1	agg._pronto
Pronuntiatio	1	s.f._pronuntiatio
pronuntiatio	4	s.f._pronuntiatio
pronunziagione	1	s.f._pronunciagione
propensato	1	v._propensare
propi	1	
propî	1	poss._proprio

Tav. 3b

Con questi primi tre risultati (testi con sporadiche annotazioni, lemmario provvisorio e formario generale) più un altro numero di *files* secondari (formari inversi, formari suddivisi per forme mai, sempre e solo talvolta annotate, ecc.) abbiamo comunque una buona base di partenza per affrontare il completamento dell'etichettatura secondo abbiamo prospettato nel § 3. Sennonché ci si affaccia un'ulteriore preliminare questione: la scelta del sistema di etichette (*tagset*).

5. Il sistema di etichette grammaticali

La lista delle categorie grammaticali[33] che abbiamo ereditato è presentata nella tavola seguente (Tav. 4), dove a fianco di ogni sigla è anche dato il suo numero di occorrenze (estratto dalla DIMA Logic):

a.g.	13	agg.	3902
antr.	6	art.	8
avv.	2233	cong.	1012
cong.s.	3	corp.	1
dim.	512	escl.	16
fest.	36	indef.	939
interr.	67	lat.	3
n.	2	n.g.	423
n.op.	40	n.p.	82
n.p.i.	11	num.	171
pers.	177	poss.	30
prep.	693	rel.	137
s,f,.	1	s.f.	5915
s.f.pl.	12	s.i.	15
s.m.	6193	v.	14707

Tav. 4

La scelta, indubbiamente, sembra alquanto casuale; vi sono però precise ragioni per questo stato di cose. In primo luogo, il nostro corpus di partenza non è altro che un ritaglio di un corpus molto più ampio, e pertanto anche le nostre etichette non sarebbero altro che una selezione casuale da un sistema più ampio; in secondo luogo (P. Beltrami, c.p.), non si può neppure dire che *vi sia* un «sistema» unitario per tutta la base dati del *TLIO*, quanto piuttosto un complesso palinsesto di *tagsets*, ereditati dalle varie epoche e gestioni del *Vocabolario*, che viene ormai sistemato solo in fase di redazione di voce.

Anche tenendo conto di tale situazione, la granularità di questa batteria di etichette è comunque troppo bassa e disomogenea: ad esempio, disporre di un'unica etichetta «verbo» e di un'unica «congiunzione» non è compatibile con ricerche sintattiche di profondità appena più che minima;

ed a fronte di un'unica etichetta per il verbo, in area nominale si hanno invece ben dieci etichette, alcune delle quali abbastanza curiose ed incoerenti al loro interno[34]. Nella tavola seguente sono raccolte tutte le etichette nominali presenti attualmente nel *Padua corpus* con alcuni esempi (per meglio valutarne la portata) dei lemmi cui sono state attribuite[35]:

s.m.	mantello – mantenimento – mantovano – marzo – miele
s.f.	campana – conoscienza – margherita – settimana
s.i.	berbice – fine – fonte – erede – oste *[tutti]*
s.f.pl.	calende – capita – nozze *[tutti]*
n.p.	Altissimo – Antichristo – Babèl – Christo – Dio – Domenedio – Donna – Figlio – Maestà – Madonna Sancta Maria – Misericordioso – Spirito Santo – Vergine Maria Domenedio – Donna – Figlio – Maestà – Madonna Sancta Maria – Misericordioso – Spirito Santo – Vergine Maria
n.p.i.	Acerra – Altafronte – Cestella – Carmino – Malta – Spagnata
n.g.	Babillonia – Bari – Borgo di Piazza Oltr'Arno – Borgognia – Campo Marzio – Charon – Corneto – Fiandra – Fiorenza – Laterano – Magna – Mantua – Massa – Mercato Vecchio – Mirra – Napoli – Osstia – Pavia – Ponte Vecchio – Rodano – Sacta Fiore – Spagna – Terra Santa – Torre della Volpe – Via Nuova
n.op.	Apochalipx – Bibbia – Castel Sant' Agnolo – Chiesa di Sancto Martino – Chiesa di Sancto Salvatore – Chiesa di Santo Petro – Credo in Deo – Culiseo di Roma – De Civitate Dei – Delli Offici – Eneida – Gloria in excelsis Deo – Libro d'Amerigo – Libro di Remedio d'Amore – Paternostro – Poetria – Requiem etterna – San Lorenzo fuor le mura – Spedale di Sancto Spiritu – Tesoro – Topica – Vecchio Testamento
corp.	Leonista *[tutti]*
fest.	Ascensione – Befanie – Candelora – Innocenti – Natale – Ogni Sancti – Pasqua – San Salvadore

Tav. 5

5.1 Struttura gerarchica di etichette.

Oltre alle osservazioni già fatte, le etichette che abbiamo ereditato non risultano conformi ai più recenti standard elaborati dalla *corpus linguistics* (come quelli fissati da EAGLES) anche perché sono «compatte» (caratteristica probabilmente non problematica nel progetto lessicografico dell'OVI), ossia non costruite in base ad una struttura gerarchica di «etichette tipate»[36].

L'uso di etichette tipate, va detto, è ormai pressoché universale in *corpus linguistics*; tra i vari approcci e modelli seguiti, poi, il modello della famiglia di tipi a struttura gerarchica è risultato senz'altro il più sintetico ed efficiente[37], in quanto sostituisce un gran numero di regole lessicali incor-

porando nella struttura dei tipi anche molte delle relazioni linguistiche esistenti all'interno del lessico.

L'utilizzo di etichette analitiche nella annotazione di un corpus (quali sono quelle, ad esempio, già presenti in GATTO per l'area nominale), in effetti, ne permette una descrizione dettagliata e ricerche specifiche; tuttavia l'analiticità risulta dispersiva se non viene sussunta in un sistema di generalizzazioni gerarchiche, fondata sull'ereditarietà.

Quel che pertanto ci riproporremmo è disperdere il meno possibile il lavoro contenuto in GATTO per riportarlo, nell'ottica del futuro lavoro sul corpus, ad un sistema gerarchico di tipi con ereditarietà[38]. Ma la contestuale proposta di un nuovo *tagset* con le caratteristiche preconizzate sarà argomento di un altro contributo, dato che in questa sede intendevamo limitarci ai soli problemi della *acquisizione* dei dati da un progetto preesistente e con diverso orientamento (lessicografico anziché grammaticale).

Manuel Barbera e *Carla Marello*
Università di Torino

Note

* Anche se quanto espresso nel presente articolo è frutto di un lavoro condotto in stretta collaborazione ed è pertanto pienamente condiviso da entrambi gli autori, i §§ 1, 4.1, 4.2 e 5 sono specialmente dovuti a M. Barbera ed i §§ 2, 3 e 4 a C. Marello.

1. Si tratta, in breve, dei seguenti testi: Maestro Rinuccino, *Rime*; Brunetto, *Rettorica*, *Favolello* e *Tesoretto*; *Fiori di filosafi*; *Libro del dare e dell'avere di Castra Gualfredi*; *Capitoli della Compagnia di San Gilio*; Jacopo Cavalcanti, *Tre Sonetti*; *Lettere di Consiglio de' Cerchi*; Bono Giamboni, *Libro de' Vizî e delle Virtudi* e *Trattato di Virtù e di Vizi*; *Libro del dare e dell'avere di Lapo Riccomanni*; Dante, *Vita nuova* (ed. Barbi); *Capitoli della Compagnia della Madonna d'Orsammichele*; *Libro della Compagnia di Santa Maria del Carmine*; Guido Cavalcanti, *Rime*; *Novellino* (ed. Favati); *Disciplina Clericalis*; *Cronica fiorentina.*
2. La restrizione cronologica, in effetti, ha meno bisogno di giustificazioni: la lingua dei peraltro pochi testi fiorentini precedenti gli anni '60 del Duecento è nettamente più arcaica, mentre quella trecentesca, già a partire dall'ultimo Dante, tende, via via più marcatamente, a ricostruirsi come paradigma letterario di se stessa. Solo, semmai, si sarebbe preferito che fosse incluso nel *Padua Corpus* anche il *Convivio*, la cui importanza nella prosa delle Origini difficilmente può essere sottostimata. Diciamo che, restando nell'idea che una opportuna simulazione di sincronia non debba superare il cinquantennio, la definizione cronologica «dalla *Rettorica* di Brunetto al *Convivio*» ci sarebbe sembrata una scelta forse più opportuna di «dal 1251 al 1300».
3. Senza dimenticare, naturalmente, gli studi dello stesso Renzi (cfr. Renzi 1987, pp. 172-173 e 1998a).

4. Cfr. Renzi (1998a, pp. 23-28), che riassume anche tutta la questione e la bibliografia precedente.
5. Parliamo al plurale perché, ovviamente, nel Duecento il fiorentino non ha ancora vinto: per Dante, d'altra parte e più in generale, la caccia alla «pantheram quam sequimur» del «vulgare illustre» era ancora ben aperta prima che lui stesso si lanciasse sull'usta e che, in un certo qual senso, creasse la selvaggina medesima.
6. L'iniziativa non sarebbe forse pertinente data l'impostazione amichevolmente «anti-filologica» data al progetto di *ITALANT* (cfr. Renzi 1998, pp. 21-22); bisogna però ribadire che, per una «lingua morta» come l'italiano antico, noi ci possiamo solo confrontare con testi, e che non possiamo mai prescindere da una loro comprensione filologicamente corretta anche quando non è questo il nostro fine ultimo.
7. Avrebbe, d'altro canto, senso una grammatica di antico francese che descrivesse il solo franciano, astraendolo completamente dalla complessa dialettica letteraria e culturale che intrattiene con normanno, champenoise e piccardo? Chrétien era pur sempre uno champenois ...
8. Come diversamente anche suggerito dalla relazione di Minne De Boer, *La struttura della parola nel Novellino*, all'incontro *ITALANT* di Padova, *Italiano antico e corpora elettronici*, 19-20 febbraio 1999.
9. All'interno del progetto cofinanziato dal MURST nel 1997 «Ricerche linguistiche sull'italiano antico», sorto con l'intento di fiancheggiare i lavori di *ITALANT*, il gruppo di ricerca di Torino, a cui appartengono gli autori del presente contributo, si propone di studiare aspetti testuali dell'italiano antico, fra cui gli aspetti sintattici legati al tipo di testo. Da questo punto di vista, la presenza dei testi documentari, la cui grammatica testuale è sostanzialmente divergente da quella degli altri tipi testuali, comporta problemi ancor maggiori della presenza dei testi poetici.
10. Interpretiamo così l'osservazione di Renzi sul fatto che *ITALANT* «nasce da una costola della *Grande Grammatica Italiana di Consultazione*» (Renzi 1998b, p. 7) ed è «la scommessa che i linguisti-linguisti siano capaci di mettere in azione nell'italiano antico quelle strategie di studio che già hanno saputo impiegare sull'italiano moderno» (Renzi 1998a, p. 30).
11. Di cui fanno parte oltre alla Prof.ssa Mortara Garavelli, coordinatrice, Carla Marello, Manuel Barbera, Valerio Allegranza e, per l'elaborazione informatica, i ricercatori della DIMA Logic (dott. Cesare Oitana e Daniele Rizzo).
12. Il riferimento diretto è al titolo del libro curato da Klavans & Resnik (1996), che contiene i contributi presentati ad un *workshop* dal medesimo titolo tenutosi a Las Cruces, New Mexico, nel 1994. Tuttavia si può risalire ben più indietro: la linguistica dei corpora è da sempre terreno di confronto fra grammatiche basate sull'intuizione e grammatiche basate sull'osservazione (cfr. Aarts 1991) e recentemente c'è chi vi ha visto la humus della «*approximate linguistics*» (Grefenstette 1998, p. 27). Più in generale, si può anche notare come l'uso (preliminare) di modelli grammaticali in varia misura «rozzi» (tradizionali, superficiali, ecc.), non sia intrinsecamente pregiudiziale ad un (successivo) impiego di modelli descrittivi «forti»: si potrebbe, ad esempio, richiamare in proposito l'utile distinzione adottata da Giorgio Graffi tra concetti «ingenui» e concetti «teorici» (cfr. Graffi 1991).

13. Per una panoramica dettagliata sugli *HMM* nell'elaborazione informatica del linguaggio orale e scritto si veda Knill & Young (1997).
14. Calcolando la frase formata in media da 20 parole: cfr. Abney (1997, p. 121).
15. Rinviamo ad Abney (1997) per ulteriori dettagli statistico-matematici. Mancini (1993, pp. 64-85) descrive i risultati dell'impiego di una classificazione grammaticale automatica al corpus del *LIP* (Lessico di frequenza dell'italiano parlato).
16. Fra i più conosciuti annotatori ibridi c'è CLAWS, cfr. Garside & Smith (1997).
17. Per un diagramma dei rapporti tra dimensione del *training corpus* e percentuale statistica di accuratezza con diversi tipi di annotatori cfr. Schmid (1994, tav. 3).
18. Tale affermazione è una generalizzazione dei dati di Schmid (1994, tav. 3).
19. Siamo a conoscenza di testi coevi o di poco posteriori (ad es. la *Divina Commedia*) annotati su supporto elettronico per il progetto CIBIT, coordinato dal Prof. Mirko Tavoni. Tali testi sono stati annotati con *PITagger*, un software elaborato da Eugenio Picchi (cfr. l'URL http://www.ilc.pi.cnr.it/pisystem/intro.htm), e sono interrogabili attraverso una nuova versione di DBT (*Data Base Testuale*) che consente di interrogare un corpus anche per parti del discorso oltre che per lemmi e forme. Speriamo di poter avere accesso a tali testi in un prossimo futuro, ma per ora abbiamo dovuto procedere senza il loro aiuto.
20. Così recita la prima pagina della *Guida all'uso* (Iorio-Fili 1998). Per una descrizione dettagliata di GATTO, si vedano anche i due manuali di riferimento (Iorio-Fili 1998a e b).
21. Il sistema avviene mediante una gestione separata di testo ed etichette: i riferimenti ai lemmi e le etichette grammaticali non sono direttamente incapsulate nei testi, ma sono mantenute in una serie di *files* distinti (uno per testo) e correlate nei testi con cifre, dette «codici di lemma», il cui valore numerico è espresso da ventun serie numeriche distinte, tutte originanti da «1» in ordine sequenziale di occorrenza nel testo.
22. Delle 20.247 forme del testo (i dati numerici che abbiamo estratto [17.II.99] sono ancora lievemente sporchi per eccesso, ma in misura statisticamente non apprezzabile) 9.734 (per 11.688 occorrenze) sono sempre lemmatizzate, 6436 sono solo parzialmente lemmatizzate (per 25.672 occorrenze lemmatizzate e 128.843 non) e 4.077 (per 51.898) non sono mai state lemmatizzate. Globalmente, pertanto, su 218.801 occorrenze 37.360 sono lemmatizzate e 180.741 non lo sono. I lemmi finora assegnati sono invece 6.421 (6.827 contando gli «omografi»).
23. EAGLES, cioè «Expert Advisory Group on Language Engineering Standards», è un'iniziativa della Commissione Europea all'interno del «DG XIII *Linguistic Research and Engineering programme*». Il gruppo ha un sito web da cui si possono ricavare gli standards per l'italiano (URL: http://www.ilc.pi.cnr.it e ftp://ftp.ilc.pi.cnr.it/pub/eagles/).
24. In italiano si mantiene una certa distinzione fra *acquisizione del lessico,* appannaggio degli studi sull'appprendimento spontaneo da parte di umani e *acquisizione (di informazione) lessicale,* espressione tipica degli studi di linguistica computazionale, anzi più precisamente di *linguistic engineering*, che accompagnano le ricerche di linguistica dei corpora e di linguistica cognitiva.
25. Un campione esemplare: il libro a cura di Gleitman & Landau (1994) dal titolo *The Acquisition of the Lexicon.*

26. Il programma è attualmente giunto alla versione 2.3.17.
27. Per una breve introduzione al progetto di GATTO cfr. Iorio-Fili (1997).
28. Cfr. la precedente nota 21.
29. Che sono, nella procedura GATTO, almeno i «riferimenti organici» (titolo del capitolo, numero di paragrafo ecc.), i «riferimenti di pagina», i «riferimenti topografici ed a versi» (numeri di volume, riga, verso), i «campi formula» (aree delimitate da due particolari codici «che l'interprete non controlla e che risultano 'non appartenenti al testo indicizzato'»), ed altre codifiche minori: cfr. Iorio-Fili (1998a, 41-49).
30. La nuova versione 3.01 di GATTO dovrebbe avere in parte ovviato a tali inconvenienti (P. Beltrami, c.p.).
31. Ci limitiamo, infatti, ad accennare che, per varie ragioni, raramente le cose procedono così lisce: talvolta si verificano, infatti, alternanze di due diversi codici per un medesimo grafema; altre volte si nota una certa trascuratezza nella notazione degli spazi, specie in relazione ad apostrofi, punti di clisia e segni interpuntivi. Tutti difetti minuscoli, se vogliamo, ma che creano gravi problemi in una gestione completamente automatica dei dati. Tali incoerenze, che sono il naturale portato della lunga storia della base dati dell'OVI (per le dimensioni di questo aspetto informatico cfr. Marinelli 1997), stanno lentamente venendo corrette man mano che procede l'allestimento del Vocabolario.
32. Naturalmente così si presenta per chi lavora a completare l'annotazione del corpus, in ispecie per la formulazione delle regole di disambiguazione, e certo non per l'utente finale (sia esso di GATTO od altro *software*), al quale tutti i materiali qui esibiti propriamente non sono destinati.
33. Una volta, naturalmente, uniformate le sigle non puntate alle puntate, che nel corpus di partenza appaiono in alternanza libera.
34. A tale situazione caotica contribuisce anche il fatto che tutta l'onomastica è ormai un «ramo secco», escluso dalla redazione del TLIO e pertanto da tempo non più implementato e riorganizzato nella base dati.
35. Quando è presente l'indicazione [Tutti] significa che quelli sono nel corpus tutti i lemmi a cui è stata assegnata l'etichetta; negli altri casi, invece, sono riportati solo pochi esempi rappresentativi.
36. Cfr. scritti di Cesare Oitana per uso interno della DIMA Logic. Sull'argomento cfr. Krieger & Nerbonne (1991) e Schmidt (1993).
37. Infatti nella linguistica contemporanea, dalla LFG («Lexical Functional Grammar»), alla HPSG («Head-Driven Phrase Structure Grammar»; cfr. Pollard & Sag 1987) ed al sofisticato CUF («Comprehensive Unification Formalism»; cfr. Dörre & Dorna 1993), sono largamente diffusi sistemi di tipi gerarchici che si estendono dal lessico alle strutture sintattiche e semantiche.
38. La DIMA Logic, che fa parte del gruppo di ricerca torinese, ha creato un sistema analogo per le annotazioni lessicali (morfosintattiche e semantiche) dei suoi programmi applicativi.

Bibliografia

Aarts, J. (1991): Intuition-Based and Observation-Based Grammars, in : Aijmer K. & B. Altenberg (eds.) : *English Corpus Linguistics. Studies in Honour of Jan Svartvik*. Longman, London and New York, pp. 44-62.

Abney, S. (1997): Part-of-Speech Tagging and Partial Parsing, in: Young S. & G. Bloothooft (eds.) : *Corpus-based Methods in Language and Speech Processing*. Kluwer, Amsterdam, pp. 118-136.

Ascoli, G. I. (1882-1885): L'Italia dialettale. *Archivio Glottologico Italiano* VIII, Torino-Firenze, pp. 98-128. (Parzialmente riprodotto anche in Ascoli, G. I. (1975) : *Scritti sulla questione della lingua* a cura di Corrado Grassi, Einaudi, Torino, pp. 57-62.)

Boguraev, B. & J. Pustejovsky (eds.) (1996): *Corpus Processing for Lexical Acquisition.* MIT Press, Cambridge (Mass.) – London (England).

Church, K. & R. L. Mercer, (1993): Introduction to a special issue on computational linguistics using large corpora.*Computational Linguistics* XIX, pp. 1-24.

Dörre, J. & M. Dorna (1993): CUF – *A Formalism for Linguistic Knowledge Representation*, Deliverable R.1.2A, DYANA 2, August 1993. (Postcript version available also on the FTP server of Stuttgart IMS, http://www.ims.uni-stuttgart.de/).

Garside, R. & N. Smith, (1997): A Hybrid Grammatical Tagger: CLAWS4, in: Garside R., G. Leech & A. Mc Enery (eds.), pp.102-121.

Garside, R., G. Leech & A. Mc Enery (eds.) (1997): *Corpus Annotation. Linguistic Information from Computer Text Corpora*. Longman, London & New York.

Gleitman, L. & B. Landau, (1994): *The Acquisition of the Lexicon*, MIT Press, Cambridge (Mass.) – London (England).

Graffi, G. (1991): Concetti «ingenui» e concetti «teorici» in sintassi. *Lingua e stile* XXVI, Bologna, pp. 347-363.

Grefenstette, G. (1998): The Future of Linguistics and Lexicographers: Will there be Lexicographers in the years 3000?, in: Fontenelle, T., Ph. Hiligsmann, A. Michiels et al. (eds.), *euralex '98 Proceedings. Papers* submitted to the Eighth EURALEX International Congress on Lexicography in Liège, Belgium, vol. 1, Liège, Université de Liège – Département d'anglais et de néerlandais, 1998, pp. 25-41.

Iorio-Fili, D. (1997): Un nuovo software lessicografico: GATTO, Opera del Vocabolario italiano. *Bollettino,* II, Firenze, pp. 259-270.

Iorio-Fili, D. (1998): *GATTO. Guida all'uso. Versione 2.2, revisione 47* (allo sviluppo del programma ha collaborato Francesco Leoncino). Firenze, CNR – Centro Studi OVI, 2 novembre 1998. File DOC incluso nel pacchetto.

Iorio-Fili, D. (1998a)*: GATTO. Manuale di riferimento: gestione dati. Versione 2.2, revisione 47* (allo sviluppo del programma ha collaborato Francesco Leoncino). Firenze, CNR – Centro Studi OVI, 2 novembre 1998. File DOC incluso nel pacchetto.

Iorio-Fili, D. (1998b): *GATTO. Manuale di riferimento: ricerche. Versione 2.2, revisione 47* (allo sviluppo del programma ha collaborato Francesco Leoncino). Firenze, CNR – Centro Studi OVI, 2 novembre 1998. File DOC incluso nel pacchetto.

Kahrel, P., R. Barnett & G. Leech, (1997): Towards cross-linguistic standards or guidelines for annotation of corpora, in Garside, R., G. Leech & A. Mc Enery (eds.), pp.231-242.

Klavans, J. L. & Ph. Resnik (eds.) (1996): *The Balancing Act. Combining Symbolic and Statistical Approaches to Language*. MIT Press, Cambridge (Mass.) – London (England).

Knill, K.& S. Young, (1997): Hidden Markov Models in Speech and Language Processing, in: Young, S. & G. Bloothooft (eds.): *Corpus-based Methods in Language and Speech Processing*. Kluwer, Amsterdam, pp. 27-68.

Krieger, H.-U. & J. Nerbonne, (1991): *Feature-Based Inheritance Networks for Computational Lexicons*. Saarbrücken, Deutsches Forschungszentrum für Künstliche Intelligenz, Research Report 31.

Mancini, F. (1993): L'elaborazione automatica del corpus, in: De Mauro, T., F. Mancini, M. Vedovelli & M. Voghera: *Lessico di frequenza dell'italiano parlato*. Etaslibri, Milano, pp. 54-85.

Marinelli, R.(1997): Evoluzione delle procedure e dei supporti magnetici per la gestione di un archivio di dati: l'esempio dei nastri dell'OVI nell'archivio dell'ILC, Opera del Vocabolario italiano. *Bollettino* II, Firenze, pp. 251-258.

Pollard, C. & I. A. Sag, (1987): Information-Based Syntax and Semantics, Stanford, Stanford University Center for the study of language and information. *CSLI lecture notes* 13.

Renzi, L.(1987): *Nuova introduzione alla filologia romanza*. Con la collaborazione di Giampaolo Salvi, Il Mulino, Bologna, (rifacimento di IDEM, *Introduzione alla filologia romanza*, Il Mulino, Bologna, 1976).

Renzi, L. (ed.), (1998): *ITALANT: per una grammatica dell'italiano antico*. Centrostampa Palazzo Maldura, Padova.

Renzi, L. (1998a): Perché una grammatica dell'italiano antico: una presentazione, in: Renzi, L. (1998), pp. 21-32.

Renzi, L. (1998b):Premessa, in: Renzi, L. (1998), p. 7.

Schmid, H. (1994): *Probabilistic Part-of-Speech Tagging Using Decision Trees*, paper presented at the International Conference on New Methods in Language Processing, Manchester (UK), 1994; revised postcript version available on the FTP server of Stuttgart IMS (URL: http://www.ims.uni-stuttgart.de/).

Schmidt, P.(1993): *Basic Decisions on Type and Feature System and on Core Formalism*. Deliverable DI, 1993, Commission of EEC.

Riassunto

Per chi si occupa di sintassi poter interrogare corpora elettronici per sequenze di parti del discorso è molto più utile che non l'interrogare per sequenze di forme o lemmi. Dal momento che *ITALANT* si propone soprattutto come studio sintattico dell'italiano antico, il gruppo torinese del progetto ha cominciato ad annotare morfosintatticamente il *Padua Corpus*. Le caratteristiche del corpus e il fatto di poter acquisire informazioni utili per l'annotazione da una preesistente base di dati (GATTO, Gestione degli Archivi Testuali del Tesoro delle Origini) hanno fatto scegliere un annotatore basato su regole. L'insieme di annotazioni (*tagset*) adottato intende soddisfare gli standard richiesti per entrare nel virtuale insieme di corpora della ricerca internazionale annotati secondo le indicazioni EAGLES.