L''Ancient Greek Dependency Treebank'. Un nuovo strumento per lo studio della lingua greca

Nel novembre 2009 il Perseus Project ha pubblicato sul proprio sito internet la prima versione dell'Ancient Greek Dependency Treebank (AGDT)[1]. L'obiettivo dell'ambizioso progetto, che affianca l'analogo latino (Latin Dependency Treebank, LDT) già disponibile dal 2007, è quello di creare il primo treebank dei testi letterari della Grecia antica, a partire dalle opere poetiche di età arcaica e classica.

Nella linguistica contemporanea, si definisce con il termine inglese 'treebank' un *corpus* annotato in cui i testi che lo compongono, suddivisi in unità strutturali via via più piccole (token), sono arricchiti da un corredo di meta-informazioni che descrivono la morfologia delle singole parole e le loro relazioni sintattiche all'interno della frase. La natura delle annotazioni aggiunte può comprendere anche ulteriori livelli di analisi linguistica e variare a seconda degli scopi cui il *corpus* è destinato[2].

Nel caso dell'AGDT, così come per il LDT, le opere antiche sono divise in frasi seguendo la punteggiatura forte (punti, punti in alto e interrogativi) delle edizioni di riferimento[3]. Tutti i token di ciascuna frase, che corrispondono nella grande maggioranza dei casi alle singole parole[4], sono descritti, oltre che da una serie di identificatori univoci, da cinque etichette di analisi morfologica e sintattica. Esse comprendono: la forma attestata nel testo, il lemma a cui la parola si riferisce, un codice che ne descrive in dettaglio la morfologia, la testa da cui la parola dipende e la relazione sintattica implicata nella dipendenza[5]. Al momento in cui scriviamo, il treebank, liberamente accessibile e scaricabile dal sito internet del progetto, comprende un totale di 309.096 parole annotate, che corrispondono al testo integrale di alcune opere di poesia esametrica (*Iliade* e *Odissea*; *Teogonia*, *Opere e Giorni*, *Scudo di Eracle* di Esiodo) e tragica (le sette tragedie tradizionalmente attribuite ad Eschilo, *Aiace* di Sofocle). Le edizioni utilizzate per la creazione del *corpus* sono quelle già riprodotte nella biblioteca

1 Accessibile all'indirizzo: http://nlp.perseus.tufts.edu/syntax/treebank/index.html.

2 Per una discussione più dettagliata sui *corpora* annontati cf. *infra* § 1.1; per l'annotazione semantica e pragmatica dei *corpora* cf. quanto detto al § 2.2.

3 Cf. n. 6 per un elenco.

4 L'eccezione più rilevante, anche a causa dell'importanza strutturale delle congiunzioni coordinanti nel sistema di annotazione adottato (su cui cf. *infra*, § 1.1 e n. 25), è rappresentata dalle congiunzioni composte come οὔτε o μήτε, che sono divise in due token. Sfortunatamente, tale separazione delle congiunzioni non è stata ancora resa sistematica: εἴτε, per esempio, o i casi di crasi con καί costituiscono ancora un singolo token.

5 La relazione sintattica è descritta attraverso l'impiego di un numero finito di tag (quali 'soggetto', 'oggetto', 'attributo', eccetera): cf. Bamman – Crane 2006, 69 per una lista. Questa forma di descrizione delle relazioni sintattiche, basata sull'individuazione dei rapporti di dipendenza diretta tra le parole di una frase, si fonda su di una particolare versione della 'grammatica di dipendenza': cf. *infra* § 1.1 per una discussione più articolata.

digitale del Perseus Project[6]. Il numero dei testi annotati che saranno pubblicati sul sito del progetto è destinato ad aumentare rapidamente, dato l'elevato numero di studenti, ricercatori e istituzioni universitarie coinvolte[7].

L'AGDT colma un vuoto nel panorama della linguistica computazionale, poiché arricchisce l'elenco dei *corpora* annotati con testi in una lingua comparativamente povera di risorse per il trattamento automatizzato del linguaggio come il greco. D'altro canto, il progetto di Perseus mette a disposizione della comunità degli studiosi della lingua greca uno strumento di lavoro supplementare. Le dimensioni e la qualità dei testi inclusi nel *corpus* già pubblicato autorizzano già da subito ad interrogarsi sui potenziali impieghi cui la risorsa potrà prestarsi anche in ambito filologico.

Allo stesso tempo, è bene ricordare un assunto acquisito della linguistica computazionale: i *corpora* annotati possono essere utilizzati con pieno profitto solo a patto che i principi strutturali e le linee guida operative seguite nella loro progettazione siano esposti con la massima chiarezza[8]. Perché i risultati di un'interrogazione, come del resto di qualunque operazione compiuta sul *corpus*, siano ben compresi dagli utenti è necessario che i principi metodologici che sottintendono alla costruzione della risorsa siano resi espliciti.

Nel caso di uno strumento destinato a oltrepassare i limiti della linguistica computazionale e dei *corpora*, la necessità della documentazione coinvolge un aspetto metodologico più ampio. L'illustrazione di alcuni dei concetti fondamentali della linguistica dei *corpora* può consentire, infatti, un uso più consapevole degli strumenti anche ad un pubblico che, come gli antichisti, è naturalmente lontano dai dibattiti metodologici di tale disciplina.

1. *Corpora* annotati e linguistica dei *corpora*

In quanto *corpora* annotati, i treebank nascono nell'alveo e dall'esperienza della *corpus linguistics*.

A partire dalla seconda metà del Ventesimo secolo, la ricerca empirica sulle manifestazioni concrete del linguaggio ha conosciuto un mutamento decisivo grazie alla creazione delle prime grandi raccolte di testi disegnate appositamente per lo studio linguistico[9]. L'impiego delle risorse informatiche ha contribuito a tale sviluppo in due

[6] Ovvero: Allen – Monro 1920 per i poemi omerici; Evelyn-White 1914 per Esiodo; Smyth 1922 per Eschilo; Jebb 1896 per l'*Aiace* di Sofocle.

[7] Il sito internet del progetto (http://nlp.perseus.tufts.edu/syntax/treebank/contributors.html) menziona più di duecento persone coinvolte. Dato che l'intero processo di annotazione avviene per mezzo di software a interfaccia web, i lettori e gli utenti sono esplicitamente invitati a collaborare contattando i responsabili.

[8] Cf. Leech 2004.

[9] Progetti che vengono di solito menzionati come pionieri in questa vicenda sono il Brown Corpus (Francis – Kucera 1967, 1964), pensato specificamente per lo studio linguistico dell'inglese contemporaneo, e l'Index Thomisticus (Busa 1974-80), orientato piuttosto allo studio dell'opera di Tommaso D'Aquino. Per un agile e utile orientamento su *corpus linguistics*, *computational linguistics* e *digital humanities*, cf. Zeldes – Lüdeling 2007; più in dettaglio cf. anche McEnery – Wilson 2001.

aspetti. Da un lato, l'evoluzione tecnologica ha permesso di immagazzinare e processare una quantità di dati sempre più ampia. Dall'altro, una disciplina come la linguistica computazionale, orientata al trattamento automatizzato del linguaggio e al raggiungimento di obiettivi operativi, tra cui soprattutto la traduzione automatica, ha favorito l'applicazione di metodi matematico-statistici per processare i materiali. Parallelamente alla crescita delle dimensioni del materiale raccolto e delle possibilità di impiego, ha preso piede il dibattito teorico, due aspetti del quale, in particolare, possono essere qui ricordati.

Un primo nodo cruciale è il concetto di 'rappresentatività'. Un *corpus*, infatti, aspira ad essere un microcosmo capace di riflette in scala ridotta le caratteristiche generali di una lingua o di un suo particolare sottoinsieme. Le opzioni metodologiche che guidano la scelta dei testi (scritti o trascrizioni di comunicazioni orali), delle fonti cui attingere i materiali, delle dimensioni dell'intera collezione e delle sue singole parti rivestono un'importanza fondamentale. Questioni in apparenza puramente pratiche, quali la preferenza per testi già digitalizzati o la selezione di materiale a seconda del diritto d'autore, possono spesso avere ricadute impreviste sulla qualità delle risorse prodotte[10].

Un secondo elemento del dibattito sui *corpora* è di natura più prettamente teorica e riguarda l'uso delle collezioni come strumenti della ricerca. Se è ovviamente possibile che linguisti appartenenti alle più disparate scuole ricorrano sporadicamente ai *corpora* a fini esemplificativi, collezioni disegnate secondo solidi criteri teorici possono consentire approcci empiristici anche più radicali; esse, comunque, invitano il linguista a ripensare il rapporto tra teoria e dati nella ricerca.

In un'importante messa a punto Tognini-Bonelli[11] ha distinto due diverse tipologie operative fondamentali, a loro volta molto articolate nelle diverse possibili realizzazioni, di ricorso ad un *corpus*: una modalità denominata *corpus-based* e un approccio *corpus-driven*.

Nel primo caso la riflessione teorica precede e orienta l'accesso ai dati empirici; il *corpus* rappresenta il banco di prova attraverso il quale la teoria viene verificata e, nel caso, integrata, modificata o rigettata[12].

La seconda metodologia, invece, aspira ad una maggiore neutralità teorica. L'astrazione e la spiegazione dei fenomeni non sono indipendenti dal materiale empirico, bensì vengono inferite direttamente dai dati: intere grammatiche possono essere indotte processando i *corpora* attraverso approcci statistici e matematici. In termini operativi, le due modalità di concepire il rapporto fra dato fenomenico e teoria sono anche definiti, sulla base della direzione della ricerca che procede dall'astratto al concreto o viceversa, rispettivamente *top-down* e *bottom-up*.

[10] Per una buona introduzione si veda Sinclair 2004. Il caso, lì discusso, della Bank of English è esemplare per le ricadute scientifiche che scelte motivate da fattori editoriali o di copyright possono avere.

[11] Tognini-Bonelli 2001.

[12] In altri casi, invece, diverse spiegazioni concorrenti dello stesso fenomeno linguistico possono essere statisticamente misurate nella loro capacità di dar conto dei dati; in analisi quantitativa è il cosiddetto metodo della multifattorialità: cf. Gries 2003 per un esempio sull'inglese.

Entrambi gli approcci hanno in comune un'istanza metodologica fondamentale, nella misura in cui, sia per chi opera *top-down* sia per chi segue la direttrice opposta, l'interpretazione dei dati dei *corpora* non è una fase opzionale o una verifica supplementare, bensì rappresenta la componente fondamentale del lavoro. Lo statuto metodologico della *corpus linguistics*, intesa come disciplina che riconosce al lavoro sui *corpora* un ruolo centrale, è dibattuto dagli stessi studiosi che, a vario titolo, vi si riconoscono. Recentemente, tuttavia, Gilquin[13] ha individuato almeno due fattori che unificano i diversi approcci che ambiscono a riconoscersi all'interno di tale disciplina. Uno di essi è la tendenza a enfatizzare l'analisi quantitativa, oltre che qualitativa, dei fenomeni. Il secondo, più importante ancora, è un impegno al rispetto dell'integrità dei dati, ovvero a rendere conto il più possibile, in fase di riflessione teorica, della totalità dei fenomeni osservati nel *corpus*, senza operare selezioni arbitrarie di sottoinsiemi.

Benché forse approcci di tipo diverso diverranno possibili e praticati con l'aumentare (in quantità e qualità) delle risorse disponibili, un'applicazione di metodologie *corpus-based* ai testi latini e greci appare a chi scrive già fin da oggi del tutto percorribile[14]. Possediamo, infatti, una ricca tradizione grammaticale, normativa e storica, i cui assunti teorici possono essere arricchiti, rivisti o corretti attraverso l'interrogazione dei *corpora* annotati[15]. Da un punto di vista metodologico, inoltre, la concezione che vede nella linguistica teorica e nella critica testuale dei classici due ambiti indipendenti, ciascuno dotato del suo campo d'indagine specifico, ma posti in relazione di mutua collaborazione, data agli albori della *grammatiké techne* occidentale[16].

Il problema della rappresentatività del *corpus*, invece, cela implicazioni più complesse.

Da un lato, si può facilmente arguire che l'insieme delle opere letterarie greche e latine sopravvissute rappresenta un *corpus* chiuso, prodotto di un processo di canonizzazione secolare o di eventi traumatici che hanno determinato una selezione a volte casuale dei testi conservati. L'aspirazione ad includere la totalità della produzione, dati dei limiti cronologici su cui si può discutere, è dunque legittima in questo caso e, da sola, sufficiente a risolvere la questione della rappresentatività del *corpus*. Quanto il *corpus*, a sua volta, sia rappresentativo del fenomeno linguistico in questione (la produzione letteraria in lingua greca antica) è un problema che non concerne il lavoro di costruzione del *corpus* linguistico in particolare, quanto piuttosto la scienza filologica in generale[17].

[13] Gilquin 2010.

[14] Cf. anche *infra*, § 2.1.

[15] Così Passarotti 2009.

[16] Apollonio Discolo (1.1-5 ed. Lallot), nel primo paragrafo del suo trattato sulla costruzione delle parole, afferma di intraprendere lo studio della *syntaxis* proprio perché essa appare uno strumento indispensabile per l'esegesi dei poeti (Lallot 1997).

[17] A questo punto, tuttavia, si pone il difficile problema delle opere sopravvissute in forma frammentaria, che sono testimonianza di incalcolabile valore, in particolare per certi generi e per determinate forme letterarie o epoche (si pensi alla lirica arcaica). La loro integrazione in un *corpus* accanto alle opere sopravvissute in forma integrale pone notevoli problemi teorici e pratici (Berti

Il problema del canone degli autori o delle opere antiche da riprodurre, naturalmente, non esaurisce il problema della rappresentatività del *corpus* di quei testi particolari che sono le opere letterarie antiche sopravvissute fino ai giorni nostri. Nel processo secolare di trasmissione, questi testi sono stati riprodotti ed editi in una moltitudine di formati e vesti differenti. Quale delle molte edizioni di un testo antico dovrà essere riprodotta nel *corpus*? Si dovranno privilegiare questioni pratiche di diritto di accesso e di diffusione, o si dovranno piuttosto scegliere differenti edizioni per ogni singolo autore in base al loro 'valore' scientifico, con evidenti rischi di soggettività?

La questione, in apparenza molto pratica, cela un problema teorico di fondo. Se in linguistica è naturale considerare un *corpus* come una semplice collezione di testi rappresentativi di un dato linguaggio, tale definizione non può essere sufficiente dal punto di vista di un filologo che si occupi di opere dell'antichità greco-romana. La nozione stessa di 'testo', in questa prospettiva, appare problematica.

A ben guardare, se la linguistica dei *corpora* per le lingue contemporanee privilegia l'aspetto sintagmatico, ovvero la sequenza delle diverse manifestazioni linguistiche incluse, dal punto di vista di un filologo su ognuna delle posizioni di quella sequenza possono insistere le diverse forme che il testo ha conosciuto nella storia della sua tradizione[18]. In altre parole, per utilizzare la terminologia di de Saussure[19], il *corpus* delle opere antiche non esiste solamente sul piano sintagmatico, ma anche su quello paradigmatico. L'oggetto del lavoro di un filologo non può limitarsi ad un testo, ma deve prendere in considerazione un'intera tradizione testuale.

Per tale ragione, la scelta di un testo di riferimento (sia essa fondata su motivi contingenti, sia essa dettata da criteri scientifici o dal prestigio di cui una particolare edizione gode al momento della creazione del *corpus*) non potrà che essere effimera, recando la chiara impronta del luogo, dell'epoca e della personalità di chi ha operato la selezione.

Un *corpus* linguistico 'tridimensionale', che includa cioè anche la dimensione della tradizione testuale ed ermeneutica degli artefatti che lo compongano, è ancora un prodotto in via di elaborazione, tanto in termini teorici, quanto nei dettagli operativi necessari per la sua concretizzazione[20]. L'aggiunta di annotazione linguistica (su cui

et Al. 2009; Romanello *et Al.* 2009). Curiosamente, la questione, che pare così specifica della filologia classica, non è del tutto ignota a chi progetta *corpora* di lingue moderne, in particolare laddove è previsto l'uso di opere letterarie per le quali i detentori del copyright autorizzano solamente una riproduzione parziale. In simili casi, ragioni d'uso e di rappresentatività consigliano, secondo Sinclair 2004, di privilegiare l'integralità del testo: nel limite del possibile è più opportuno che i *corpora* siano costituiti da testi interi, poiché l'integralità del messaggio costituisce un valore aggiunto in sé. Si tratta, a mio avviso, di un criterio ragionevole anche per i *corpora* di lingue antiche, per lo meno in fase preliminare; solamente, di nuovo, è assolutamente necessario essere chiari su tale opzione.

[18] Fondamentale, per i paradigmi della scienza filologica, il rinvio a Pasquali 1952.

[19] de Saussure 1916.

[20] Di grande interesse in questo senso è l'Homer Multitext Project, promosso dal Center for Hellenic Studies, su cui si veda più di recente Smith 2010. Per il Digital Aeschylus cf. Boschetti 2009. Il progetto Musisque Deoque (http://www.mqdq.it), *corpus* di poesia latina con apparati critici compilati manualmente da collaboratori, consente di comprende anche le varianti nelle ricerche effettuate.

vedi § 1.1) eleva naturalmente al quadrato i problemi, poiché comporta l'aggiunta di un apparato interpretativo ad un *corpus* testuale già stratificato e storicamente complesso.

Non stupisce, pertanto, che le fondamenta per un treebank dei testi greci e latini siano state gettate in accordo all'abituale prassi di progettazione dei *corpora*, a partire, cioè, da un testo di riferimento scelto nel panorama delle edizioni critiche del XIX e XX secolo guardando in particolare alle esigenze di riproducibilità e copyright. La necessità di un approccio originale alla natura storicamente 'aperta' del lavoro esegetico e critico sui testi antichi, ben chiara ai curatori del progetto, viene, in questa fase iniziale, affrontata su di un altro terreno, quello dell'organizzazione del lavoro di annotazione[21].

1.1 Treebank e annotazione dei corpora

Il materiale grezzo raccolto nei *corpora* presenta alcuni svantaggi che ne limitano le potenzialità di impiego. In primo luogo, esempi di omografia o ambiguità lessicale possono compromettere il trattamento (soprattutto automatizzato) dei dati[22]. Un simile inconveniente può essere aggirato aggiungendo a ciascun elemento del *corpus* informazioni linguistiche fondamentali. I tipi di analisi che possono essere codificati attraverso metadati aggiunti ai token possono appartenere ai più diversi livelli linguistici, dalla fonetica alla semantica[23]. Storicamente, probabilmente anche a causa della natura tipologica dell'inglese che rende casi di omografia come quello citato più sopra particolarmente frequenti, la classificazione delle parole in termini di 'parte del discorso' (*part of speech*, *POS*) è stata prioritaria.

Un treebank, come il nome stesso (grosso modo: 'banca dati di alberi') suggerisce, comporta l'integrazione di un ulteriore livello di analisi. In questo tipo di *corpora* anche le relazioni sintattiche tra i costituenti della frase sono codificate e descritte attraverso l'annotazione.

L'aggiunta di un simile apparato di informazioni ripropone in forma ancora più evidente il problema, già accennato, della reciproca relazione tra teoria linguistica e *corpora*[24]. Se nemmeno il lavoro di *POS-tagging* è immune dalla necessità di compiere scelte precise (cosa e quante sono le parti del discorso, come distinguerle eccetera), è nella sintassi che le differenze di approccio possono essere più profonde, a seconda del quadro teorico di riferimento adottato per classificare e descrivere i fenomeni attraverso

[21] Cf. Bamman *et Al.* 2009, sull''ownership model' e sulla 'scholarly annotation', discussi più sotto; cf. *infra* le parole citate alla n. 35.

[22] Un esempio tratto da Dickinson – Meurers 2003: l'inglese *can* potrebbe essere tanto un sostantivo (italiano 'lattina'), quanto voce del verbo denominativo corrispondente (it. 'mettere in scatola') o del più comune ausiliare servile (it. 'potere'). La quantità di lavoro necessario per disambiguare le diverse forme rischia facilmente di essere troppo elevata.

[23] Leech 2004.

[24] Si tratta di un passo che non tutti sono, in effetti, disposti a compiere; secondo alcuni, l'annotazione introdurrebbe un'adulterazione indebita dei dati del *corpus*: cf. ad esempio Sinclair 1987. In prospettiva contraria, secondo Hajičová – Sgall 2006 l'annotazione di un *corpus* rappresenta un banco di prova per le teorie linguistiche, nonché un'opportunità per scoprire punti deboli o aspetti da migliorare nei modelli grammaticali.

un numero finito di etichette.

Nella progettazione dei treebank, si possono distinguere due modelli di rappresentazione delle relazioni sintattiche, assurti a standard *de facto* (cf. Abeillé 2003: xvi-xviii). La *constituency annotation* analizza la frase scomponendola in costituenti intermedi, che possono a loro volta ricorsivamente dividersi in altri sintagmi fino a giungere al livello delle parole. Nel modello opposto (*dependency annotation*), invece, le parole stesse sono poste direttamente in relazione di dipendenza tra di loro: secondo tale modalità di rappresentare la sintassi, un soggetto dipende direttamente dal proprio verbo, un aggettivo dal sostantivo con cui concorda, eccetera[25].

Dal punto di vista teorico, il primo approccio si apparenta alla grammatica generativo-transformazionale e all'influenza di Chomsky[26]; un punto di riferimento per la grammatica di dipendenza, invece, è l'opera di Tesnière[27].

Storicamente, l'analisi a costituenti ha dominato a lungo il campo fin dai primi progetti, poiché essa ben si adatta a descrivere una lingua caratterizzata da un ordine dei costituenti rigido e dalla scarsa morfologia flessionale come l'inglese. Viceversa, le diverse formalizzazioni della grammatica di dipendenza si rivelano più adatte a rappresentare una lingua fortemente flessa, i cui costituenti possono presentare un alto grado di discontinuità. Per tale ragione la grammatica di dipendenza è la cornice teorica favorita, oltre che da diversi progetti su lingue a noi contemporanee, anche per le lingue antiche[28].

Fra i diversi 'dialetti' della grammatica di dipendenze, l'AGDT adotta la 'Functional Generative Description' nel cui alveo, in uno stretto rapporto di interscambio tra teoria linguistica e pratica dell'annotazione, è nato il Prague Dependency Treebank per la lingua ceca[29].

[25] Tale approccio, all'apparenza molto intuitivo, non è tuttavia esente da una certa artificialità nella rappresentazione di determinati costrutti. Nel sistema di regole adottato da Perseus, ad esempio, in caso due parole siano coordinate attraverso l'uso di una congiunzione o di un segno di interpunzione, è proprio l'elemento coordinante (e nel caso di coordinazione polisindetica l'ultimo coordinante) ad assumere la funzione di testa. Nel caso, poi, che una parola sia logicamente riferita ad entrambi i *cola* coordinati (ad esempo un soggetto comune a due verbi, come nell'esempio: *Paolo studia e lavora*), anche tale parola dipenderà dalla congiunzione. Nell'esempio, *Paolo* sarà governato da *e* con funzione di soggetto.

[26] Chomsky 1957 e 1965.

[27] Tesnière 1959. Cf. per una presentazione più dettagliata, anche in relazione all'applicabilità dei modelli alle lingue antiche cf. Boschetti 2005, 53-64 e Passarotti 2009, 7 s.

[28] Cf. Bamman *et Al.* 2007 per il LDT e l''Index Thomisticus Treebank' (IT-TB), che condividono le medesime regole per l'annotazione; anche il progetto PROIEL (http://www.hf.uio.no/ifikk/english/research/projects/proiel), *corpus* parallelo di traduzioni del Nuovo Testamento nelle lingue indoeuropee, adotta la grammatica di dipendenze. Boschetti 2005, 74-80, invece, opta per un sistema misto, ispirato al NEGRA *corpus* della lingua tedesca: in esso la frase è descritta attraverso i rapporti di dipendenza tra costituenti, i quali tuttavia possono presentarsi in ordine variabile ed essere discontinui fra loro; tali costituenti sono comunque rappresentati da *phrasal nodes* intermedi, invece che dalle parole stesse.

[29] Sulla 'Functional Generative Description' il rinvio è a Sgall *et Al.* 1986; per la reciproca interdipendenza tra teoria e annotazione il già citato Hajičová – Sgall 2006; sul PDT Böhmová *et Al.* 2003.

In linea con l'interpretazione stratificata del fenomeno linguistico offerta dal paradigma generativo-funzionale, il PDT offre un'annotazione su tre livelli: uno di analisi morfologica, ed uno rispettivamente per la struttura sintattica di superficie (analitico) e per quella profonda (tectogrammaticale)[30]. I due treebank di Perseus adottano, salve poche modifiche, i medesimi tag e le medesime regole del livello analitico praghese[31]. Secondo questo formalismo, la frase è descritta come una struttura gerachica incentrata sulla funzione di predicazione; i diversi complementi vengono poi analizzati sulla base della funzione, distinguendo in particolare tra argomenti necessari e complementi circostanziali[32]; tutte le parole ricevono un'annotazione morfologica e sintattica, compresi i segni di punteggiatura e le parti del discorso che svolgono una funzione più strutturale (preposizioni e congiunzioni).

Il processo di annotazione, in particolare per *corpora* di grandi dimensioni, comporta numerosi problemi pratici che possono avere importanti conseguenze sulla natura e sulla qualità del treebank. Per i *corpora* di lingue che possono vantare più risorse, solo parte del lavoro viene generalmente svolta da annotatori umani. Per lo più, l'annotazione manuale svolge un ruolo preliminare e interessa una parte ridotta del *corpus*: l'output prodotto dagli annotatori umani funge soprattutto da *training set* su cui vengono allenati *part-of-speech taggers* o *parsers* sintattici, che concludono in forma automatizzata (o semi-automatizzata) la marcatura delle parti restanti[33].

Per quanto riguarda l'annotazione manuale, inoltre, una prassi molto comune consiste nell'utilizzare due persone per le medesime frasi, con un terzo revisore (di solito dotato di competenze più avanzate) impiegato per armonizzare le differenze e correggere gli errori. Tale flusso di lavoro (denominato *standard method*) è stato adottato anche per il LDT e per parte dell'AGDT. Nel paradigma della *corpus linguistics*, una simile organizzazione del lavoro mira a ridurre al minimo la soggettività dell'interpretazione, proponendo l'annotazione prodotta alla fine del processo come una rappresentazione oggettiva delle strutture morfo-sintattiche soggiacenti alle singole manifestazioni linguistiche.

Per il *corpus* dei tragici, caratterizzato da situazione testuale complessa e dall'elevato grado di conflittualità nelle possibili interpretazioni, i creatori hanno invece preferito sperimentare un approccio diverso. Un solo annotatore, con maggiore esperienza di studio filologico dei testi, è responsabile del lavoro di marcatura. Il risultato si configura, di conseguenza, come un'interpretazione originale e fortemente individualizzata del testo.

Il modello cui gli autori del progetto mirano esplicitamente per questo metodo (da essi chiamato *scholarly annotation*) è quello dell'edizione annotata, prodotto principe

[30] Sul livello tectogrammaticale cf. le osservazioni svolte più sotto al § 2.2.

[31] Si noti che LDT e IT-TB adottano, in uno sforzo congiunto, il medesimo tagset per il livello analitico. Per l'IT-TB, inoltre, è in corso l'annotazione semi-automatica a livello tectogrammaticale: cf. Passarotti c.s.

[32] Altro sulla valenza al § 2.3.

[33] Se per il latino dell'IT-TB è in corso di sperimentazione l'adozione di un flusso di lavoro simile (Passarotti 2009), l'annotazione sintattica del greco è ancora integralmente manuale.

della critica testuale fin dalla fondazione della Biblioteca di Alessandria. Un treebank dovrebbe, in questa prospettiva, presentarsi come un testo in cui è incorporato un commento linguistico e sintattico sotto forma di metadati, espresso nel metalinguaggio del sistema di annotazione adottato dal progetto. Creare una versione sintatticamente annotata di un testo, in altre parole, si presenta come una forma di lavoro critico, condotta con i metodi della scienza filologica e in continuo confronto con la tradizione esegetica millenaria che interessa le opere antiche[34].

In questa fase iniziale del lavoro, la differenza fra i prodotti dei due metodi non appare immediatamente apprezzabile. I curatori del progetto, tuttavia, esprimono apertura verso la possibile inclusione di una pluralità di differenti versioni di una medesima opera che la *scholarly annotation* sembrerebbe incoraggiare; non è da escludere che diverse versioni di uno stesso testo, prodotto di due annotatori differenti e discordanti, come è naturale, nella lettura di passi problematici, siano incluse nel treebank e pubblicate sulla pagina del progetto[35].

2 Gli scenari futuri: *corpus* linguistics e NLP

2.1 Treebanks e ricerca linguistica: alcuni esempi

I *corpora* annotati si sono imposti all'attenzione dei linguisti in particolare grazie a due ambiti di applicazioni in cui sono stati utilizzati, il trattamento automatizzato del linguaggio (*Natural Language Processing* o NLP) e la ricerca linguistica.

L'impiego dei *corpora* come strumento di lavoro per testare teorie precedentemente elaborate rappresenta, come si è detto (§ 1), una delle vie più battute e meglio fondate da un punto di vista metodologico della linguistica dei *corpora*.

Nell'ambito dei fenomeni sintattici, questo tipo di approccio è già praticabile sui dati pubblicati dal Perseus Project. Il treebank, in effetti, consente di raccogliere rapidamente i dati eseguendo ricerche orientate alla morfo-sintassi anche molto sofisticate. Reggenze dei verbi, tipi di predicazione (frase nominale *vs* uso della copula), realizzazione dei nessi nomi-attributo, ordine dei costituenti e delle parole sono esempi di informazioni che possono essere estraibili sulla totalità del *corpus* annotato messo a disposizione dal progetto[36].

Al momento di scrivere, tale ricerca non è ancora alla portata di chi non abbia competenze piuttosto avanzate in linguistica computazionale, benché l'obiettivo del Perseus Project per il prossimo futuro sia quello di mettere a disposizione strumenti di interrogazione accessibili a tutti gli utenti[37]. Alcuni tentativi di sfruttamento delle

[34] Bamman *et Al.* 2009.

[35] Scrivono Bamman *et Al.* 2009, 13: «by publicly releasing the data with citable attributions of ownership, we hope to provide the core around which other interpretations of the data can be layered – a scholar who disagrees with a single annotation decision need not start from scratch to contribute a new annotation, but can simply build on the existing data and change only the elements subject to debate.»

[36] Alcuni esempi di analisi del *word order* latino sono illustrati da Bamman – Crane 2006, 72-6.

[37] Cf. Bamman – Crane 2007, 38.

risorse pubblicate fino ad ora per studi *corpus-based*, tuttavia, sono già possibili.

Fenomeni prettamente sintattici, molto rilevanti per l'interpretazione di passaggi o di intere opere, possono essere indagati con profitto attraverso un'interrogazione del treebank. È il caso, per esempio, dell'alternanza fra la frase nominale (del tipo frequente nelle iscrizioni vascolari: ὁ παῖς καλός) e la predicazione per mezzo della copula, su cui esiste abbondante letteratura[38].

I dati che possono essere estratti in relazione a questo fenomeno nel *corpus* delle tragedie di Eschilo, per esempio, paiono rilevanti tanto da un punto di vista linguistico quanto da quello filologico. La complessa predicazione nominale di Aesch. *Eum*. 381-6[39] è apparsa sospetta, anche perché essa viola la tendenza ad esprimere sempre il pronome soggetto nelle frasi nominali ogni qual volta esso è rappresentato da un pronome di prima o seconda persona[40]; la struttura sintattica è alterata da Sommerstein (1989), che stampa la congettura di Heath διέπομεν per il tradito διόμεναι.

Eppure i dati ricavabili dal treebank ci informano che è proprio con gli aggettivi composti di δυσ- e εὐ- (rispettivamente 11 contro 5 e 21 contro 10), nonché con gli aggettivi che, come μνήμονες κακῶν e δυσπαρήγοροι βροτοῖς, governano un complemento (37 contro 20) che Eschilo predilige l'uso della costruzione nominale rispetto alla copula[41]. Il testo tradito, dunque, pare rispondere ad un *usus* stilistico ben visibile nel *corpus*; tra le tante difficoltà del testo, non sembra prudente alterare la costruzione sintattica.

La frequenza della costruzione nominale con i composti di δυσ- e εὐ- o con aggettivi in grado di governare complementi è già stata intuita dalle grammatiche. L'interrogazione di un *corpus* annotato, tuttavia, permette di verificare molto facilmente, e di misurare con precisione, l'esistenza di particolari tendenze stilistiche.

Anche da un punto di vista linguistico più generale il *corpus* di Eschilo offre dati interessanti sul fenomeno. La teoria secondo cui esisterebbe un legame prioritario, anche in termini cronologici, tra la frase nominale e la predicazione impersonale e atemporale[42], avanzata forse sulla base di una lettura un po' frettolosa di Benveniste[43], appare poco persuasiva alla luce della chiara tendenza in Eschilo ad usare la frase

[38] Per il greco e le lingue indoeuropee si vedano almeno Meillet 1906-08, Benveniste 1950, Guiraud 1962.

[39] Il testo riprodotto ed annotato nel treebank è il seguente: Μένει γάρ. εὐμήχανοί / τε καὶ τέλειοι, κακῶν / τε μνήμονες σεμναὶ / καὶ δυσπαρήγοροι βροτοῖς, / ἄτιμ' ἀτίετα διόμεναι λάχη / θεῶν διχοστατοῦντ' ἀνηλίῳ λάμπᾳ; il passo è tormentato da una serie di difficoltà e da almeno una sicura corruttela, testimoniata dalla mancata responsione del testo tradito della strofe con l'antistrofe; il testo di Smyth, su cui si basa il treebank, non rappresenta certamente la migliore soluzione a queste difficoltà: cf. Sommerstein 1989, 147-50.

[40] Ma la costruzione senza pronome soggetto, benché chiaramente in declino, è ben attestata come tratto sintattico arcaico nelle lingue indoeuropee più conservative: cf. Meillet 1906-08.

[41] Si noti poi che grazie al treebank è possibile ottenere facilmente il quadro complessivo di un fenomeno con cui comparare le cifre delle specifiche sotto-classi che interessano: nel caso in questione, il *corpus* fornisce 342 (57%) frasi nominali contro 242 (43%) casi di uso della copula.

[42] Cf. per esempio le conclusioni di Moorhouse 1982, 340-2 sui dati di Sofocle.

[43] Benveniste 1950.

nominale di preferenza quando il soggetto è il pronome deittico ὅδε (25 casi contro 11).

Tale tendenza assume significato, a mio parere, se interpretata sulla base della più generale definizione della funzione verbale offerta da Benveniste[44]; accanto ad un ruolo coesivo (che fornisce coerenza sintattica alla frase), la predicazione svolge anche una funzione assertiva, ovvero di realtà. Ora, nella frase nominale atemporale e impersonale un elemento diverso dal verbo, e perciò privo di marche morfologiche di tempo, aspetto e persona, può assumere il ruolo di predicato; l'intelligibilità del discorso o la sua coerenza sintattica non ne risultano comunque compromesse. Il pronome deittico soggetto di una proposizione nominale, d'altro canto, potrà facilmente rimpiazzare il verbo nell'esercizio dell'asseverazione assertiva della proposizione stessa[45]. Una lettura del fenomeno della frase nominale attraverso la sola lente dell'opposizione tra predicazione generale *vs* predicazione individuale[46] non può, viceversa, rendere conto a pieno di questi dati.

Benché svolte su di un *corpus* tutto sommato limitato, tali osservazioni sono state rese possibili, almeno a chi scrive, solo dal ricorso al treebank. La facilità e la rapidità con cui i dati dell'AGDT possono essere interrogati consente, infatti, di elaborare spiegazioni sempre più idonee a rendere conto dei problemi studiati, che possono successivamente essere di nuovo testate sul *corpus*. Ugualmente, la possibilità di estrarre da un *corpus* tutte le attestazioni rilevanti di un determinato costrutto, come la frase nominale, permette di osservare il fenomeno da una varietà di angolazioni, così da correggere precedenti approcci teorici parziali o da suggerire nuovi percorsi di ricerca.

2.2 Livelli supplementari di annotazione

Lo spettro di fenomeni che è possibile indagare attraverso l'uso di un treebank è piuttosto ampio. Esso non è, tuttavia, sufficiente ad esaurire la complessità dei fattori invocati dalle teorie linguistiche: raramente la sintassi, isolata dal resto dei livelli di analisi, come la semantica e la pragmatica, è in grado di rendere conto degli aspetti più importanti e studiati della lingua[47]; per alcuni fenomeni, inoltre, il limite della frase risulta troppo costrittivo[48].

L'uso sempre più sofisticato di architetture che conservano separatamente i dati dalle annotazioni (*stand-off markup*), consente di moltiplicare a piacimento i livelli di analisi,

[44] Benveniste 1950, 154.

[45] Cf. la formulazione di Benveniste 1950: «À la relation grammaticale qui unit les membres de l'énoncé s'ajoute implicitement un "cela est!" qui relie l'agencement linguistique au système de la réalité». Va da sé che un deittico è perfettamente in grado di aggiungere il «cela est!» necessario alla predicazione.

[46] Cf. Guiraud 1962.

[47] Si prenda ad esempio il già citato problema dell'ordine dei costituenti. Partendo da Erodoto e dalle sezioni giambiche delle tragedie di Sofocle, Dik 1995; Dik 2007 ha sostenuto l'ipotesi che il *word order* in greco sia legato alla struttura informativa della frase (articolazione *topic-focus*). Un approccio *corpus-based* a una simile teoria è solo parzialmente aiutato dall'uso di un treebank.

[48] Cf. per esempio il problema della deissi e dell'anafora, sulla cui rilevanza per l'interpretazione di testi letterari antichi, spesso destinati alla *performance*, si veda ora Edmunds 2008.

anche conflittuale, che possono essere collegate ai dati non interpretati del *corpus*[49].

L'esempio di una delle collezioni annotate più antiche della lingua inglese, il Penn Treebank[50], può mostrare come differenti annotazioni possano operare sui medesimi dati. Recentemente il *corpus*, composto da articoli tratti dal *Wall Street Journal*, è stato il punto di partenza per lo sviluppo di altri progetti indipendenti volti a integrare nuovi tipi di informazione.

Il Penn Discourse Treebank ha implementato un approccio lessicale per marcare le relazioni discorsive tra frasi e proposizioni[51]. Connettivi esplicitamente (tramite l'uso di particelle inferenziali) o implicitamente realizzati vengono marcati nel testo ad un livello che, di necessità, supera la frase. Il *corpus* PropBank, invece, integra l'annotazione sintattica del Penn Treebank con una marcatura completa degli argomenti verbali, che vengono distinti per ruoli semantici[52].

Anche a causa della lunga tradizione di ricerca grammaticale sulla lingua greca, è chiaramente interesse della comunità scientifica ampliare il livello della sintassi con markup relativi ad ulteriori fenomeni linguistici.

In questo senso, in virtù della parentela di fondo tra AGDT e PDT di cui si è detto, un possibile riferimento potrebbe essere rappresentato dall'annotazione tectogrammaticale del treebank praghese.

Come il livello analitico, anche il tectogrammaticale rappresenta le relazioni tra i componenti in termini di gerarchie e rapporti di dipendenza incentrati sul predicato. Esso, tuttavia, intende riflettere la struttura dei significati della frase; i token, di conseguenza, vengono annotati con una serie di attributi che descrivono gli aspetti semantici e pragmatici.

A differenza di un albero analitico, dove la frase mantiene una fisionomia piuttosto ben riconoscibile, la rappresentazione tectogrammaticale non riflette specularmente la frase come essa appare nella realizzazione di superficie del testo. Se a livello analitico ad ogni parola (ivi comprese le parti del discorso puramente funzionali, come congiunzioni, preposizioni, ausiliari e persino la punteggiatura) corrisponde un nodo, nella struttura profonda solo i termini dotati di una vera e propria autonomia semantica vengono mantenuti. Alcuni nodi vengono così soppressi: congiunzioni, preposizioni, ausiliari e verbi servili[53]. Allo stesso modo, nuovi nodi possono essere introdotti. Il caso più frequente è quello degli elementi sottintesi nelle costruzioni ellittiche, sia che essi

[49] Si noti, tuttavia, che, a differenza del PDT e dell'IT-TB che adottano una pura annotazione *stand-off*, ripartendo il materiale tra quattro livelli di informazione (uno per il testo, uno rispettivamente per l'analisi morfologica, analitica e tectogrammaticale), l'AGDT segue la modalità opposta (*inline markup)*. Il treebank greco offre il testo e tutti i livelli di annotazione in un unico file.

[50] Taylor *et Al.* 2003.

[51] Prasad *et Al.* 2008.

[52] Palmer *et Al.* 2005.

[53] La presenza di un verbo servile viene registrata attraverso uno speciale grammatema (vedi più avanti nel testo) che identifica la modalità deontica (*deontmod*) e può avere 7 valori diversi (tra cui necessità, obbligatorietà, volontà, possibilità); tale attributo caratterizza il verbo che, in superficie, è retto dal servile.

siano integrabili dal contesto, sia che invece non lo siano[54].

Ogni nodo viene definito da una serie complessa di attributi, che in particolare riportano tre tipologie di informazioni.

I *grammatemi* codificano gli aspetti semantici veicolati dalla morfologia delle parole. Alcuni di essi (ad esempio genere e numero per nomi e aggettivi; tempo e modo per i verbi) hanno una precisa corrispondenza con determinate categorie morfologiche. In questo caso, il valore espresso dalla flessione delle parole viene riprodotto nel nodo tectogrammaticale, con alcune eccezioni[55]. Altri, come la modalità deontica (cf. n. 52) o l'aspetto dei verbi (che può assumere due valori: progressivo/imperfettivo e complesso/perfettivo), non corrispondono necessariamente a precise categorie morfologiche.

Una lista di una quarantina di *functori*, inoltre, descrivono le relazione di dipendenza fra i nodi in termini di ruoli semantici. Ad esempio, il ruolo di *attore* definisce, in particolare, l'origine umana e non umana di un'azione o di uno stato di cose: a livello di sintassi superficiale, corrisponde al primo argomento di un verbo, ovvero al soggetto di un verbo attivo, all'agente di un predicato passivo e a un genitivo soggettivo. Un terzo gruppo di codici descrive la struttura informativa della frase, classificando i nodi in base ai fattori di *contextual-boundness* e contrastività. Dalle tre diverse realizzazioni (*t*: contextually-bound non contrastive; *c*: contextually-bound contrastive; *f*: contestually-non bound) può essere dedotta l'articolazione in *topic* e *focus*[56].

Infine, un'ultima serie di attributi permette di sciogliere le anafore e le co-referenze all'interno della frase e a livello di discorso; ciò si traduce, nella visualizzazione, in un collegamento che unisce i nodi che condividono il medesimo referente.

La costruzione del treebank praghese prevede strumenti automatici e semi-automatici per l'estrazione degli alberi tectogrammatici dal livello analitico[57]. Di conseguenza, la transizione tra le due strutture risulta una strada percorribile con uno sforzo relativamente contenuto.

2.3 Treebanks e lessicografia

Le applicazioni dei treebank nell'ambito lessicografico sono molteplici. Storicamente, uno dei primi benefici arrecati dai *corpora*, annotati e non, è stata la possibilità di

[54] Per un esempio di nodo lessicale integrabile a partire dal contesto cf. la frase: *Giovanni visita Maria, Paolo Lucia*; Nell'albero tectogrammaticale il nodo verbale *visita* sarebbe duplicato. Il secondo gruppo, in cui invece non è possibile estrarre nessun elemento preciso dal contesto, è rappresentato ad esempio da frasi nominali quali: *perché tanta fretta?* In questo caso un nodo identificato da un'etichetta funzionale (predicato), ma privo di un preciso lemma lessicale, prenderebbe il posto del predicato.

[55] Ad esempio, il grammatema di numero dei *pluralia tantum* riflette l'effettiva 'quantità' degli oggetti menzionati (es. lat. *castra*: si tratta di un accampamento di un esercito, o sono menzionati più accampamenti?), anziché il numero grammaticale.

[56] Hajičová – Sgall 2006.

[57] Böhmová *et Al.* 2003, 117-9.

estrarre frequenze d'uso e informazioni sulla collocazione di ogni lessema[58]. Ma è sul comportamento sintattico delle parole che l'apporto di un treebank può rivelarsi decisivo.

A partire dalla tendenza di determinati lemmi semanticamente complessi a combinarsi in locuzioni più o meno fisse è possibile, infatti, estrarre automaticamente un numero di informazioni rilevanti sulle diverse classi di significati delle parole. In questa direzione, il Perseus Project ha già intrapreso ricerche per la creazione di un lessico dinamico del latino, destinato ad integrarsi a pieno con la biblioteca digitale del progetto[59].

Utilizzando i testi annotati per allenare *parser* e *tagger* che siano in grado di marcare anche il resto del *corpus*, i responsabili del lessico dinamico hanno potuto identificare i lemmi che tendono a co-occorrere nelle medesime strutture sintattiche, classificandoli per tipologia di relazione. Tra le coppie nome-attributo risaltano locuzioni idiomatiche quali *res publica*, *patres conscripti* o *bellum civile*. Concentrandosi sulla coppia predicato-oggetto, invece, è possibile estrarre i più frequenti complementi di ogni verbo, in modo tale da fornire un'illustrazione dei più importanti usi. Nel caso dei verbi di significato più generico (come *ago*, *gero*, *do*), inoltre, questo approccio permette di ottenere i materiali necessari ad operare almeno una classificazione orientativa dei diversi sensi del lemma[60].

Quel che appare più promettente in un simile approccio è proprio il carattere 'dinamico' del lavoro lessicografico, come definito dai responsabili del progetto. I risultati ottenuti si riferiscono ad un *corpus* che comprende solo un numero limitato di opere rispetto al potenziale dell'intera latinità. L'aggiunta di nuovo materiale e nuovi testi, appartenenti o meno al canone del latino classico, aggiornerà i risultati, ma non comporterà ulteriore lavoro manuale ai creatori del lessico. Allo stesso tempo, l'utente può essere messo in grado di adattare il *corpus* di riferimento alle proprie esigenze, limitando il lessico entro un preciso arco cronologico o ad un certo novero di autori.

Un simile compito di individuazione delle diverse sotto-categorie di significati dei lemmi verbali sulla base della reggenza sintattica può essere ulteriormente sviluppato grazie ai lessici di valenza.

Con valenza si intende abitualmente il numero di argomenti che determinate classi di parole (verbi, aggettivi, nomi) sono in grado di reggere. Nel caso dei verbi, la valenza abbraccia tutti gli attanti che partecipano all'azione verbale: soggetto, oggetti diretti e indiretti, complementi predicativi.

La nozione di *oggetto* di un verbo copre, in questo contesto, uno spettro più vasto rispetto a quello di complemento oggetto di un predicato transitivo, tradizionalmente riservato dalle grammatiche scolastiche. In analogia alla distinzione fra *actants* e *circonstants* operata da Tesnière[61], essa si estende anche a quei complementi indiretti

[58] Cf. Sinclair 1987 per il progetto COBUILD.
[59] Bamman – Crane 2008.
[60] Cf. le tabelle in Bamman – Crane 2008, 15.
[61] Tesnière 1959, 102.

che sono richiesti per completare necessariamente la costruzione del verbo[62]. Nei treebank del Perseus Project e nell'IT-TB, come già nel PDT, la medesima distinzione è operata in sede di annotazione. In particolare, i complementi del verbo che non possono essere omessi senza che la frase diventi agrammaticale o che il senso del verbo cambi chiaramente, ricevono l'etichetta di *object*, i complimenti circostanziali quella di *adverbial*.

Un treebank a dipendenze permette di estrarre con precisione e completezza le informazioni relative alla valenza, distinguendo altresì i diversi costrutti del medesimo lemma. Un buon esempio di una simile applicazione è rappresentato dall'IT-Valex creato a partire dall'Index Thomisticus Treebank[63]. Per ogni lemma verbale è possibile visualizzare il numero degli argomenti governati e l'ordine usuale in cui ogni specifica costruzione è attestata. Ricerche complesse, sulla base del lemma del predicato e dei complementi governati, dei casi retti o della relazione (oggetto, soggetto, complemento predicativo) degli argomenti, sono possibili sul *corpus* delle opere annotate del filosofo aquinate. Infine, tutti passi dove è attestato il costrutto selezionato vengono visualizzati, con i diversi argomenti opportunamente evidenziati per una più immediata fruizione.

Una simile organizzazione del materiale lessicale può essere ulteriormente ampliata e resa più sofisticata grazie al contributo delle informazioni estratte da altri livelli di annotazione. Un esempio di una strutturazione più complessa può essere tratto dal PDT-Vallex, il lessico di valenza del Prague Dependency Treebank. Nella sua più recente versione, grazie alla presenza dell'annotazione tectogrammaticale (su cui cf. quanto detto al § 2.2), esso consente di operare una classificazione su diversi usi del lemma non solo in relazione al numero degli argomenti retti o alla relazione sintattica, ma anche in base ai ruoli semantici dei diversi complementi.

Il lemma *dosáhnout*, per esempio, può reggere due o tre argomenti. E tuttavia, i diversi ruoli semantici dei complementi possono aiutare a distinguere ancora le attestazioni della costruzione più comune in tre sottogruppi, a seconda che il verbo governi, oltre al soggetto-agente, un paziente (it. 'raggiungere'), un'espressione idiomatica (it. 'ottenere il proprio scopo', letteralmente 'raggiungere il suo'), o un complemento di direzione (it. 'giungere fino a')[64].

Un simile livello di complessità presuppone, naturalmente, l'esistenza dei diversi livelli di annotazione da cui attingere le informazioni ed è dunque riservato, per il greco e il latino classico, ad un futuro non immediato. Al contrario, un equivalente greco dell'IT-Valex basato sui testi già annotati e iclusi nell'AGDT appare del tutto

[62] Così ad esempio il complemento di direzione sarà da considerarsi oggetto di un verbo di moto. Si noti, inoltre, che il medesimo verbo può avere diverse reggenze e un diverso numero di argomenti richiesti a secondo delle sotto-classi di senso. Cf. per esempio il lat. *facere*, che può essere costruito con un oggetto all'accusativo, o con oggetto e predicativo quando significa 'eleggere', 'nominare'.

[63] Cf. McGillivray – Passarotti 2009. Il lessico è consultabile online all'indirizzo http://itreebank.marginalia.it/itvalex.

[64] L'esempio è tratto dal manuale disponibile online sul sito internet del progetto: http://ufal.mff.cuni.cz/pdt2.0/doc/pdt-guide/en/html/ch03.html#a-data-vallex; il lessico può essere consultato in rete all'indirizzo: http://ufal.mff.cuni.cz/pdt2.0/visual-data/pdt-vallex/vallex.html.

immaginabile. In virtù della facilità di uso, dell'efficacia della visualizzazione, della completezza e varietà degli esempi che esso renderebbe disponibile agli utenti, nonché alla possibilità di personalizzazione del *corpus* di riferimento di cui si è detto, esso rappresenterebbe senz'altro un'acquisizione di notevole importanza tra gli strumenti di ricerca lessicale del greco antico.

Boston Francesco Mambrini

RIFERIMENTI BIBLIOGRAFICI

Abeillé 2003
A. Abeillé, *Introduction*, *Treebanks. Building and Using Parsed Corpora*, Dordrecht-Boston 2003, xiii–xxvi.

Allen – Monro 1920
T.W. Allen – D.B. Monro, *Homeri Opera. Recognoverunt brevique adnotatione critica instuxerunt D.B. Monro et Th.W. Allen*, Oxford 1920.

Bamann – Crane 2006
D. Bamman – G. Crane, *The Design and Use of a Latin Dependency Treebank*, in *Proceedings of the Fifth International Treebanks and Linguistic Theories Conference (TLT 5)*, Prague 2006, 67-78.

Bamann – Crane 2007
D. Bamman – G. Crane, *The Latin Dependency Treebank in a Cultural Heritage Digital Library*, in *Proceedings of the Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2007)*, Prague 2007, 33-40.

Bamann – Crane 2008
D. Bamman – G. Crane, *Building a Dynamic Lexicon from a Digital Library*, in *Proceedings of the 8th ACM/IEEE-CS Joint Conference on Digital Libraries*, Pittsburg 2008, 11-20.

Bamman *et Al.* 2007
D. Bamman – M. Passarotti – G. Crane – S. Raynaud, *A Collaborative Model of Treebank Development*, in *Proceedings of the Sixth Workshop on Treebanks and Linguistic Theories (TLT 6)*, Bergen 2007, 1-6.

Bamman *et Al.* 2009
D. Bamman – F. Mambrini – G. Crane, *An Ownership Model of Annotation: The Ancient Greek Dependency Treebank*, in *Proceedings of the Eighth International Workshop on Treebanks and Linguistic Theories (TLT 8)*, Milan 2009, 5-15.

Benveniste 1950
E. Benveniste, *La phrase nominale*, Bulletin de la Société de Linguistique 45, 1950, 19-36.

Berti *et Al.* 2009
M. Berti – M. Romanello – A. Babeu – G. Crane, *Collecting Fragmentary Authors in a Digital*

Library, in *JCDL '09: Proceedings of the 2009 joint international conference on Digital libraries*, New York (NY) 2009, 259-62.

Böhmová *et Al.* 2003
A. Böhmová – J. Hajič – E. Hajičová – B. Hladká, *The Prague Dependency Treebank: Three-Level Annotation Scenario*, in A. Abeillé (ed.), *Treebanks: Building and Using Syntactically Annotated Corpora*, Dordrecht-Boston 2003, 1-26.

Boschetti 2005
F. Boschetti, *Saggio di analisi linguistiche e stilistiche condotte con l'ausilio dell'elaboratore elettronico sui Persiani di Eschilo*, Tesi di Dottorato, Università di Trento - Université Lille III 2005.

Boschetti 2009
F. Boschetti, *Digital Aeschylus. Breadth and Depth Issues in Digital Libraries*, in R. Bernardi – S. Chambers – S. Gottfried (eds.), *Proceedings of the Workshop on Advanced Technologies for Digital Libraries 2009 (AT4DL 2009)*, Bozen 2009, 5-9.

Busa 1974-80
R. Busa, *Index Thomisticus. Sancti Thomae Aquinatis operum omnium indices et concordantiae*, Stuttgart 1974-1980.

Chomsky 1957
N. Chomsky, *Syntactic Structures*, The Hague 1957.

Chomsky 1965
N. Chomsky, *Aspects of the Theory of Syntax*, Cambridge 1965.

de Saussure 1916
F. de Saussure, *Cours de linguistique générale, publié par Charles Bailly et Albert Sechehaye*, Lausanne 1916.

Dickinson – Meurers 2003
M. Dickinson – W.D. Meurers, *Detecting Inconsistencies in Treebanks*, in *Proceedings of the Second International Workshop on Treebanks and Linguistic Theories (TLT 2)*, Växjö 2003, 45-56.

Dik 1995
H. Dik, *Word Order in Ancient Greek: A Pragmatic Account of Word Order Variation in Herodotus*, Amsterdam 1995.

Dik 2007
H. Dik, *Word Order in Greek Tragic Dialogue*, Oxford 2007.

Edmunds 2008
L. Edmunds, *Deixis in Ancient Greek and Latin Literature: Historical Introduction and State of the Question*, Philologia Antiqua 1, 2008, 67-99.

Evelyn-White 1914
H.G. Evelyn-White, *Hesiod. The Homeric Hymns and Homerica with an English Translation*, London 1914.

Francis – Kucera 1964
W.N. Francis – H. Kucera, *Manual of Information to Accompany a Standard Corpus of Present-day Edited American English, for Use with Digital Computers*, Providence 1964.

Francis – Kucera 1967
W.N. Francis – H. Kucera, *Computational Analysis of Present Day American English*, Providence 1967.

Gilquin 2010
G. Gilquin, *Corpus, Cognition and Causative Constructions (Studies in Corpus Linguistics)*, Amsterdam-Philadelphia 2010.

Gries 2003
S.T. Gries, *Multifactorial Analysis in Corpus Linguistics. A Study of Particle Placement*, London-New York 2003.

Guiraud 1962
C. Guiraud, *La phrase nominale en grec d'Homère à Euripide*, Paris 1962.

Hajičová – Sgall 2006
E. Hajičová – P. Sgall, *Corpus Annotation as a Test of a Linguistic Theory*, in *Proceedings of the Fifth International Language Resources and Evaluation (LREC'06)*, Genoa 2006, 879-84.

Jebb 1896
R.C. Jebb, *Sophocles. The Plays and Fragments. Part VII. The Ajax*, Cambridge 1896.

Lallot 1997
J. Lallot, *Apollonius Dyscole. De la Construction*, Paris 1997.

Leech 2004
G. Leech, *Adding Linguistic Annotation*, in M. Wyne (ed.), *Developing Linguistic Corpora: A Guide to Good Practice*, Oxford 2004.

McEnery – Wilson 2001
T. McEnery – A. Wilson, *Corpus Linguistics*, Edinburgh 2001[2] .

McGillivray – Passarotti 2009
B. McGillivray – M. Passarotti, *The Development of the Index Thomisticus Treebank Valency Lexicon*, in *Proceedings of the EACL 2009 Workshop on Language Technology and Resources for Cultural Heritage, Social Sciences, Humanities, and Education (LaTeCH-SHELT&R 2009)*, Athens 2009, 43-50.

Meillet 1906-08
A. Meillet, *La phrase nominale en indo-européen*, Mémoires de la Société de Linguistique de Paris 14, 1906-08, 1-26.

Moorhouse 1982
A.C. Moorhouse, *The Syntax of Sophocles*, Leiden 1982.

Palmer *et Al.* 2005
M. Palmer – D. Gildea – P. Kingsbury, *The Proposition Bank: An Annotated Corpus of Semantic Roles*, Computational Linguistics 31.1, 2005, 71-106.

Pasquali 1952
G. Pasquali, *Storia della tradizione e critica del testo*, Firenze 1952.

Passarotti 2009
M. Passarotti, *Theory and Practice of Corpus Annotation in the "Index Thomisticus Treebank"*, Lexis 27, 2009, 5-24.

Passarotti c.s.
M. Passarotti, *When Praguian Functionalism Meets Latin. From Analytical to Tectogrammatical Annotation of Latin Syntax*. in *Proccedings of the International Conference on Linguistics and Classical Languages (LCL)*, Rome c.s.

Prasad *et Al.* 2008
R. Prasad – N. Dinesh – A. Lee – E. Miltsakaki – L. Robaldo – A. Joshi – B. Webber, *The Penn Discourse Treebank 2.0*, in *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*, Marrakech 2008, 2961-8.

Romanello *et Al.* 2009
M. Romanello – M. Berti – F. Boschetti – A. Babeu – G. Crane, *Rethinking Critical Editions of Fragmentary Texts By Ontologies*, in S. Mornati – T. Hedlund (eds.), *Proceedings of 13th International Conference on Electronic Publishing: Rethinking Electronic Publishing: Innovation in Communication Paradigms and Technologies*, Milan 2009, 155-74.

Sgall *et Al.* 1986
P. Sgall – E. Hajičová – J. Panevová, *The Meaning of the Sentence and Its Semantic and Pragmatic Aspects*, Prague-Dordrecht 1986.

Sinclair 1987
J.M. Sinclair, *Looking up: An Account of the COBUILD Project in Lexical Computing and the Development of the Collins COBUILD English Language Dictionary*, London 1987.

Sinclair 2004
J.M. Sinclair, *Corpus and Text – Basic Principles*, in M. Wyne (ed.) *Developing Linguistic Corpora: A Guide to Good Practice*, Oxford 2004.

Smith 2010
N. Smith, *Digital Infrastructures and the Homer Multitext Project*, in G. Bodard – S. Mahony (eds.), *Digital Research in the Study of Classical Antiquity*, Farnham 2010, 121-38.

Smyth 1922
H.W. Smyth, *Aeschylus. With an English Tranlsation*, Cambridge (MA) 1922.

Sommerstein 1989
A.H. Sommerstein, *Aeschylus. Eumenides*, Cambridge 1989.

Taylor *et Al.* 2003
A. Taylor – M. Marcus – B. Santorini, *The Penn Treebank: an Overview*, in A. Abeillé (ed.), *Treebanks: Building and Using Syntactically Annotated Corpora*, Dordrecht-Boston 2003, 5-22.

Tesnière 1959
L. Tesnière, *Éléments de syntaxe structurale*, Paris 1959.

Tognini-Bonelli 2001
E. Tognini-Bonelli, *Corpus Linguistics at Work*, Amsterdam-Philadelphia 2001.

Zeldes – Lüdeling 2007
A. Zeldes – A. Lüdeling, *Three Views on Corpora: Corpus Linguistics, Literary Computing, and Computational Linguistics*, Jahrbuch für Computerphilologie 9, 2007, 151-80.

***Abstract*:** This article aims to provide an overview of the Ancient Greek Dependency Treebank (AGDT) that has been developed by the Perseus Project (Tufts University). The AGDT is the first corpus that includes complete morphological and syntactical annotation for the ancient Greek language. Currently in its first release, it includes more than 300,000 words and this corpus could potentially serve as a major asset for linguistic and philological research on the ancient Greek language. This article will also introduce some of the methodological principles that guided the treebank's construction, as well as outline a number of the main directions in which the tool can be expanded. Some practical applications of the treebank will also be presented, including: querying of the treebank for linguistic research, expansion of the annotation to other levels of linguistic analysis (semantics, pragmatics), the potential for lexicography and the creation of valency lexicons. The goal of this discussion is to show how fruitful the interaction between computational linguistics and Classical philology can be, one from which both disciplines can greatly benefit.

***Keywords*:** linguistics, treebank, Greek syntax, annotated *corpora*, lexicography.