Misurare Memorata Poetis: prime statistiche

Silvia Arrigoni, Università "Ca' Foscari" di Venezia, silvia.arrigoni@unive.it
Fahad Khan, Università "Ca' Foscari" di Venezia - CNR-ILC, fahad.khan@ilc.cnr.it
Monica Monachini, CNR-ILC, monica.monachini@ilc.cnr.it
Federico Boschetti, CNR-ILC, federico.boschetti@ilc.cnr.it

1 Memorata Poetis e l'annotazione tematica

Il progetto di ricerca *Memorata Poetis 'memoria poetica e poesia della memoria'*[1], ha avuto come scopo precipuo la creazione di un ampio *corpus* di testi, in prevalenza di natura epigrammatica ed epigrafica, per favorire lo studio dell'intertestualità attraverso le differenti lingue e tradizioni poetiche, oltre che nei diversi generi letterari di riferimento. L'interesse per questi testi si è rivolto all'indagine automatica della presenza di medesimi temi e motivi all'interno di tradizioni linguistiche e culturali che possano essersi influenzate a vicenda, lungo un arco cronologico piuttosto ampio; per questo motivo, il database di testi raccolti nel sito nel progetto è stato dotato di un motore di ricerca che, avvalendosi della marcatura tematica manuale effettuata sugli stessi, permetta di compiere analisi di tipo tematico e, quindi, semantico. Il bacino linguistico raggruppa testi nelle due lingue classiche per eccellenza, greco e latino, quest'ultima con un'estensione cronologica individuabile dalle origini alla produzione poetica in lingua latina di età umanistica e rinascimentale, senza trascurare la produzione epigrafica medievale, ma anche testi in lingua araba, italiana delle origini, e inglese. Proprio per la molteplicità di lingue interessate, si è reso necessario concentrarsi su una tipologia testuale ben definita e con caratteri di ricorsività tematica, oltre che formulare, tali da individuare chiaramente temi e motivi propri di ciascun testo

Il progetto è stato condotto nell'ambito di un PRIN (2010-11) in grado di coinvolgere differenti istituzioni universitarie in tutta Italia, in modo tale che ciascuna delle unità di ricerca interessate fosse specializzata in una delle differenti lingue dei testi inseriti nel *corpus*. Al progetto hanno partecipato circa 45 esperti, che hanno curato a vario titolo l'operazione di marcatura tematica dei testi.

L'eterogeneità fra testi prodotti in lingue e culture diverse, unita alla natura manuale dell'operazione di taggatura tematica dei testi comporta, tuttavia, un certo inevitabile grado di arbitrarietà, dovuto alle differenti interpretazioni che ciascun testo offre quotidianamente agli studiosi e alle singole letture che di esso avvengono.

Nell'ambito delle Digital Humanities e della Linguistica Computazionale, i task di annotazione prevedono generalmente che i medesimi documenti siano annotati da più sog-

[1] http://www.memoratapoetis.it

getti: i modelli più frequenti richiedono di norma tre annotatori oppure due annotatori e un supervisore che armonizzi i casi discordanti.

Rispetto a questo protocollo, per l'annotazione dei testi di *Memorata Poetis* si è deciso di operare in modalità differenti:

si è privilegiata, in questa prima fase, la quantità di testi annotati rispetto alla qualità dell'operazione stessa, per ottenere la sufficiente massa critica su cui poter operare in una seconda fase del progetto; data la natura fortemente innovativa del progetto, si è rinunciato a fornire linee guida stringenti, limitando invece l'adozione di una tassonomia di temi e motivi stabilita a priori per l'annotazione, basata sugli *Indices Rerum Notabilium* presenti nelle antologie di poesia classica e ulteriormente raffinata da esperti filologi. Il tagset o 'Indice dei temi e motivi' propone circa 1250 voci, suddivise in sei raggruppamenti principali (*Animalia, Arbores et virentia, Dei et heroes, Homines, Loca, Res*) e organizzate su tre livelli gerarchici, da concetti più generali e in grado di racchiuderne altri al proprio interno (es. *Animalia* per 'Animali'), a un secondo livello più specifico (es. *Genera animalium*, vale a dire 'Specie animali' incluso in *Animalia* e a sua volta produttivo quanto a temi), al terzo e più puntuale, concernente temi altamente specialistici (es. *Amphibia*, gli 'Anfibi', voce dipendente da *Genera animalium*). Altri esempi di tag possono essere *Vsus animalium, Vsus in medicina, Aetates animalium, Flores, Metamorphosis in arbores, Evocationes, Dei artis medicae, Simulationes et dissimulationes*[2].

L'intervento che proponiamo fornirà statistiche relative al *modus operandi* degli annotatori e alla distribuzione dei tags. Lo scopo è quello di studiare l'omogeneità dell'annotazione attraverso differenti tipologie testuali, in particolar modo propri di generi diversi, come ad esempio gli epigrammi funebri, etc. Per l'analisi statistica vorremmo considerare la densità dei tag per ciascun testo nei tre *corpora* in lingua latina; il numero di tag tematici per verso, la frequenza di annotazioni per ciascun testo; la granularità dei tag (livello impiegato fra i tre attualmente presenti nella tassonomia dell'*Index* di temi); l'informatività dell'annotazione.

Le statistiche di *Memorata Poetis*: primi risultati.

Il *corpus* complessivo del progetto contiene circa 12.500 testi marcati. La loro generale distribuzione nelle differenti lingue è rappresentata nella Fig. 1; nella Fig. 2 è invece presentata la ripartizione dei testi in relazione al parametro della lunghezza (numero di versi): si tratta in prevalenza di epigrammi brevi. Quanto alle statistiche più generali, si è riscontrata la presenza media di solo 1.4 temi (tag) per verso e, nell'intero *corpus*, circa il 33% dei versi di ciascuna composizione è stata marcata. Gli esperti che si sono occupati della tematizzazione potevano assegnare un tema al testo intero in aggiunta o in alternativa a quelli per i singoli versi; l'annotazione dei testi interi è presente nell'85% dei casi. Nel computo figuravano più di 85.000 annotazioni tematiche, corrispondenti a più di 64.000 tipologie di annotazione. Abbiamo deciso di classificare i tag in 5 differenti categorie, vale a dire sostantivi astratti, concreti, nomi propri, espressioni formulari, e abbiamo creato anche una generica categoria 'altro' per quei temi che non siamo riusciti a ricondurre a nessuna delle precedenti. La distribuzione di queste categorie nei testi di *Memorata Poetis* è alla Fig. 3.

Nel corso dello studio dei dati ci siamo resi conto del fatto che vi era una considerevole variazione concernente le modalità di annotazione; ciò è in parte dovuto all'istituzione

[2]L'elenco completo dei temi è disponibile alla pagina l'elenco completo dei temi è disponibile alla pagina http://www.memoratapoetis.it/public/memorata/ricerca/index.

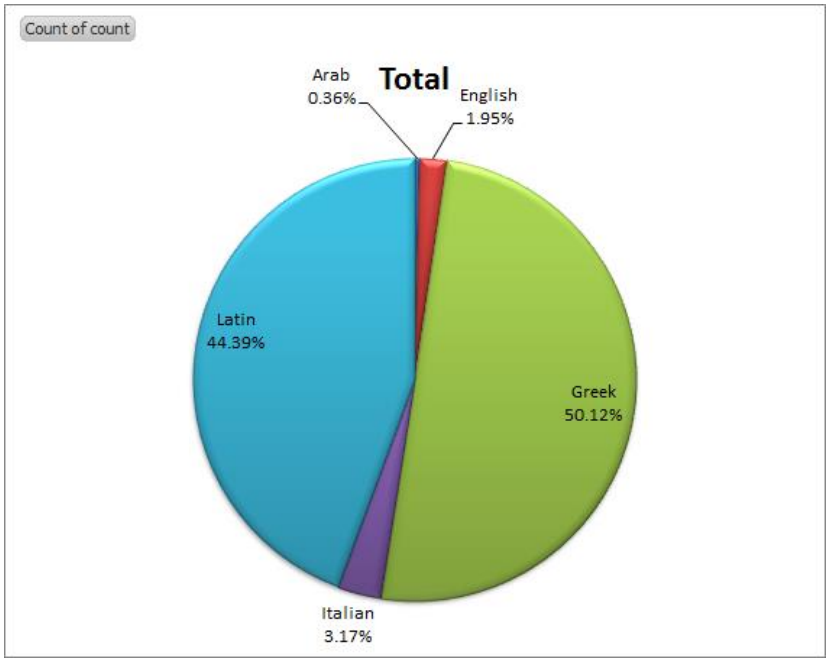

Figura 1

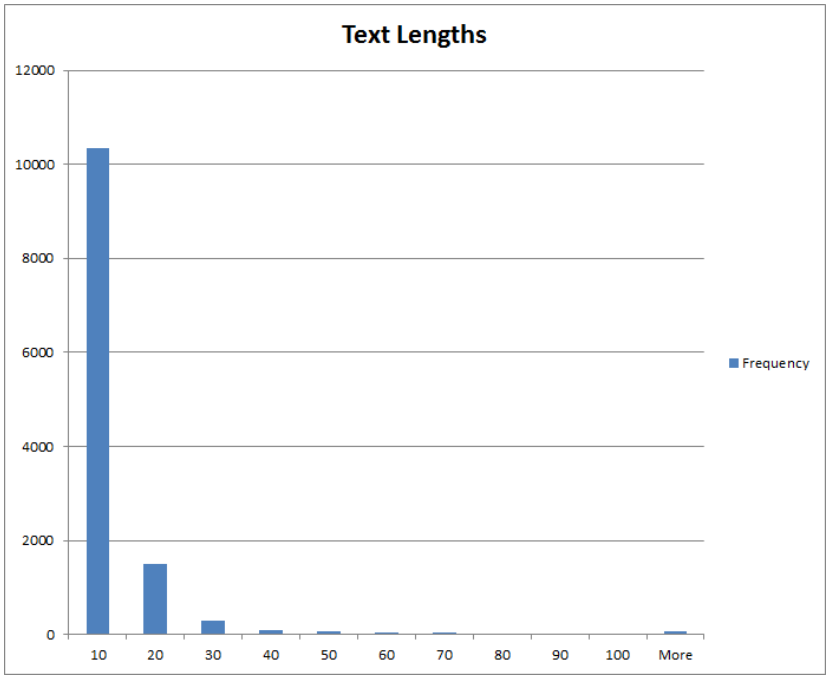

Figura 2

universitaria di riferimento e, conseguentemente, alla differente 'scuola di pensiero filologico'. Per fare alcuni esempi, una medesima formula ricorrente con frequenza in alcuni testi di ambito funerario latino, quale è *hic situs est* ('qui è sepolto' o 'qui giace'), è talvolta resa tramite l'utilizzo del tema *Tumulus (monumentum non profanandum)* contenuto nel ma-

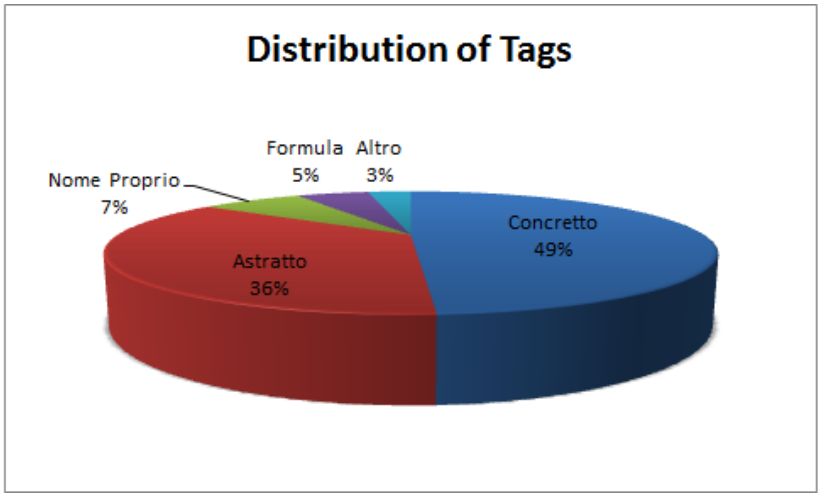

Figura 3

crogruppo *Homines* e nella sezione *Mors* (CLE 317, 3), talvolta con *Tumulus (mortuus vel sepulcrum adloquuntur)*[3] come nel caso di CLE 1398, altre ancora si privilegia il contesto complessivo del testo a discapito dell'annotazione puntuale della formula, come si evince da CLE 472, 1[4].

Per poter studiare questa divergenza, non disponendo di medesimi testi annotati da più di un esperto (cf. *supra*), abbiamo deciso di concentrare la nostra attenzione su una sezione omogenea di circa 1380 testi, i *Carmina Epigraphica*, redatti in una sola lingua, il latino, e tematizzati da due differenti unità di ricerca, che indicheremo rispettivamente con T1 e T2. In maniera abbastanza sorprendente, nonostante T1 abbia marcato il 72,5% dei testi, e T2 solo il 27, 5%, T2 ha utilizzato un numero complessivo di tag ben superiore rispetto a T1. Ciò non è tuttavia dovuto alla maggiore lunghezza dei testi di T2, come si può vedere nelle Fig. 4, che è un istogramma relativo alla lunghezza dei testi rispettivamente per T1 e T2, normalizzato per la lunghezza dei *corpora* di T1 e T2. Vi è inoltre una netta divergenza fra T1 e T2 in termini di quantità percentuale dei versi annotati in ogni testo, come mostra la Fig. 5.

2 Conclusioni

Il nostro scopo per il presente contributo è quello di analizzare nel dettaglio le statistiche prodotte per *Memorata Poetis* e di studiarne il significato per poter valutare la funzionalità di un corpus annotato di testi come strumento per l'analisi semantica applicata alla poesia.

[3] Collocato sempre in *Homines* > *Mors*.

[4] I testi sono visibili alla pagina http://www.memoratapoetis.it/public/memorata/pagine/testi, selezionando dal menu 'Poesia latina (origini – VII sec.)' e quindi *carmina epigraphica*.

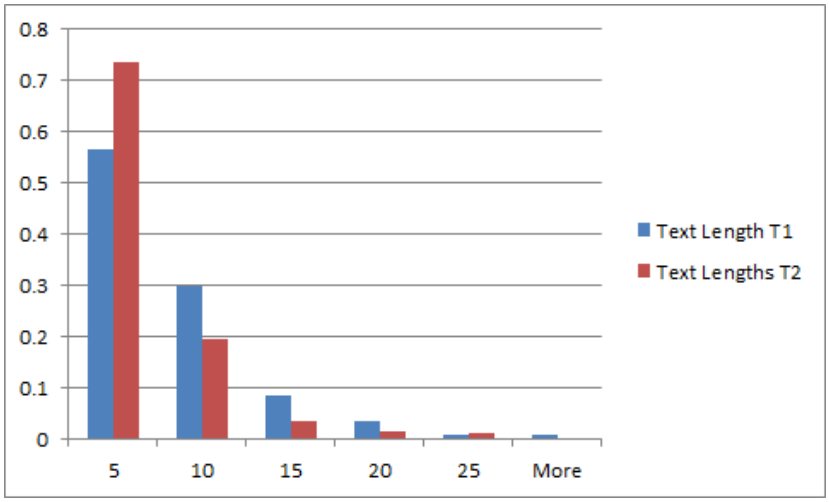

Figura 4: La lunghezza dei testi annotati da T1 e T2 (normalizzato per il numero totale dei testi taggati da T1 e T2).

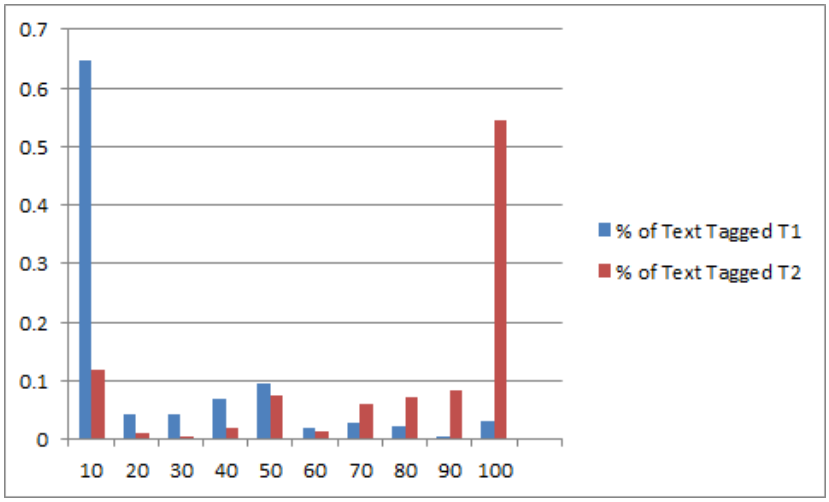

Figura 5: La frequenza della percentuale di ogni testo taggato (normalizzato).

Bibliografia

Boschetti, F., R. Del Gratta e M. Lamé. 2014a. «Computer Assisted Annotation of Themes and Motifs in Ancient Greek Epigrams: First Steps». In *Proceedings of the First Italian Computational Linguistics Conference (CLIC-it)*, 83–86. Pisa. http://www.fileli.unipi.it/projects/clic/proceedings/Proceedings-CLICit-2014.pdf.

Boschetti, F., et al. 2016. «Strumenti, Risorse e Linguistic Linked Open Data per le Lingue Antiche». In *Proceedings of the 4th Conference of the Associazione per l'informatica Umanistica e la Cultura Digitale (AIUCD)*, forthcoming. Torino.

Jiang, J. J., e D. W. Conrath. 1997. «Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy». In *Proceedings of International Conference Research on Computational Linguistics (ROCLING X)*, 1–15. Taiwan. http://arxiv.org/pdf/cmp-lg/9709008.pdf.

Khan, A. F., et al. 2016. «Restructuring a Taxonomy of Literary Themes and Motifs for More Efficient Querying». *MATLIT* 4 (2): 11–27. doi:10.14195/2182-8830. http://iduc.uc.pt/index.php/matlit/article/view/2354/2252.