AIUCD 2016

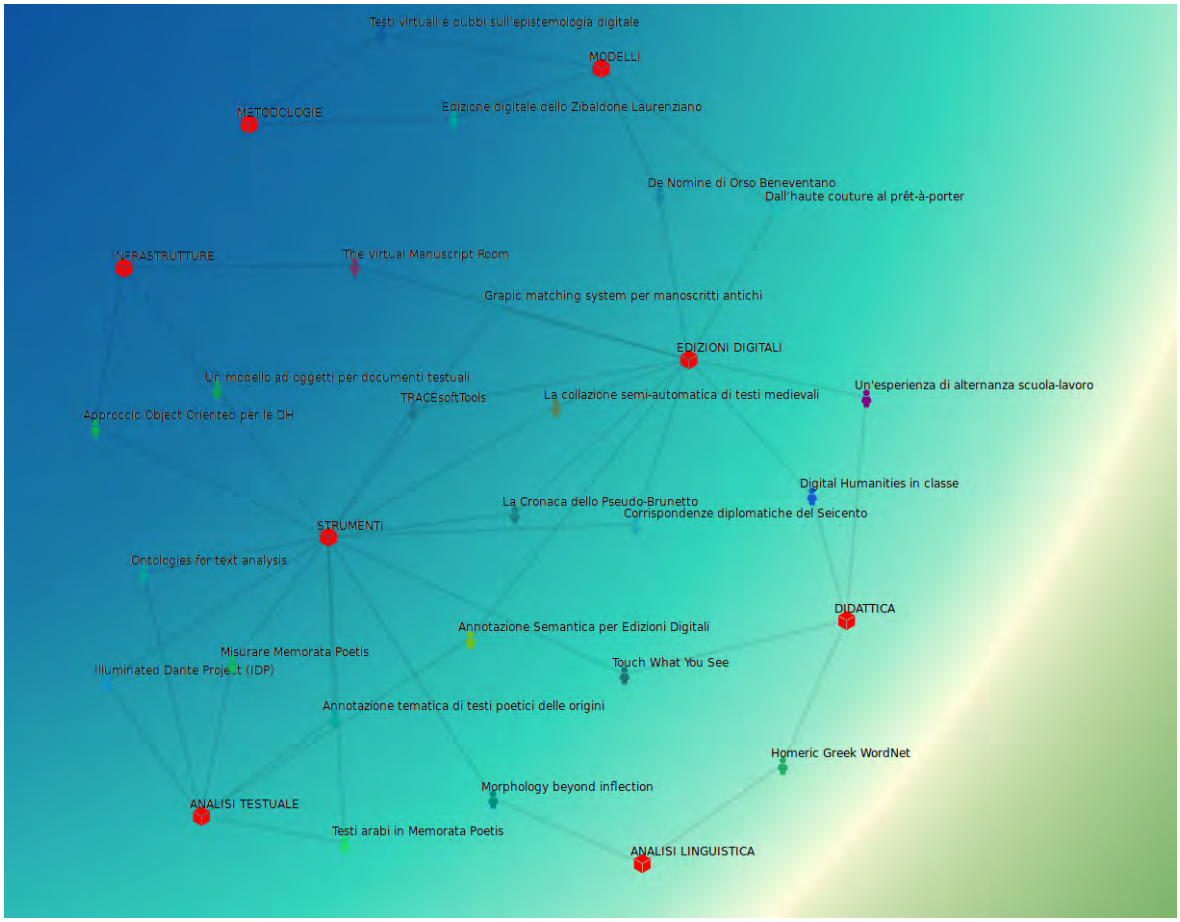
Venezia, 7–9 Settembre 2016 / Venice, 7th–9th September 2016

Book of Abstracts

Edizioni digitali: rappresentazione, interoperabilità, analisi del testo e infrastrutture

Digital editions: representation, interoperability, text analysis and infrastructures

edited by Federico Boschetti



ASSOCIAZIONE PER L'INFORMATICA UMANISTICA E LA CULTURA DIGITALE

Firenze 2017

AIUCD 2016

Book of Abstracts

Firenze 2017

ISBN: 978-88-942535-0-4

AIUCD 2016

Venezia, 7–9 Settembre 2016 / Venice, 7th–9th September 2016

Book of Abstracts

Edizioni digitali: rappresentazione, interoperabilità, analisi del testo e infrastrutture
Digital editions: representation, interoperability, text analysis and infrastructures

edited by Federico Boschetti

ASSOCIAZIONE PER L'INFORMATICA UMANISTICA E LA CULTURA DIGITALE

Firenze 2017

Gli abstract pubblicati in questo volume hanno ottenuto il parere favorevole da parte di valutatori esperti della materia, attraverso un processo di revisione anonima sotto la responsabilità del Comitato Scientifico di AIUCD 2016.

Questo volume non comprende gli interventi degli *invited speakers*, ma le presentazioni e i video autorizzati sono disponibili online: `http://aiucd2016.unive.it`

All abstracts published in this volume have received favourable reviews by experts in the field of DH, through an anonymous peer review process under the responsibility of the AIUCD 2016 Scientific Committee.

This book does not contain the key-note lectures, but authorized videos and slides are available online: `http://aiucd2016.unive.it`

Overview

Il convegno AIUCD 2016 è dedicato alla rappresentazione e allo studio del testo sotto vari punti di vista (risorse, analisi, infrastrutture di pubblicazione), con lo scopo di far dialogare intorno al testo filologi, storici, umanisti digitali, linguisti computazionali, logici, informatici e ingegneri informatici.
Sono stati presentati al convegno 45 contributi sottoposti a peer review. 23 contributi orali su 35 (66%) sono stati accettati nella forma originaria, mentre 12 su 35 (34%) sono stati accettati dopo la conversione in poster. I 10 interventi presentati dagli autori stessi come poster sono stati tutti ammessi nella forma proposta.

Ad AIUCD 2016 hanno partecipato non solo autori italiani (70%), ma anche di altri paesi europei (Svizzera: 10%; Germania: 7%; Francia: 5%; Spagna: 2%; Olanda: 1%) e nordamericani (Canada: 3%; Stati Uniti: 1%). Questo del resto è in linea con la composizione del comitato scientifico, formato da studiosi che operano in Italia (86%), Germania (7%) e Stati Uniti (7%).

Da un lato le Digital Humanities, oltre alla creazione e al mantenimento di risorse (digitalizzazione, annotazione, etc.), devono prendere in considerazione il modo in cui tali risorse saranno usate.
Dall'altro lato la linguistica computazionale, oltre allo sviluppo di strumenti di analisi (parser, riconoscitori di entità nominate, etc.), deve tener conto della qualità delle risorse su cui far operare tali strumenti.

The AIUCD 2016 conference is devoted to the representation and study of the text under different points of view (resources, analysis, infrastructures), in order to bring together philologists, historians, digital humanists, computational linguists, logicians, computer scientists and software engineers and discuss about the text.
At the conference 45 contribution have been submitted and have undergone peer review. 23 oral contributions out of 35 (66%) have been accepted in the original form, whereas 12 out of 35 (34%) have been accepted after the conversion in poster. The 10 interventions presented as posters by the authors have all been accepted in the original form.
At AIUCD 2016 paticipated not only authors coming from Italy (70%), but also from other European countries (Switzerland: 10%; Germany: 7%; France: 5%; Spain: 2%; Holland: 1%) and from North America (Canada: 3%; United States: 1%). This is in line with the composition of the scientific committee, formed of scholars operating in Italy (86%), Germany (7%) and United States (7%).
On the one hand, Digital Humanities, in addition to the creation and maintenance of resources (digitization, annotation, etc.), must take into account how they will be used.
On the other hand, Computational Linguistics, in addition to the development of computational tools (parsers, named entity recognizers, etc.), must take into account the quality of the resources on which the same tools are applied.

Questi aspetti: formale (modelli), digitale (risorse), computazionale (strumenti), infrastrutturale (piattaforme) e sociale (comunità) coinvolgono professionalità diverse che il convegno si pone l'obiettivo di far interagire.
La creazione di risorse e lo sviluppo di strumenti dovrebbero avanzare di pari passo ed essere fondati su modelli solidi che soddisfino i requisiti stabiliti dalla comunità degli esperti di dominio. Soltanto sapendo come usare il testo, cosa ricavare dal testo e come ricavarlo, è possibile rappresentarlo in modo adeguato.

Ora che le grandi iniziative di digitalizzazione mettono a disposizione molteplici edizioni delle medesime opere, abbondante letteratura secondaria e numerose opere di consultazione (dizionari, enciclopedie, etc.), il filologo che opera nell'era digitale dovrebbe essere in grado di passare senza soluzione di continuità dalla gestione di fenomeni prettamente ecdotici (gestione delle varianti) all'analisi del testo secondo altri principi (linguistica computazionale).
Gli strumenti di analisi e le metodologie statistiche sviluppati per essere usati su un intero corpus di testi letterari o su ampie collezioni di letteratura secondaria devono integrarsi con gli strumenti per comparare varianti testuali e valutare varianti interpretative.
È tempo che le infrastrutture di ricerca diventino garanti di interoperabilità e integrazione degli strumenti per lo studio filologico con gli strumenti per l'analisi di ampi corpora testuali, abbattendo le barriere rigide fra filologia digitale e computazionale da un lato e linguistica dei corpora dall'altro.

These aspects: formal (models), digital (resources), computational (tools), infrastructural (platforms) and social (communities) involve different skills that the conference aims to make interact with each other.

The creation of resources and the development of tools should advance hand in hand, and should be based on solid models that meet the requirements established by the experts of the field. It is necessary that resources and tools be developed in parallel: only if you know how to use the text, what can be extracted from it and how to do it, can you adequately represent it.
Now that the major digitization initiatives provide multiple editions of the same works, abundant secondary literature, as well as numerous reference books (dictionaries, encyclopedias, etc.), the philologist who works in the digital age should be able to seamlessly switch from handling purely philological phenomena (variant studies) to text analysis performed according to different methods (computational linguistics).
The analysis tools and statistical methods developed to be used on an entire corpus of literary texts or extensive secondary literature collections must be integrated with the tools for comparing textual variants and evaluating possible interpretations.

It is time for research infrastructures to be able to guarantee interoperability and integration between the instruments for philological studies and the instruments for the analysis of large textual corpora, breaking down the rigid barriers between digital and computational philology on the one hand, and corpus linguistics on the other hand.

Sommario

Modelli e Metodologie

Models and Methods

Posters

The virtual representation of the World City Project: from text to 3D model

Valentina Bova, Università della Calabria, valentina.bova@dimes.unical.it
Francesca Parisi, Università della Calabria, francesca.parisi@dimes.unical.it
(The authors are listed in alphabetical order)

1 Introduction

People have the natural ability to depict scenes thanks to language descriptions. Sometimes the setting really exists but, other times, they are born from the imagination of the writer. In the first case the association between textual description and three-dimensional image is simple, while, when the scene is the result of the writer's fantasy, the reader can comprehend what the author writes only using his imagination. Furthermore, the comprehension is even more difficult if the text refers to a specific domain. Let us consider, for example, an architectural scene described by textual descriptions and accompanied by 2D architectural drawings: one can imagine that the three-dimensional visualization could be very difficult for non-expert readers.

On this basis, this work illustrates the possibility to assign a shape to textual descriptions in order to materialize them in a virtual space. We have focalised our work on many documents related to the "World City" project (Otlet 1929), a utopian idea dated around 1927 and conceived by the biographer Paul Otlet and the architect Le Corbusier. Our aim, therefore, is to allow the users to visualize an automatically generated 3D representation of a text, even if they are not specialists of the architectural domain.

2 State of Art

Prior works that use the *3D scene generation from text* methodology are present in the literature. For instance, **WordsEye** (Coyne and Sproat 2001) is a tool that allows to create three-dimensional scenes starting from a simple textual description. In particular, users describe an object, an environment or an action through a short text and the program, firstly, carries out a syntactic and semantic analysis, and subsequently assigns a 3D model to each semantic element. In the end it assembles all the models to create a three-dimensional scene that approximates the textual description.

An interesting aspect of text-to-scene conversion that regards the knowledge of how human actions, described in a text, are effectively realized in the reality has also been dealt in the paper (Coyne et al. 2012). A computer system, in fact, cannot know that a person assumes a different position when he sleeps (-he is lying-) or when he walks (-he is standing-), so, the authors of the paper have created a set of vignettes that acts as a link between

function and form. In other words, they find a connection between the meaning of the scene and its representation.

An extension of the pioneering WordsEye system was presented by (Chang, Savva, and Manning 2014). The authors of the paper described a method to represent unstated facts: they showed how their system can infer the presence of missing objects in the text description in order to insert them in the 3D scene.

A further study carried out by Cropper (2014), concerns some textual descriptions extracted from books. In particular, the paper is focused on a text-to-scene conversion system (TTSCS) that is able to generate 2D or 3D scenes from a short text of fiction books. The problem connected to book's descriptions is that authors generally do not provide detailed descriptions of the scenes, but they prefer to let the readers use their imagination. On this basis, to generate a fairly accurate 3D model it is necessary to identify not only the explicit objects but also the implicit ones. Therefore, the paper describes an approach that permits to infer implicit objects through the identification of explicit elements in the text.

Another method presented in Chang et al. (2015) is able to ground many lexical terms to concrete referents and to improve 3D scene generation. The authors created a dataset of 3D scenes annotated with natural language descriptions and, starting from this data, the system can learn how to ground textual descriptions to physical objects. Furthermore, Dessai and Dhanaraj (2016) proposed a 3D scene generation system supported by user interaction. The user, in fact, takes part in the process in two different steps: at the beginning and in the end of it. Specifically, he provides a short natural language text as input that allows the system to find explicit constraints on the objects of the virtual scene. Afterwards the user can interact again with the system because he can modify the candidate scene through textual commands or by direct control. In that way the system can acquire more spatial knowledge and add new learned relationships or implicit constrains to the existing ones. In the end, the final output is a 3D virtual scene that the user can observe from all points of view and can also render to create 2D images.

3 Methodological proposal

The purpose of our methodology is to find a set of rules that, starting from unstructured textual information and (previously modelled) three-dimensional elements, will generate a 3D scene. In particular, we aim to find a set of lexical expressions (e.g.: combination of names, verbs and adverbs) through semantic annotation techniques (Kiryakov et al. 2003) that will suggest the shapes and the locations of the 3D elements in the virtual space. In other words, we purpose to formalize some associations between words and 3D models that will represent an important starting point for the construction of semi-automatic algorithms. These latter will allow the readers to automatically visualize the three-dimensional scene associated to the textual documents.

To sum up, below we illustrate a set of steps that explains our methodological approach:

1. Collection of all the relevant sources concerning the "World City" project (e.g.: publications, 2D architectural drawings, letters, notes, etc.).

2. Creation of a "3D library" containing three-dimensional objects referring to the project;

3. Identification of existing lexical resources for the architectural domain (e.g.: ontologies, thesaurus, etc.);
4. Creation of semantic annotation rules to assign a specific meaning to words or expressions in the texts;
5. Formalization of the link between annotated elements, 3D models and ontology;
6. Evaluation of the 3D models obtained.

To date, the approach tested on the case study is at its starting point. We are working on the collection of textual resources and 2D plans concerning the "World City" project. The texts considered are: the original publications of Paul Otlet (primary sources, written in French), texts or books written by authors who studied the project and published on it (secondary sources, written in different languages), architectural drawings or drafts of Le Corbusier, notes and letters of correspondence between Otlet and the architect, and so forth. The documents are principally kept in two institutes that are the archive centre of "Mundaneum", in Mons (Belgium), and the "Foundation Le Corbusier", in Paris (France).

After having collected the documentation we have modelled some 3D objects that will be archived in a "3D library" and, when necessary, they will be taken by the system and used for the construction of the scene. Considering that not all the buildings of the "World City" have architectural drawings in support of the project, but only textual descriptions, the 3D library will also contain generic three-dimensional elements related to the architectural domain in general. Logically, the resulting 3D scene will be very different because, in presence of technical drawings the system will generate detailed models, similar to the idea of the authors, conversely, without drawings, the resulting 3D models will be less detailed and based on suppositions (one of the possible models).

The subsequent phases will cover textual semantic annotation and scene template parsing; they will consist of the following steps:

1. Automatic corpus cleaning to remove useless parts of text (e.g.: punctuation, extra spaces, etc.);
2. Application of Part-of-speech (POS) tagging techniques (Van Halteren 1999) to identify the grammatical category of each term of the text;
3. Identification of verbs to indicate a specific position (e.g.: "emprunter", "engager");
4. Classification of the words in the text according to semantic-conceptual categories such as : people, position, place, event, physical object, etc.;
5. Identification of the objects through the detection of the nouns present in the phrase and their semantic-conceptual categories;
6. Identification of objects properties thanks to the extraction of adjectives and nouns;
7. Creation of dependency patterns between words such as "X is composed by Y";
8. Construction of spatial relations between objects like: inside(A,B), outside(A,B), near(A,B), left_of(A,B), above(A,B) etc.;

Once having annotated the documents, all the resources (e.g.: single terms, expressions, position indicators[1], etc.) identified in the texts and associated to the elements of the "3D library", will be described by means of an existing ontology. In particular, we propose to collect specific domain entities (e.g. : column) in generic ontological classes (e.g.: architectural elements) and to identify proprieties and relations between them (figure 1).

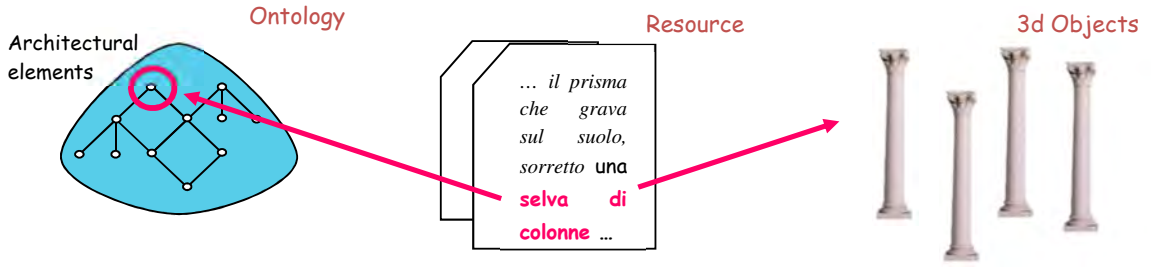


Figure **1** – Association between a resource, an element of the 3D library and the ontology.

Figure 1: Association between a resource, an element of the 3D library and the ontology

9. Identification of keywords extracted from the text (nouns and adjectives referring to the objects) and their application to all the three-dimensional objects contained in the database. These keywords will be useful for querying the 3D elements from the database.

Expected results

The methodology described explains how the system will represent a three-dimensional scene through the interpretation of textual descriptions. The principal difficulty of this process is that the machines do not have the capabilities of analysis like humans do, but they can only analyze the code.

Therefore, it is necessary to transfer adequate knowledge to the system, so that it will be able to interpret human language in a specific domain.

The process of semantic annotation that we will use is one of the ways so make an unstructured textual description explicit. Thanks to this procedure we will assign a specific meaning to each word in the phrases and we will guide the system to a correct interpretation of the text.

In particular, once having collected all the useful documents related to the "World City", we will apply the abovementioned methodology to generate a 3D virtual scene that will describe the project from an architectural point of view. All the single three-dimensional elements will be extracted from the 3D database and assembled by the system through specific rules to construct a more complex scene. The examples below have been obtained through human modelling and show the expected result using the methodology proposed in this paper.

Once having annotated the documents, all the resources (e.g. : single terms, expressions, position indicators1, etc.) identified in the texts and associated to the elements of the "3D library", will be described by means of an existing ontology. In particular, we propose to collect specific domain entities (e.g. : column) in generic ontological classes (e.g.: architectural elements) and to identify proprieties and relations between them (Fig. 1).

4 Expected results

The methodology described explains how the system will represent a three-dimensional scene through the interpretation of textual descriptions. The principal difficulty of this process is that the machines do not have the capabilities of analysis like humans do, but they can only analyze the code.

Therefore, it is necessary to transfer adequate knowledge to the system, so that it will be able to interpret human language in a specific domain.

The process of semantic annotation that we will use is one of the ways so make an unstructured textual description explicit. Thanks to this procedure we will assign a specific meaning to each word in the phrases and we will guide the system to a correct interpretation of the text.

In particular, once having collected all the useful documents related to the "World City", we will apply the abovementioned methodology to generate a 3D virtual scene that will describe the project from an architectural point of view. All the single three-dimensional elements will be extracted from the 3D database and assemblehuman modelling and show the expected result using the methodology proposed in this paper.

[1] For "position indicators" we intend all expressions that give information about the specific position of the object ("au pied de", "à pic sur" etc.).

"[...] il prisma che grava sul suolo, sorretto da una selva di colonne [...]. Di qui i sette livelli della sezione sono raggiungibili alla quota livello primo mediante un tunnel orizzontale proveniente dal corpo di connessione [...]." (Gresleri and Matteoni 1982)

Figure 2: 3D representation of the "Musée Mondial" backside

"Imaginons le visiteur de ce musée : Il est entré dans la parvis de Musée Mondial ; [...] la pyramide, à pic sur le vide émouvant de péristyle, le domine. Il emprunte l'un des grands rampants de gauche ou de droite, il est sur la première grande platef orme." (Otlet and Le Corbusier 1928)

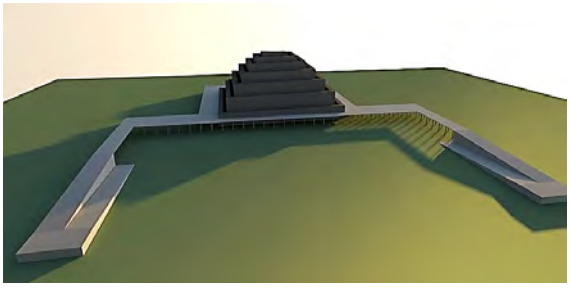


Figure 3: 3D representation of the "Musée Mondial" entrance

5 Conclusions and perspectives

The documents related to "The World City" project are written in many languages (the primary sources in French, and the secondary ones in English/Spanish/Italian), hence, the methodology proposed will work with texts in different languages maintaining the same rules for the 3D model construction. This means that regardless of language descriptions, the correspondence between text and 3D representation will remain the same. This multilingual aspect will be an important characteristic of the work because it will be able to make the 3D model interoperable.

This work aims to disseminate knowledge of the "World City" project thanks to three-dimensional virtual models derived from unstructured information. The proposed methodology may provide the opportunity to increase the users' cultural knowledge as they will learn by playing. Unlike prior works, in fact, the users will be knowledge consumers because the methodology will not be applied to generate generic 3D scenes, but it will work on a specific architectural domain. The final result will be an automatic virtual reconstruction of the "World City".

In addition, the method will lay the basis for future implementations because the learning technique could be applied to other domains to increase user knowledge in diversified fields.

References

Chang, A., M. Savva, and C. D. Manning. 2014. «Learning Spatial Knowledge for Text to 3D Scene Generation». In *EMNLP 2014*, 2028–2038.

Chang, A., et al. 2015. «Text to 3d scene generation with rich lexical grounding». Preprint: arXiv:1505.06289.

Coyne, B., and R. Sproat. 2001. «WordsEye: An Automatic Text-to-Scene Conversion System». In *ACM SIGGRAPH 2001*, 487–496. Addison Wesley.

Coyne, R. E., et al. 2012. «Annotation Tools and Knowledge Representation for a Text-To-Scene System». In *COLING 2012*. http://bit.ly/2ctqMzo.

Cropper, A. 2014. *Identifying and inferring objects from textual descriptions of scenes from books.* OASIcs-OpenAccess Series in Informatics 43. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

Dessai, S., and R. Dhanaraj. 2016. «Text to 3D Scene Generation». *International Journal of Latest Trends in Engineering and Technology (IJLTET)* 6 (3).

Gresleri, G., and D. Matteoni. 1982. *La Città Mondiale. Andersen Hébrard Otlet Le Corbusier*. Venezia: Marsilio.

Kiryakov, A., et al. 2003. «Semantic annotation, indexing, and retrieval». In *The Semantic Web-ISWC 2003*, 484–499. Berlin-Heidelberg: Springer.

Otlet, P. 1929. *Cité Mondiale. Geneva: World Civic Center: Mundaneum*. publication n.133 de l'Union des Associations Internationales. Palais Mondial, Bruxelles- Février 1929.

Otlet, P., and Le Corbusier. 1928. *Mundaneum*. publication n.128 de l'UAI Bruxelles - Août 1928.

Van Halteren, H., ed. 1999. *Syntactic wordclass tagging*. Dordrecht: Kluwer.

I Benedettini: un modello di rappresentazione delle informazioni per un modello di riuso dei beni architettonici

Claudia Cantale, Dipartimento di Scienze Umanistiche dell' Università degli Studi di Catania, claudiacantale.oc@gmail.com
Daniele Francesco Santamaria, Dipartimento di Matematica e Informatica dell'Università degli Studi di Catania, santamaria@dmi.unict.it

1 Cenni storici: il Monastero dei Benedettini e l'intervento di G. De Carlo

A partire dal 1977 il Monastero dei Benedettini di San Nicola l'Arena di Catania ha ricevuto una serie di interventi di recupero e di riadeguamento con l'obiettivo di destinare il plesso ad una nuova funzione d'uso, ovvero di convertirlo in sede della Facoltà di Lettere e Filosofia dell'Università di Catania. Il progetto di recupero porta la firma dell'architetto Giancarlo De Carlo. Per ragioni differenti e in relazione tra di loro, il recupero del Monastero dei Benedettini si è prolungato per 25 anni (formalmente dal 1979 al 2005). Nel 2008 il Progetto Guida di G. De Carlo è stato riconosciuto dalla Regione Siciliana come opera di architettura contemporanea[1], un riconoscimento che sottolinea la grande capacità dell'architetto di interpretare i segni del passato donandogli un nuovo significato ed una nuova funzione d'uso.

Questa lunga fase, in cui si sono alternati l'apertura dei cantieri e i trasferimenti degli uffici universitari, è caratterizzata dalla produzione ingente di documenti (dai rilievi fino alle riprese fotografiche, dai verbali di cantiere fino agli affidamenti degli appalti) che per ragioni pratiche sono stati organizzati in un apposito archivio, nato come atto spontaneo del Geometra Antonino Leonardi, allora Responsabile dell'Ufficio Tecnico Universitario – Sezione Benedettini. Dichiarato dal 2003 "Archivio del Museo della Fabbrica dei Benedettini", è stato inserito nella rete museale d'Ateneo. Da qualche anno sono stati redatti e portati avanti alcuni progetti di valorizzazione delle risorse archivistiche custodite al suo interno, avviando progetti di ricerca o assegnando tesi di laurea sullo stesso. Nella logica di una sua maggiore istituzionalizzazione e per migliorare il suo sistema di comunicazione verso gli studenti del Dipartimento di Scienze Umanistiche si sta lavorando alla realizzazione di un inventario digitale da condividere on line che permetta alla raccolta documentale

[1]DECRETO 23 maggio 2008, Gazzetta Ufficiale della Regione Siciliana, 20 giugno 2008, Anno 62° Numero 28

di essere maggiormente utilizzata ai fini della ricerca. È crescente, infatti, la richiesta di consultazione degli atti e dei disegni custoditi al suo interno da parte di ricercatori e amatori provenienti da differenti atenei e da diversi enti di ricerca.

2 Il progetto: ricavare da un archivio il significato di un restauro

Il progetto di ricerca "*Le Digital Humanities per il modello dei Benedettini*", ancora in corso, punta a organizzare e rappresentare le informazioni contenute in Archivio per avviare un percorso di analisi di alcune delle tensioni politiche e sociali nate attorno al restauro dell'edificio a seguito del cambiamento di destinazione d'uso. De Carlo è considerato in Italia il padre dell'architettura della partecipazione, che lui stesso definisce un'utopia realistica (De Carlo 1972). A Catania sperimentò, però, la progettazione "tentativa". «Tentativa: nel senso che tenta di raggiungere la soluzione procedendo per prove e verifiche, ma anche nel senso che mette in tentazione la situazione con la quale si confronta, per fare emergere i suoi squilibri e per capire come e fino a che punto può cambiare, senza snaturarsi, e raggiungere nuovi equilibri» (De Carlo 1996).

Questa deviazione sul percorso precedentemente tracciato permette di porsi domande sulle cause e opportunità che indussero l'architetto a portare a termine il compito arduo di realizzare una sede idonea per gli studenti e al tempo stesso di ridare una seconda giovinezza ad uno degli edifici più rappresentativi della storia del sudest siciliano, in assenza di un processo di coinvolgimento e di reale partecipazione dei futuri fruitori. Pur non volendo interpretare i segni architettonici, si può comunque affermare che la relazione tra l'architettura e coloro che usano l'architettura, nel caso dei Benedettini, sia stata consolidata in un processo lungo e quasi mai semplice di cui però ancora oggi possono essere riconosciute le cicatrici e le rinunce.

Dall'analisi dei dati tratti da alcuni dei documenti si tenterà, infatti, di dare una possibile interpretazione alle dinamiche nate attorno ad alcuni cantieri (prevalentemente quelli siti nelle zone di confine/frontiera) e quali furono le ragioni sociali e politiche e/o tecniche che hanno reso necessarie alcune modifiche dei progetti già in corso d'opera.

Si è scelto di trattare in questa prima fase tre tipologie di testo: il Progetto Guida "Un progetto per Catania" redatto da De Carlo in cui vengono presentate le soluzioni per il recupero dell'intero edificio e pubblicato nel 1989; i 92 diari di cantiere che l'architetto usava stilare a rientro da ogni visita al Monastero che venivano inviate agli uffici universitari; 105 lettere - in corso di pubblicazione – indirizzate al Geometra Leonardi che descrivono in maniera più intima e scevra da condizionamenti istituzionali le scelte, le ipotesi progettuali e i desideri professionali dell'architetto, in una dimensione amicale e personale. La restituzione vuole mostrare in quali epoche e su quali cantieri si sono concentrati i maggiori investimenti in termini di risorse umane, economiche e intellettuali per risolvere e dirimere conflitti, trovando soluzioni possibili. L'obiettivo finale è quello di pubblicare in *Open Access* e *Open Data* i verbali di cantiere con il riferimento alle fonti archivistiche custodite nell'Archivio del Museo della Fabbrica.

Le tecnologie di processamento dei linguaggi naturali consentono di individuare attori, soggetti e oggetti con precisione nei testi, di individuare le relazioni che tra essi occorro-

no e digitalizzare le informazioni estrapolate. Ma nel caso specifico la realizzazione di un database classico non consentirebbe di ricostruire automaticamente le relazioni che intercorrono tra attori e agenti, ovvero tra i professionisti e i cantieri su cui hanno agito in una arco temporale dato, in quanto la collezione delle informazioni è nata in maniera non organica attraverso la testimonianza e la mediazione del fondatore dell'archivio, il geometra Antonino Leonardi. Per questo motivo si è scelto di strutturare la ricerca dei rapporti e dei nessi tra persone o ditte e i luoghi presenti nei testi sottoposti al processo di analisi attraverso le indicazioni dei *Linked Open Data*, ed in particolare del *Semantic Web*, una implementazione dei *Linked Data* orientata ad internet, che nell'ultimo decennio ha conosciuto un'ampia diffusione in molti campi, sia scientifici che umanistici. In questa visione di diffusione della conoscenza, le informazioni classicamente prodotte per essere interpretate, comprese ed usufruite dall'uomo, diventano portatrici di un significato esplicito e coerente che vengono processate, integrate e condivise automaticamente da agenti software, mantenendo comunque la struttura originaria. Per tale motivo si parla di "dati intelligenti". Di fatto, la semantica delle informazioni consente non solo una migliore diffusione e collaborazione tra ricercatori, ma anche un sistema di ricerca efficiente, preciso ed esaustivo, ampiamente utilizzabile nei più diversi contesti.

La modellazione di queste informazioni avviene tramite *Ontologia*. L'ontologia, in quanto rappresentazione formale, condivisa ed esplicita di concettualizzazione di un dominio di interesse, che permette agli elementi di essere connessi attraverso relazioni stabilite dalla loro semantica, farà sì che le informazioni possano essere modellate da una rappresentazione formale non ambigua. L'applicazione di tecniche di deduzione logica, o più in generale, di ragionamento automatico, consente di estrarre la conoscenza implicita nel modello ma non presente nei dati stessi. La *distant reading*, infatti, a dispetto della *close reading*, consentirà in maniera più rapida e precisa di estrapolare le informazioni contenute nelle differenti fonti da noi trattate migliorando i tempi e il lavoro di analisi del ricercatore. Sulle tre tipologie di testo scelte per avviare la ricerca sulla storia contemporanea del Monastero dei Benedettini e del territorio su cui insiste, si sta tentando di creare un modello, frutto di astrazione e della riduzione in unità minime, basandoci su un processo analitico che ha già portato alla creazione di OntoCeramic (Cantone et al. 2015, 122-127), un sistema (modello) di classificazione automatico della ceramica attraverso le sue forme e i materiali.

In tale maniera il "Modello Benedettini", riferito alle metodologie, alle idee, ai processi di partecipazione che hanno permesso al Monastero di riavere una seconda vita divenendo uno dei luogo più significativi per la città, diviene "modello" di astrazione ontologica per lo studio e l'analisi di sociologia urbana e di urbanistica.

Bibliografia

Bennato, D. 2015. *Il computer come macroscopio. Big Data e approccio computazionale per comprendere i cambiamenti sociali e culturali*. Milano: Franco Angeli.

Cantone, D., et al. 2015. «OntoCeramic: an OWL ontology For ceramics classification». In *Proceedings of the 30th Italian Conference on Computational Logic, CILC 2015, Genova, Italy, July 1-3, 2015, CEUR Workshop Proceedings*, 1459:122–127.

Dato, G., e G. Pagnano. 1984. «Il convento dei Benedettini può ospitare il Magistero?» In *Urbanistica e città meridionale*, a cura di G. Dato. Catania: Culc.

De Carlo, G. 1988. *Un progetto per Catania. Il recupero del Monastero di San Nicolò l'Arena per l'Università*. A cura di D. Brancolino. Genova: Sagep.

Doerr, M. 2003. «The CIDOC CRM - An ontological approach to semantic interoperability of metadata.» *AI Magazine* 24 (3): 75–92.

Felicetti, A., et al. 2013. «Mapping ICCD Archaeological Data to CIDOC CRM: the RA Schema». In *Practical Experiences with CIDOC CRM and its Extensions (CRMEX), Malta, 26 September 2013.*

Frezza, G., cur. 2008. *L'arca futura. Archivi mediali digitali, audiovisivi, web*. Roma: Meltemi Editore.

Guercio, M., S. Pigliaccio e F. Vallacchi. 2010. *Archivi e Informatica*. Lucca: Civita Editoriale.

Guerrini, M., e T. Possemato. 2015. *Linked data per biblioteche, archivi e musei*. Milano: Editrice Bibliografica.

Iacono, A. 2014. *Linked Data*. Milano: Associazione Italiana Biblioteche.

Lamagna, R. 2001. «Riuso del Monastero Benedettino di San Nicolò l'Arena a Catania. Progetto guida per il recupero del monastero: proposte progettuali e nuove configurazioni,» in *ARCO, Manutenzione e recupero della città storica, Atti del IV convegno nazionale*. Roma: Gangemi.

Lazzari, M., et al. 2010. *Informatica umanistica*. Milano: McGraw-Hill.

Magliano, C. 2004. «Metadati: dibattito nazionale e internazionale». In *Futuro delle memorie digitali e patrimonio culturale, Atti del Convegno internazionale, Firenze 16-17 ott. 2003,* a cura di V. Tola e C. Castellani. Roma.

Mannino, F., cur. 2015. *Breve storia del Monastero dei Bendettini di Catania*. Catania: Giuseppe Maimone editore.

Numerico, T., D. Fiormonte e F. Tomasi. 2010. *L'umanista digitale*. Mulino.

Orlandi, T. 1990. *Informatica Umanistica*. Roma: La Nuova Italia Scientifica.

Raieli, R. 2010. *Nuovi metodi di gestione dei documenti multimediali*. Milano: Editrice Bibliografica.

Tola, V., e C. Castellani, cur. 2006. *Archivi informatici per il patrimonio culturale, Atti del Convegno internazionale, Roma 17-19 nov. 2003*. Roma.

Tomasi, F. 2008. *Metodologie informatiche e discipline umanistiche*. Carocci.

Vitali, S. 2004. *Passato digitale. Le fonti dello storico nell'era del computer*. Milano: Bruno Mondadori.

Zevi, B. 1974. «Il convegno di Catania. L'Università mobilita gli architetti». In l'"Espresso", 3 marzo, ripubblicato con il titolo *Dove situare l'università di Catania. No all'emarginazione di professori e studenti*, in "Cronache di architettura" vol. IX, Laterza Roma – Bari, 1975. Roma, pag. 225.

VicoGlossia: Annotable and Commentable Library as a Bridge between Reader & Scholar

Michail Maiatsky, Dr., Scientific Associate in University of Lausanne (Faculté des Lettres), mmaiatsky@gmail.com
Alexey Boyarsky, Dr., Assistant Professor in Leiden University (Lorentz Center); Scientific Associate in CERN, Switzerland; Invited Professor in EPFL, Lausanne; ScienceWISE, alexey.boyarsky@epfl.ch
Natalia Boyarskaya, Dr., Scientific Associate in University of Lausanne (Faculté des Lettres, CROSS), natalia.boyarskaya@epfl.ch

1 Idea

VicoGlossia is a free-access free-content collaborative internet platform. It is a digital library in Humanities (mostly literary and philosophical texts[1]) with semi-automatically established references, cured and refined collaboratively by scientific commenting. For the urgency of such a type of library see (Boot 2009).

2 Aim

The aim of VicoGlossia is to offer to the reader a possibility to be assisted by explicative notes, references, commentaries to the relevant scholarship, facilitating the understanding of the text. The corpus of "classical" texts will form a hypertext serving an interface to different types of references, glosses and comments to various text units (from a word to a whole text).

3 Philosophy

Guidelines, Policies, Ethics of VicoGlossia still remain to be elaborated, using the experience of Wikipedia and other collaborative platforms.

Modesty VicoGlossia is focused on commented works, and commentators are mentioned and celebrated.

Respect to Pioneers VicoGlossia tries to detect forerunners of any given interpretation.

Traceability Chronology of commentaries and interpretations permits to reconstruct the history of interpretive reading of a given text.

[1] The corpus can potentially be enlarged and include visual, cinematographic, choreographic, musical, scientific and other works.

Competence VicoGlossia could become a bridge between scholars and wide-range public. Curious reader can be both a recipient and a supplier of commentaries.

Language Policy Two languages will be dominant within VicoGlossia: a vernacular language of the text-source, and English as *lingua franca*.

4 Advantages

No (premature) consensus No need to try to come to a Wikipedia-like consensus. VicoGlossia aims to keep all different pertinent contributions to understanding of a given text. Instead of Wikipedia's "Neutral Point of View" principle, VicoGlossia advocates rather "Fair Account of Differences". If "Wikipedia is not the place to insert personal opinions, experiences, or arguments", VicoGlossia is such a place.

Reality of science The consensus image of knowledge allows to neutralize passions, but gives a distorted image of science where controversies are *frequent*, *normal* and *fruitful*.

Translingual added value Contrary to Wikipedia, where versions in different languages exist independently, VicoGlossia will permit to form a joint community.

Improvability The result will never be considered as ultimate, and always *in progress*.

Hermeneutic democracy VicoGlossia allows to escape from power abuse frequent in the academic sphere and to assure certain visibility to interpretative minorities.

5 Risks

… are similar to those Wikipedia is constantly fighting against (vandalism, etc.). They are still fewer than in paper editions.

6 Services & possibilities offered by the platform

Comfortable access VicoGlossia aims at an easy and comfortable access to various texts considered as belonging to a "classical heritage", with different possibilities of comparison, cross-commenting etc.

Translations alignment The user will get an opportunity to check all available translations of a given text fragment and its close environment.

Symbolic capital cartography A user will have a possibility to obtain an automatically generated rapid overview of the expert field around a given text: the most respected scholars, the most cited studies, the evolution of hermeneutic of the work.

7 Use

Texts will be accessible under three regimes: 1) Clean text, 2) Text with (non-intrusive) hints, 3) Rich text (in a simplified didactic or a scholar version). According to the specificity of the text, various types of comments will be distinguished, e.g. those pertinent for: a) authenticity, b) chronology, c) genealogy, d) impact, etc. There will be intensive (explicatie) and extensive (recommending further reading) commentaries. According to the types of

anchoring in the text-1, a distinction between 1) holist, 2) supra-phrastic, 3) phrastic & 4) lexical will be drawn. Further classification will be presented in the poster.

In our poster we will present a prototype of the platform: a draft of VicoGlossia facilities inoculated to the Tolstoy Digital project. We will show two vectors of our project:

1. automated mapping of research literature onto original text, i.e. establishing links between text units in the 'classical text' and secondary literature; text mining and other semantic technologies, well elaborated within ScienceWISE / LSIR, at the EPFL, Lausanne (Aberer et al. 2011; Astafiev et al. 2012; Magalich et al. 2016);

2. providing of collaborative tools permeating to comment the text, to amend existing commentaries, to administrate interventions, etc.

8 Technical details

The system can be understood consisting of 3 interfacing solutions:

1. digital library and paper repository

2. multiplatform applications for users (readers)

3. annotation platform for contributors

Digital library and paper repository Software implementations of a digital library are designed for managing large corpora of texts (books, papers, etc.) and corresponding metadata. Centralized repository for academic texts provides higher accessibility of information and faster academic exchange of ideas.

Multiple open source projects exist — most notably, Invenio developed and used by multiple natural sciences institutions (CERN, DESY, EPFL, FNAL, SLAC).

Applications for users The most visible part of the system is the interface that allows users to access texts together with notes, references and commentaries provided by scholars. It is similar to existing open-access digital libraries (Wikisource, Project Gutenberg, etc.), but provides additional features usually met only in single-book annotation projects (commentaries, references, links to other texts, etc.)

Annotation platform The bulk of information presented in user applications is extracted semi-automatically for scientific papers both by parsing explicit references and by ontology-based analysis similar to the ScienceWISE platform. Results then are validated by contributors of the platform.

Additionally, platform provides a semantic workspace for contributors willing to annotate texts manually (possibly, as a part of the paper submission process to the paper repository).

As for hosting and support, locally hosted servers or national cloud-system can be used for hosting. Once the detailed goals and the principles of the mapping are defined and the algorithms to establish links are determined. This algorithm should be implemented in the form allowing robust and powerful performance. The system will include back-end (data base and analytical part connected to it) and a front end (system of interfaces allowing to get inputs from user, organise crowdsourcing and present the results to users

in a convenient and friendly form). Creation of such system requires involvement of a professional envelopment team. A constant technical and humanist support is needed to assure its lasting existence and improvement.

References

Aberer, K., et al. 2011. «ScienceWISE: A Web-based Interactive Semantic Platform for Scientific Collaboration». In *10th International Semantic Web Conference*. Bonn, Germany. http://exascale.info/assets/pdf/ScienceWISEdemo_ISWC_2011.pdf.

Astafiev, A., et al. 2012. «ScienceWISE: A Web-based Interactive Semantic Platform for Paper Annotation and Ontology Editing». In *Extended Semantic Web conference 2012*. Greece. http://2012.eswc-conferences.org/sites/default/files/eswc2012_submission_316.pdf.

Boot, P. 2009. *Mesotext. Digitised Emblems, Modelled Annotations & Humanities Scholarship*. Amsterdam: Amsterdam University Press.

Courtney, N., ed. 2007. *Library 2.0 & beyond. Innovative technologies & tomorrow's user*. Connecticut: Libraries unlimited.

Lang, A., ed. 2012. *From Codex to Hypertext : Reading at the Turn of the 21st Century*. Amherst, MA: University of Massachusetts Press.

Magalich, A., et al. 2016. «ScienceWISE: Topic Modeling over Scientific Literature Networks». Submitted to ACM SIGMOD Record.

Marshall, C. C. 1997. «Annotation: From Paper Books to the Digital Library». In *Proceedings of the Second ACM International Conference on Digital Libraries*, 131–140. DL '97. Philadelphia, Pennsylvania, USA: ACM. ISBN: 0-89791-868-1. doi:10.1145/263690.263806. http://doi.acm.org/10.1145/263690.263806.

Sinatra, M., and M. Vitali-Rosati, eds. 2014. *Pratiques de l'édition numérique*. Montréal: Presses de l'Université de Montréal.

Snyder, I., ed. 1998. *Page to Screen: Taking Literacy into the Electronic Era*. London: Routledge.

Unsworth, J., K. O'Brien O'Keeffe, and L. Burnard, eds. 2006. *Electronic Textual Editing*. New York, NY, USA: MLAA. http://www.tei-c.org/About/Archive_new/ETE/Preview.

“Semplicemente ti ci abitui”. Le molte culture dell’edizione elettronica (con a seguire alcune considerazioni pratiche a riguardo)

Federico Meschini, Università degli Studi della Tuscia, fmeschini@unitus.it

“Young man, in mathematics you don’t understand things. You just get used to them.”

Questa frase, attribuita a John Von Neumann, e dal forte sapore pragmatico, così come l’approccio ingegneristico della macchina omonima, contrapposto alla natura astratta della macchina di Turing, ben si adatta ai numerosi tentativi di definizione dell’edizione elettronica. Scopo di questa relazione è di spostare il focus dei vari tentativi di analisi in relazione sia al “Computational thinking” (Wing 2006), data la sua recente diffusione, sia, riconducendosi a un livello più generale, alla disciplina dell’Informatica Teorica (Papadimitriou 1993); l’importanza di questo approccio inizia ad essere presa in considerazione anche nel settore delle Digital Humanities, e, più in generale, nell’ottica del superamento della divisione tra le due culture (Burnard 2000) o, andando a ritroso, della separazione gutenberghiana dei saperi tra scienze umane e scienze esatte (Chandra 2014).

Le edizioni critiche digitali sono da sempre uno degli aspetti più rilevanti dell’informatica umanistica, per la loro storia, diffusione, e combinazione di riflessioni teoriche e applicazioni pratiche (Robinson 2013a). Quale può essere quindi l’interazione tra questi due aspetti: da un lato quali sono gli strumenti che il pensiero computazionale può mettere a disposizione per una migliore comprensione della natura delle edizioni digitali, e dall’altro come possono queste ultime aiutare a diffondere questa attitudine, giudicata come una delle componenti essenziali per gli “abitanti” del 21° secolo? Da un punto di vista ontologico un’edizione digitale ha lo stesso status di un distributore automatico, dato che entrambi possono essere ridotti ad automi a stati finiti, anche se dal punto di vista epistemologico sono totalmente diversi e rispondo ad esigenze opposte, conoscitive da un lato e di sussistenza dall’altro.

Nonostante la necessità di un approccio astratto sia stata più volte sottolineata in diverse riflessioni teoriche (Meschini 2008), l’uso degli strumenti messi a disposizione dall’informatica teorica è stato fino ad ora pressoché nullo, probabilmente più per una mancanza di familiarità che per effettivi dubbi riguardo un progresso scientifico. D’altro canto, data l’importanza strategica dell’edizione digitale, sia nel mondo accademico e culturale sia in quello editoriale, una migliore comprensione e divulgazione dei suoi aspetti computa-

zionali potrebbe aiutare a ridurre lo spessore del muro che attualmente divide le discipline umanistiche da quelle esatte, aiutandole così ad uscire dalla nicchia sempre più stretta in cui vengono al momento collocate, e recuperando quindi il ruolo di primo piano cui hanno diritto (Thaller 2014).

"Sulla prima pagina è scritto: Nell'affresco sono una delle figure di sfondo."

Basandosi su queste premesse teoriche, il resto della relazione si concentrerà sulla presentazione di un'istanza concreta di un'edizione elettronica, cercando quindi di declinare a un livello idiografico ciò che è stato definito in maniera nomotetica. Il testo preso in esame è "Q", scritto dal collettivo Luther Blissett (1999). Questa scelta è giustificata da diversi motivi: la rilevanza culturale, in quanto è considerata come una delle opere principali del movimento New Italian Epic (Ming 2009b); l'ambientazione europea e nello specifico durante la riforma protestante e la controriforma, e quindi l'estrema rilevanza dal punto di vista storico, culturale e sociologico; l'aderenza con numerosi personaggi ed eventi storici che rendono possibile una contestualizzazione e un arricchimento semantico attraverso fonti di dati esterne; la presenza di diversi livelli di lettura, come le forti analogie tra gli eventi storici e i movimenti di protesta nati a partire dalla fine degli anni '60; la disponibilità di audiolibri e traduzioni in diverse lingue e, non ultima, la modalità copyleft con cui è stato distribuito il romanzo.

Inoltre la recente pubblicazione di un altro libro con in comune molti dei personaggi di Q, "Altai" (Ming 2009a), dà la possibilità di estendere la sottostante piattaforma digitale di pubblicazione in modo da gestire universi narrativi condivisi (Jenkins 2006). La densità sia culturale sia letteraria dell'opera in questione permetterebbe la creazione di un'edizione elettronica, e del relativo modello sottostante, in grado di superare lo status di "digital incunabula" e di sfruttare adeguatamente le effettive possibilità dello strumento computazionale e di un approccio orientato ai dati. Il progetto prevede quindi le seguenti caratteristiche: modalità di lettura lineare e semantica (Tomasi 2013a); localizzazione e contestualizzazione geografica e temporale, rappresentazione grafica delle sequenze narrative; grafo delle relazioni tra i personaggi (Moretti 2011); interfaccia di fruizione crossmediale con sincronizzazione del testo nelle diverse edizioni, supporti e adattamenti (Jewell, Lawrence e Tuffield 2005). Ognuna di queste caratteristiche verrà presentata andando ad analizzare le criticità implicite nei vari approcci tecnologici, insieme ai dati e ai metadati necessari alle relative implementazioni.

Bibliografia

Blissett, L. 1999. *Q*. Torino: Einaudi.

Burnard, L. 2000. «From two cultures to digital culture: the rise of the digital demotic». Online. http://users.ox.ac.uk/~lou/wip/twocults.html.

Chandra, V. 2014. *Geek Sublime: The Beauty of Code, the Code of Beauty*. Minneapolis, MN, USA: Graywolf Press.

Jenkins, H. 2006. *Convergence Culture*. New York, NY, USA: NYU Press.

Jewell, M. O., F. Lawrence e M. M. Tuffield. 2005. «OntoMedia: An Ontology for the Representation of Heterogeneous Media». In *Multimedia Information Retrieval Workshop 2005. SIGIR, ACM.* http://eprints.soton.ac.uk/261009/1/OntoMedia.pdf.

Meschini, F. 2008. «Mercury ain't what he used to be, but was he ever? Or do electronic scholarly editions have a mercurial attitude?» In *International Seminar of Digital Philology*. Edinburgh. http://slideplayer.com/slide/784700.

Ming, W. 2009a. *Altai*. Torino: Einaudi.

— . 2009b. *New Italian Epic. Letteratura, sguardo obliquo, ritorno al futuro*. Torino: Einaudi.

Moretti, F. 2011. «Network Theory, Plot Analysis». *New Left Review* 68. http://newleftreview.org/II/68/franco%C2%ADmoretti%C2%ADnetwork%C2%ADtheory%C2%ADplot%C2%ADanalysis.

Papadimitriou, C. H. 1993. *Computational Complexity*. New York, NY, USA: Pearson.

Robinson, P. 2013a. «Towards a Theory of Digital Editions». *Variants*:105–131. https://www.academia.edu/3233227/Towards_a_Theory_of_Digital_Editions.

Thaller, M. 2014. «Are the Humanities an endangered or a dominant species in the digital ecosystem?» In *Proceedings of the Third AIUCD Annual Conference on Humanities and Their Methods in the Digital Ecosystem*, a cura di F. Tomasi, R. Rosselli Del Turco e A. M. Tammaro. Association for Computing Machinery (ACM). doi:10.1145/2802612.2802613. http://dx.doi.org/10.1145/2802612.2802613.

Tomasi, F. 2013a. «L'edizione digitale e la rappresentazione della conoscenza. Un esempio: Vespasiano da Bisticci e le sue lettere». *Ecdotica* 9:264–286. https://www.researchgate.net/publication/267451270_L'edizione_digitale_e_la_rappresentazione_della_conoscenza_Un_esempio_Vespasiano_da_Bisticci_e_le_sue_lettere.

Wing, J. M. 2006. «Computational Thinking». *Communications of the ACM* 49 (3): 33–35. https://www.cs.cmu.edu/~15110-s13/Wing06-ct.pdf.

Oltre la “concordanza” in un contesto di assistenza all’analisi concettuale: il concetto di “mind” in Peirce

Davide Pulizzotto, LANCI-UQAM, davide.pulizzotto@gmail.com
José A. Lopez, LANCI-UQAM, josaleg@hotmail.com

1 Introduzione, questione di ricerca, contributo

Nell’ambito delle discipline umanistiche, e ancor più particolarmente in filosofia, l’analisi concettuale (AC) è utilizzata per scandagliare le dimensioni, proprietà o componenti di un concetto (Deleuze e Guattari 2005) sviluppato da un autore o un gruppo di autori. Lo studio del concetto di “evoluzione” nell’opera di Darwin, o del concetto di “bellezza” nelle opere letterarie del XIX secolo rappresentano degli esempi d’analisi concettuale. Sono diversi gli approcci metodologici di una AC (Beaney 2015), almeno tanti quante sono le molteplici e divergenti teorie del concetto (Laurence e Margolis 2003). Se tuttavia adottiamo il postulato semiotico secondo il quale un concetto e le sue proprietà si manifestano soprattutto attraverso il linguaggio naturale, allora possiamo affermare che una delle maniere più classiche di effettuare una AC in filosofia si basa sull’analisi di un collezione di testi e, in particolare, sull’analisi della *concordanza* di una *forma canonica* del concetto. Nel caso del *L’origine della specie* di Darwin, per esempio, l’estensione più semplice della forma canonica del concetto di “evoluzione”, sarebbe lo stesso lessema “evoluzione”. La concordanza, invece, sarebbe l’insieme di tutti i segmenti di testo in cui è presente la forma canonica. Il limite principale di un tale approccio è costituito dal fatto che un concetto può manifestarsi anche all’interno di segmenti di testo in cui non è presente alcuna estensione della forma canonica. Esistono pertanto delle *forme non canoniche* del concetto che generalmente sono sconosciute, imprevedibili o difficilmente ritracciabili per mezzo di una concordanza e che sono invece contenute in porzioni di testo che chiameremo peri-segmenti. L’obiettivo del nostro lavoro è di fornire, in un contesto di informatica umanistica, una soluzione a questo tipo di problema.

Il contributo maggiore di una tale ricerca risiede nella costruzione di catene di trattamento dell’informazione che si propongano come strumenti di assistenza all’analisi concettuale in filosofia. Infatti, nonostante l’obiettivo principale di questo lavoro rimanga l’identificazione dei peri-segmenti, il metodo sviluppato fornisce una assistenza all’intero processo di AC. Inoltre, dato che la natura dell’algoritmo utilizzato è indipendente dalla lingua o dallo stile di scrittura, il metodo può facilmente essere applicato in altri contesti, purché tuttavia si presenti una problematica simile a quella dell’analisi concettuale in filosofia. Infine, tale lavoro contribuisce al trasferimento di conoscenze e strumenti dall’intelligenza artificiale alle discipline umanistiche.

2 Letteratura connessa

Le tecniche sviluppate in elaborazione del linguaggio naturale (NLP), machine learning (ML), text mining (TM) o information retrivial (IR) sono utilizzate in filosofia già da alcuni decenni (Bynum e Moor 1998). Alcune recenti ricerche hanno, per esempio, applicato i modelli topici (*topic model*) per l'estrazione della struttura argomentativa nei testi filosofici del XIX secolo (Lawrence et al. 2014); altre hanno invece provato a rispondere a questioni filosofiche, come l'espressione delle relazioni causali (Girju e Moldovan 2002) o la percezione soggettiva del tempo (Schwartz et al. 2015), tramite catene di analisi testuale.

Come è già stato sottolineato, il nostro lavoro si inscrive principalmente in un contesto di analisi concettuale assistita dal computer in filosofia. In questo quadro, alcuni strumenti (Forest e Meunier 2005) o alcune catene di trattamento per l'analisi di un concetto sono già state sviluppate. Tra i primi lavori che hanno utilizzato questo tipo di tecniche si distingue lo studio condotto da McKinnon sul concetto di "destino" in Kierkegard (McKinnon 1973). Altre ricerche più recenti hanno esplorato il concetto di "linguaggio" nell'opera di Bergson (Danis 2012), quello di "evoluzione" in Darwin (Sainte-Marie et al. 2010), quello di "mente" in Peirce (Meunier e Forest 2009) ed infine quello di "management" nell'opera del filosofo giapponese Matsushita (Ding 2013). Da quanto è a nostra conoscenza, non esiste però alcun lavoro che si sia occupato della ricerca dei peri-segmenti, cosicché la letteratura esistente ha solitamente ridotto l'AC allo studio della concordanza della forma canonica di un concetto.

3 Metodo e sperimentazioni

La nostra catena di analisi concettuale assistita dal computer si articola in tre macro-tappe : 1) costruzione di una concordanza a partire della forma canonica, che ci permette di ottenere un primo corpus di lavoro; 2) identificazione delle componenti semantiche che emergono nella concordanza, che corrispondono essenzialmente ai differenti contesti in cui la forma canonica del concetto è utilizzata; 3) ricerca dei peri-segmenti, effettuata per mezzo di un modello classico di information retrivial e di una estrazione delle peculiarità di ogni componente semantica. Questa catena è stata sperimentata sull'analisi del concetto di "mind" nei *Collected Papers* di Ch. S. Peirce.

Sulla versione del corpus ottenuta (Peirce 1994) un trattamento linguistico di base è effettuato (segmentazione, tokenizzazione, stemming, vettorializzazione, etc), al fine di prepararla al trattamento informatico. L'estrazione della concordanza viene effettuata in funzione dell'estensione più semplice della forma canonica del concetto analizzato, che è rappresentata dal lessema "mind". Due corpus di lavoro sono quindi generati. Il primo racchiude i segmenti in cui appare la forma canonica (concordanza) e ne conta 1.323. Il secondo coincide con il corpus restante e contiene tutti i potenziali peri-segmenti, che sono 14.963. L'intero corpus segmentato conta 16.286 porzioni di testo, la maggior parte dei quali si compone di tre frasi e solo una piccola parte di une o due frasi.

La seconda tappa ci permette di identificare le varie tipologie di contesto in cui appare la forma canonica, ognuna delle quali è rappresentato da un gruppo di segmenti che condividono un certo numero di caratteristiche semantiche. Sulla base delle ipotesi generate in seno alla semantica distribuzionale (Fabre e Lenci 2015; Sahlgren 2008), chiameremo questi gruppi di segmenti componenti semantiche del concetto poiché essi, condividendo

alcune proprietà semantiche (come la presenza degli stessi co-occorrenti), modellizzano alcune dimensioni, proprietà o componenti del concetto sotto esame. Questa operazione è dunque effettuata sulla sola concordanza e si basa sull’applicazione di un algoritmo di tipo *hard clustering* (*k-means*), che ci permette di produrre un partizionamento netto e senza sovrapposizioni tra i segmenti. I parametri dell’algoritmo sono determinati da alcuni test di partizionamento valutati con il *Silhouette Coefficient* (Rousseeuw 1987). Nelle nostre sperimentazioni otteniamo dieci cluster, che modellizzano altrettanto dieci componenti semantiche del concetto di “mind”; ognuna delle quali raggruppa un certo numero di segmenti che condividono informazioni simili. L’ultima tappa della nostra catena di trattamento è costituita da due importanti operazioni: a) l’estrazione della peculiarità di ciascuna componente semantica; 2) la ricerca dei peri-segmenti. La prima è realizzata da un algoritmo di classificazione supervisionata, la *Support Vector Machine* (SVM), che è applicato su ciascun cluster con una strategia di tipo one-vs-all (Rifkin e Klautau 2004). Dopo aver stabilito il miglior modello di apprendimento per ogni cluster attraverso una *10-fold-cross-validation*, estraiamo il *weight vector*. Costruito dalla SVM a dei fini di classificazione, il weight vector è già stato utilizzato per selezionare e riponderare le caratteristiche di un gruppo di osservabili. In breve, è un vettore che “sintetizza” l’informazione dei dati osservati (Brank et al. 2002; Chang e Lin 2008) e che, nel nostro caso, modellizza la peculiarità della componente semantica. Questo vettore è infine utilizzato come query in un modello classico di information retrivial (Salton, Wong e Yang 1975), in cui si selezionano gli n vettori più simili alla query per mezzo della similarità del coseno.

4 Valutazione dei risultati e conclusioni

Alla fine del processo otterremo dunque i dieci peri-segmenti più simili ad ognuna delle dieci componenti semantiche. Questi sono stati valutati da un comitato di tre esperti secondo un protocollo di annotazione composto da tre categorie: 1. peri-segmenti direttamente legati al concetto; 2. peri-segmenti indirettamente legati al concetto; 3. peri-segmenti non pertinenti. Sulla base di questo protocollo, abbiamo classificato come pertinenti per l’analisi concettuale i peri-segmenti della prima e della seconda categoria, stabilendo così la percentuale di precisione media raggiunta dal nostro metodo, che è pari a 83,6%. La seguente applicazione web è stata sviluppata per facilitare una esplorazione indipendente dei risultati : `https://pulizzottodavide.shinyapps.io/CACAO_Peirce`.

Il limite principale di questo lavoro è costituito dalla strategia di valutazione che, non potendosi basare su un corpus annotato per la valutazione di catene di trattamento come quella qui sviluppata, si limita al calcolo della precisione. I lavori futuri saranno invece incentrati sullo creazione di un campione del corpus annotato, la cui dimensione rispetterà il 95% intervallo di confidenza e il 5% di margine di errore. Grazie a questo campione potremmo dunque comparare diversi metodi e algoritmi.

Il presente studio mette in evidenza come lo sviluppo di simili metodi di assistenza informatica può agevolare la ricerca in filosofia e il trasferimento di conoscenze dall’intelligenza artificiale verso le scienze umane.

Bibliografia

Beaney, M. 2015. «Analysis». In *The Stanford Encyclopedia of Philosophy*, Spring 2015, a cura di E. N. Zalta. Metaphysics Research Lab, Stanford University. Visitato il 11/04/2016. http://plato.stanford.edu/archives/spr2015/entries/analysis/.

Brank, J., et al. 2002. «Feature selection using support vector machines». In *Proceedings of the Third International Conference on Data Mining Methods and Databases for Engineering, Finance, and Other Fields*. WIT Press.

Chang, Y.-W., e C.-J. Lin. 2008. «Feature ranking using linear SVM». *Causation and Prediction Challenge Challenges in Machine Learning* 2:47.

Danis, J. 2012. «L'analyse conceptuelle de textes assistée par ordinateur (LACTAO) : une expérimentation appliquée au concept d'évolution dans l'œuvre d'Henri Bergson». Tesi di dott., Montréal, Université du Québec à Montréal. http://www.archipel.uqam.ca/4641/1/M12423.pdf.

Deleuze, G., e F. Guattari. 2005. «Qu'est-ce qu'un concept?» In *Qu'est-ce que la philosophie?*, 21–37. Paris: Minuit. ISBN: 978-2-7073-1942-5.

Ding, X. 2013. «A Text Mining Approach to Studying Matsushita's Management Thought». In *Proceedings of The Fifth International Conference on Information, Process, and Knowledge Management*, 36–39.

Fabre, C., e A. Lenci. 2015. «Distributional Semantics Today. Introduction to the special issue.» *Traitement Automatique des Langues*, Sémantique distributionnelle, 56 (2): 7–20. Visitato il 23/07/2016. https://hal.archives-ouvertes.fr/hal-01259695.

Forest, D., e J.-G. Meunier. 2005. «NUMEXCO: A Text Mining Approach to Thematic Analysis of a Philosophical Corpus». *CH Working Papers* 1 (1).

Girju, R., e D. I. Moldovan. 2002. «Text mining for causal relations». In *FLAIRS-02 Proceedings*, 360–364.

Laurence, S., e E. Margolis. 2003. «Concepts and Conceptual Analysis». *Philosophy and Phenomenological Research* LXVI (2): 253–282.

Lawrence, J., et al. 2014. «Mining arguments from 19th century philosophical texts using topic based modelling». In *Proceedings of the First Workshop on Argumentation Mining*, 79–87. Citeseer.

McKinnon, A. 1973. «The conquest of fate in Kierkegaard». *CIRPHO* 1 (1): 45–58.

Meunier, J. G., e D. Forest. 2009. «Lecture et analyse conceptuelle assistée par ordinateur: premières expériences». *Annotation automatique et recherche d'informations* (Paris).

Peirce, C. S. 1994. *The Collected Papers of Charles Sanders Peirce*. Electronic Edition. A cura di C. Hartshorne e P. Weiss. Virginia, U.S.A.: InteLex Corp. Charlottesville.

Rifkin, R., e A. Klautau. 2004. «In defense of one-vs-all classification». *The Journal of Machine Learning Research* 5:101–141.

Rousseeuw, P. J. 1987. «Silhouettes: a graphical aid to the interpretation and validation of cluster analysis». *Journal of computational and applied mathematics* 20:53–65.

Sahlgren, M. 2008. «The distributional hypothesis». *Italian Journal of Linguistics* 20 (1): 33–54.

Sainte-Marie, M., et al. 2010. «Reading Darwin between the lines: a computer-assisted analysis of the concept of evolution in the Origin of species». In *10th International Conference on Statistical Analysis of Textual Data*.

Salton, G., A. Wong e C.-S. Yang. 1975. «A vector space model for automatic indexing». *Communications of the ACM* 18 (11): 613–620.

Schwartz, H. A., et al. 2015. «Extracting Human Temporal Orientation in Facebook Language». In *Proceedings of the The 2015 Conference of the North American Chapter of the Association for Computational Linguistics-Human Language Technologies*.

Potest. Un nuovo sistema di scrittura umanistico

Giovanni Salmeri, Università di Roma Tor Vergata, giovanni.salmeri@uniroma2.it

1 L'attuale panorama dei sistemi di scrittura

Il campo della scrittura al computer è oggi diviso tra due sistemi principali: i programmi di *videoscrittura*, di cui i principali sono Word e LibreOffice, e i programmi di *composizione tipografica*, di cui il principale è LaTeX. Ognuno dei due sistemi ha punti di forza particolari. I programmi di videoscrittura consentono di: fruire di un ambiente integrato (dai primi appunti al testo finale stampato si rimane sempre all'interno del medesimo programma); usare un editor ottimizzato per la scrittura di testi in linguaggio naturale; infine, ovviamente, visualizzare il risultato finale mentre si compone il testo (*WYSIWYG*). Un programma di composizione tipografica consente di: separare la struttura logica del testo dalla sua presentazione grafica; ottenere un risultato finale stampato molto curato; in particolare avere un'eccellente presentazione delle formule matematiche; salvare i propri testi in un formato leggibile e modificabile con qualsiasi editor.

Ciascuno dei due sistemi soffre però della mancanza dei meriti dell'altro. L'uso di LaTeX è laborioso e difficilmente proponibile a chi non sia disposto a considerare la scrittura di un testo analoga alla composizione di un codice informatico. Dall'altra parte, un programma di videoscrittura consente con molta difficoltà di ottenere un risultato tipograficamente buono, usa per i documenti un formato molto complesso e non modificabile con programmi diversi, e il suo funzionamento *WYSIWYG* rende la scrittura di testi strutturati ripetitiva e distraente. A questa inefficienza bisogna sommare un limite sempre più evidente: ad un testo scritto con un programma di videoscrittura non possono essere associati dall'interno del programma stesso metadati arbitrari. Anche quelli previsti sono del resto così invisibili all'utente normale da essere sistematicamente ignorati. In conclusione, fermo restando che nel campo scientifico per motivi comprensibili LaTeX si è imposto come lo standard, non esiste un sistema di scrittura veramente soddisfacente nel campo umanistico[1].

In realtà vi sono almeno due sistemi alternativi che hanno inteso sommare la maggior parte dei vantaggi di entrambi i sistemi: Lyx (creato nel 1995) e TeXmacs (creato nel 1998), entrambi ancora in attivo sviluppo. Sebbene con strategie diverse, entrambi sono basati

[1] Un'interessante conferma viene proprio da questo convegno dedicato alle Digital Humanities e dal modo in cui è stato chiesto di sottoporre questo abstract: è stato fornito un template nel formato proprietario Word (template che come di consueto non fa uso degli «stili» che teoricamente dovrebbero avvicinarsi ad una marcatura semantica), contemporaneamente avvertendo che «la redazione provvederà alla reimpaginazione tramite LaTeX degli abstract». Questa prassi di duplicazione del lavoro è tipica nel campo umanistico e purtroppo resa inevitabile dai limiti degli attuali strumenti.

su LaTeX: Lyx lo usa come *back end*, TeXmacs ne imita il sistema di composizione. Entrambi tuttavia eliminano quello che pare il principale ostacolo all'uso di LaTeX, cioè la sintassi estremamente invasiva: TeXmacs seguendo integralmente il principio *WYSIWYG* (senza tuttavia rinunciare ad informare della struttura logica soggiacente), Lyx scegliendo una presentazione definita *WYSIWYM*, cioè «what you see is what you mean». Nessuno dei due ha però significativamente ridotto l'uso dei programmi di videoscrittura, cosa che in teoria era possibile e che Lyx intendeva esplicitamente fare. In effetti essi, essendo basati (realmente o idealmente) su LaTeX, ne ereditano alcuni problemi fondamentali: la necessità di pensare alla scrittura come alla creazione di «codice» (i manuali di Lyx superano le 300 pagine, quello di TeXmacs le 250 pagine), la difficoltà nel definire nuovi formati e nell'isolarli come entità a sé stanti. Lyx, usando LaTeX come motore di composizione, porta con sé inoltre anche la fragilità di quest'ultimo. Sia Lyx sia TeXmacs, inoltre, condividono con LaTeX l'orientamento prevalente alla stampa su carta. Questo problema appare oggi particolarmente significativo anche per i programmi di videoscrittura: la maggior parte dei testi oggi non viene prodotto per essere stampato su carta, o perlomeno non esclusivamente, ma piuttosto per altri formati elettronici. Dividere il testo e la sua struttura logica dal formato non appare dunque più una pignoleria teorica, ma una necessità pratica.

2 I linguaggi di marcatura leggeri

Due fatti nuovi avvenuti alla metà degli anni 90 hanno posto le basi per nuovi sviluppi. Nel 1996 viene inventato il linguaggio CSS (Cascading Style Sheet) e nel 1998 l'HTML viene ridefinito come linguaggio di marcatura esclusivamente semantico; nel 1995 l'invenzione dello «wiki», che seguendo il principio «do the simplest thing that could possibly work» usa non l'HTML, ma piuttosto un «linguaggio leggero di marcatura». Il passo successivo evidentemente doveva essere applicare in questo modo il principio della marcatura semantica alla scrittura orientata alla stampa. A partire dagli stessi anni in effetti sono stati ideati linguaggi di marcatura leggeri con questo scopo: tra gli altri txt2tags, AsciiDoc, reStructuredText. Il caso di Markdown, ideato nel 2004, è un po' diverso: questo linguaggio viene ideato come una maniera semplificata di scrivere le parti più comuni dell'HTML; poi a causa dei suoi meriti di semplicità è stato progressivamente esteso. Il punto di arrivo può essere considerato Pandoc: un traduttore «universale», che usa come linguaggio principale Markdown e ne permette la traduzione in molti altri linguaggi, tra cui LaTeX e HTML; inoltre Pandoc arricchisce Markdown fino a renderlo un linguaggio di marcatura sufficiente a tutti gli usi comuni di scrittura (incorporando anche un raffinato strumento di gestione delle bibliografie, analogo a BibTeX).

Recentemente la combinazione Pandoc + LaTeX è stata proposta come il nuovo sistema di scrittura accademico. In questo modo infatti vengono conservati tutti i meriti di LaTeX aggiungendo tuttavia il vantaggio di una scrittura molto più semplice, dovuta appunto al carattere «leggero» della marcatura. Inoltre, Pandoc supera il limite di una scrittura prevalentemente orientata alla stampa su carta, avendo tra i suoi formati di uscita, per esempio, HTML, ePub e DocBook. Ciononostante, neppure queste proposte sembrano aver avuto il successo sperato: per esempio l'ambizioso progetto ScholarlyMarkdown, lanciato con lo slogan «Let's keep the dream of academic writing using Markdown alive» (Lin 2014), sembra essere stato abbandonato dopo pochi mesi.

Le cause di questo insuccesso paiono almeno di due ordini. Anzitutto, come nel caso di Lyx e di TeXmacs, vengono ereditati diversi dei problemi di LaTeX, anzitutto la complessità del sistema soggiacente e le difficoltà legate alle definizioni dei formati. Inoltre, proponendo semplicemente Pandoc + LaTeX come sistema di scrittura si omette di riconoscere quelli che abbiamo citato come i due *primi* vantaggi di un sistema di videoscrittura: il fatto che in esso si usufruisca di un ambiente integrato, al cui interno vi è un editor ottimizzato per il linguaggio naturale. «Usate il vostro editor preferito per scrivere il testo Markdown» e «dopo aver scritto il testo, date il seguente comando da terminale» sono le istruzioni che condannano all'insuccesso un sistema di scrittura: esse infatti escludono dai possibili utenti tutti coloro che usano un programma di videoscrittura e che *non* hanno un editor preferito e *non* sanno che cos'è un terminale. In ogni caso, quasi tutti gli editor di testo sono progettati avendo in mente le esigenze della scrittura di codice, e non di linguaggio naturale.

3 Un nuovo sistema: Potest

Potest (Salmeri 2016) rappresenta il tentativo di trarre le giuste lezioni dai tentativi passati e di giungere ad una soluzione che sia in grado di rispondere, meglio di qualsiasi altro sistema attualmente esistente, alle esigenze della scrittura di testi strutturati (articoli, tesi, libri) in materie umanistiche. Il punto di partenza è costituito dall'uso di Pandoc come motore di conversione e di Markdown come linguaggio leggero di marcatura. I motivi sopra esposti mantengono infatti la loro validità.

Riguardo ai problemi legati a LaTeX, si è concluso che essi potevano essere superati in un solo modo: abbandonando LaTeX come motore di composizione per la stampa su carta. La scelta è caduta dunque su una via completamente diversa: HTML+CSS come formato principale di uscita. Fino a poco tempo fa questa scelta per la stampa su carta sarebbe stata impossibile o temeraria. Da quando però con la versione 3 del linguaggio CSS e soprattutto con lo sviluppo del motore di composizione Prince (già PrinceXML) si ha la possibilità di una conversione di eccellente livello di un sorgente HTML in un file PDF impaginato, la scelta appare naturale o almeno possibile. Il fatto che Prince incorpori al suo interno alcuni algoritmi fondamentali di LaTeX fa sì che il risultato possa essere nei casi comuni praticamente indistinguibile dal punto di vista della qualità tipografica. Scegliere CSS come linguaggio di descrizione della pagina significa inoltre avere la strada spianata verso una facile definizione di nuovi formati. Il fatto che il file HTML+CSS sia generato da Pandoc permette inoltre una flessibilità senza limiti: Pandoc permette infatti la scrittura di *back end* in Lua, nei quali può essere incorporata qualsiasi logica (è per esempio banale riordinare in qualsiasi modo i metadati di un testo, o far in modo che a partire da essi ne vengano calcolati altri). I file di supporto di Prince sono stati inoltre modificati in maniera da permettere, contemporaneamente a qualsiasi altra lingua, la corretta spezzatura in fin di riga del greco classico e, insieme con l'italiano, anche del latino: una facilitazione che per quanto ci risulta non è finora prevista in nessun altro sistema e che guarda direttamente ad alcuni frequenti usi umanistici. Per creare invece un ambiente di lavoro è necessario considerare un editor come parte integrante del sistema di scrittura. Dopo averne preso in esame molti, si è trovata la soluzione ideale in Textadept, un editor «minimalista» basato su Scintilla, disponibile per tutti i maggiori sistemi operativi. Il minimalismo di Textadept ha risolto fin dall'inizio due problemi degli attuali programmi di videoscrittura: la lentezza

e lo sviluppo di interfacce sempre più complesse e sovraccariche (che sottraggono spazio spesso prezioso nello schermo): Potest si presenta così con un editor dall'interfaccia semplicissima ed estremamente veloce. Ma soprattutto il fatto che Textadept sia scritto quasi interamente in Lua significa che in esso non c'è limite alla personalizzazione: intere parti possono essere in maniera relativamente facile modificate, tolte, aggiunte o sostituite. È così che un editor destinato principalmente alla scrittura di codice si è potuto trasformare in un editor ottimizzato per la scrittura di testi in linguaggio naturale e avente sott'occhio le esigenze di usabilità.

Alcune funzioni sono state aggiunte per facilitare l'inserimento di caratteri utili per l'uso umanistico: per esempio un sistema estremamente efficiente per scrivere in greco, cirillico o IPA, oppure combinazioni di tasti che inseriscono lo «spazio non separabile» e il «trattino opzionale», oppure una raffinata funzione che inserisce le virgolette tipograficamente corrette secondo il livello e la lingua prescelta (distinguendo tra tredici diverse consuetudini nazionali). Altre funzioni ancora sono destinate esplicitamente alla manipolazione di testi Markdown: per esempio il comando per inserire i metadati, o per trasformare le note da «interne» ad «esterne» e viceversa (una funzione questa che rende la scrittura ed elaborazione delle note più semplice e flessibile che in un programma di videoscrittura), oppure infine per trasformare una tabella nella matrice trasposta (permettendo quindi di operare in maniera immediata sulle colonne, un'operazione notoriamente impossibile o laboriosissima negli attuali editor di testo).

Una categoria a parte è costituita infine dai comandi di importazione e soprattutto esportazione: il primo permette di trasformare in formato Markdown un testo LaTeX o Word; il secondo di trasformarlo in uno dei formati previsti: PDF per la stampa (via Prince), oppure HTML per la pubblicazione in rete, docx per i casi in cui esso è richiesto, oppure in futuro altri ancora (DocBook, TEI ecc.). In questo modo l'utente si troverà sempre all'interno del medesimo ambiente di lavoro, senza dover sapere nulla del fatto che dietro le quinte il sistema Potest è costituito dalla collaborazione di tre programmi principali (una versione molto modificata di Textadept, Pandoc, Prince) e altri secondari.

4 Conclusioni

È in grado dunque il sistema Potest di rispondere alle necessità per le quali è stato pensato? La risposta sembra positiva. Finora sono stati elaborati principalmente due formati per la composizione di tesi di laurea del corso di laurea in Filosofia di Tor Vergata, e gli esperimenti condotti hanno mostrato che fin da oggi Potest, rispetto ai sistemi di scrittura esistenti, almeno offre risultati di qualità tipograficamente migliore e in tempi molto più rapidi. Nel tempo più rapido consideriamo anche il periodo di apprendimento: con l'editor di Potest si può cominciare a scrivere immediatamente (tutti i principali comandi di modifica sono stati intenzionalmente resi identici a quelli dei programmi di videoscrittura) e apprendere la marcatura Markdown non richiede più di qualche minuto. La guida completa di Potest occupa solo una quarantina di pagine, che includono però molti consigli di scrittura generici. Alla fine una «Guida telegrafica per l'impaziente», sufficiente per scrivere la maggior parte dei testi, è costituita da appena una pagina. Notiamo infine che il sistema Potest è multipiattaforma: i programmi su cui si basa sono disponibili per Linux, OSX e Windows e tutto il codice aggiunto tiene conto delle rispettive differenze. Coloro che finora hanno

provato Potest, provenendo tutti da una lunga abitudine con i programmi di videoscrittura, lo hanno definito più facile da usare, più veloce, più adatto. Diversi hanno notato che la composizione di testi diventa più «piacevole» e che in essa si è meno distratti.

Oltre a soddisfare meglio le esigenze del singolo utente, Potest sembra rispondere meglio anche a quelle della comunità scientifica. I testi composti in questo modo possono infatti non solo essere immediatamente trasformati in qualsiasi formato strutturato di uscita (dunque per basi di dati, depositi istituzionali e così via), ma sono anche pronti per accettare metadati arbitrari, necessari per un'archiviazione e condivisione efficiente. Il dialetto Pandoc di Markdown permette infatti di includere dati in formato YAML, e in Potest essi sono automaticamente inseribili nel documento e anzi necessari per il suo completamento: qualsiasi utente non avrà dunque nessuna difficoltà ad inserire quelli necessari, perché fin dall'inizio è abituato ad usarli per le informazioni essenziali (autore, titolo, ecc.) del documento che sta scrivendo. Un problema cruciale che non è stato ancora affrontato è quello della gestione automatica dei riferimenti bibliografici. Essa è già fornita di per sé da Pandoc, ma si è scelto di non inserirla in Potest finché non si sarà valutato quale sia la forma di uso più facile e con un risultato più adatto all'uso umanistico, anche a costo di doverla riprogrammare daccapo: non sembra infatti opportuno prevedere un sistema di riferimenti bibliografici che sia, anche solo per il primo uso, più laborioso di un inserimento manuale.

Un grande interesse nelle Digital Humanities è dedicato giustamente all'edizione *digitale* di testi, alla loro analisi, alle infrastrutture che ne favoriscano l'accessibilità e lo studio. È però paradossale che il mondo umanistico nella sua vita quotidiana, di cui una parte importante è costituita dalla semplicissima produzione digitale di testi, sia costretto ad affidarsi a strumenti che hanno poco a che fare sia con le sue specifiche esigenze, sia con quelle più vaste della condivisione della cultura, e che generano testi che riproducono, benché su scala diversa, le stesse difficoltà delle opere prodotte prima dell'avvento dell'informatica. Mentre ovviamente pochi di coloro che operano nel campo umanistico sono direttamente interessati alle Digital Humanities, praticamente tutti *scrivono*: e per questo immenso campo molte cose ancora non sono abbastanza pensate. Potest vuole costituire un piccolo passo avanti in questa direzione.

Bibliografia

Cottrell, A. 1999. «Word Processors: Stupid and Inefficient». Online. http://ricardo.ecn.wfu.edu/~cottrell/wp.html.

Cunningham, W. 2014. «WikiWikiWeb». Online. http://c2.com/cgi/wiki/FrontPage.

Flynn, P. 2014. *Human Interfaces to Structured Documents. The usability of software for authoring and editing*. Cork: University College Cork. http://hdl.handle.net/10468/1690.

Grainger, C. 2014. «Writing academic papers in Markdown using Pandoc». Online. http://blog.cigrainger.com/2014/07/pandoc-markdown.html.

Gruber, J. 2002. «Markdown». Online. https://daringfireball.net/projects/markdown.

Healy, K. 2014. «Plain Text, Papers, Pandoc». Online. https://kieranhealy.org/blog/archives/2014/01/23/plain-text.

Hilburn, B. 2014. «LaTeX Needs To Be Reborn». Online. http://bhilburn.org/latex-needs-to-be-reborn.

Hodgson, T. 2015. «Try Pandoc instead of Word for your research writing». Online. https://doctoralwriting.wordpress.com/2015/10/06/try-pandoc-instead-of-word-for-your-research-writing.

Hoeven, J. van der. 2001. «Gnu TeXmacs. A free, structured, wysiwyg and technical text editor». *Cahiers GUTenberg* 39–40:39–50.

Iosad, P. 2015. «Getting off the Word standard: Your academic life in plain text». Online. http://www.anghyflawn.net/teaching/2015/ilw-pandoc.

Kastrup, D. 2002. «Revisiting WYSIWYG Paradigms for Authoring LaTeX». *TUGboat* 23 (1): 57–64.

Knauff, M., e J. Nejasmic. 2014. «An Efficiency Comparison of Document Preparation Systems Used in Academic Research and Development». *PLoS ONE* 9 (12). doi:10.1371/journal.pone.0115069.

Levi, J. 2013. «Why text software is outdated. My thoughts on why (and how) typing could be better». Online. https://medium.com/@drummerjolev/why-text-software-is-outdated-265d23d91254.

Lin, T. T. 2014. «ScholarlyMarkdown. Make Academic Writing Less Frustrating». Online. http://scholarlymarkdown.com.

LYX Team. 2006. «Introduction to LYX». Online. http://wiki.lyx.org/LyX/LyXHelpDocuments.

MacFarlane, J. 2016a. «Pandoc. A universal document converter». Online.

MacFarlane, J. 2016b. «CommonMark. A strongly defined, highly compatible specification of Markdown». Online. http://commonmark.org.

Mitchell. 2016. «Textadept. A fast, minimalist, and remarkably extensible cross-platform text editor». Online. http://foicica.com/textadept.

Pathirage, M. 2015. «Academic Writing With Markdown, Pandoc and Emacs». Online. https://milinda.svbtle.com/academic-writing-with-markdown-pandoc-and-emacs.

Salmeri, G. 2016. «Potest». Online. http://mondodomani.org/potest.

Schuetzler, R. 2015. «LaTeX vs Word (Again)». Online. https://www.schuetzler.net/blog/latex-vs-word-again.

YesLogic Pty. Ltd. 2016. «Prince». Online. http://www.princexml.com.

Interrogare la *varia lectio*. La Commedia del Boccaccio

Elena Spadini

Tre manoscritti conservano il testo della Commedia dantesca copiata da Giovanni Boccaccio: Toledo, Archivo y Biblioteca Capitulares, Zelada 104, 6; Firenze, Biblioteca Riccardiana, 1035; Città del Vaticano, Biblioteca Apostolica Vaticana, Chigiano L VI 213. L'analisi sistematica del testo del poema dantesco nei tre manoscritti del Certaldese, svolta attraverso la collazione integrale dei tre codici, consente uno studio dell'attività di Boccaccio come copista, editore e interprete della Commedia; essa permette inoltre di gettare nuova luce, entro un quadro completo della variantistica, sulle relazioni genetiche attraverso l'analisi puntuale della varia lectio interna a questo ristretto gruppo di codici, di notevole interesse storico-letterario e linguistico.

Per organizzare e interrogare i dati emersi dalla collazione, è stata creata una basedati relazionale, in cui ad ogni categoria di analisi delle varianti corrisponde una tabella. La basedati permette di registrare sia le informazioni relative alla singola lezione (es. in rima o meno, errore di anticipazione, ripetizione da un verso precedente, presente nel resto della tradizione), sia quelle relative al rapporto tra le varianti (es. categoria di cambiamento).

I primi risultati del progetto, in corso, sono visionabili alla pagina accessibile all'indirizzo: http://boccacciocommedia.it.

Bibliografia

Spadini, E. 2016a. «Annotating document changes». In *DChanges '15: Proceedings of the 3rd International Workshop on (Document) Changes: modeling, detection, storage and visualization*. New York: ACM.

Tempestini, S. 2016. «Boccaccio copista e interprete della 'Commedia'». In *Intorno a Boccaccio/Boccaccio e dintorni. Atti del Seminario internazionale di studi. Certaldo Alta, Casa di Giovanni Boccaccio, 11 settembre 2015*. In press.

Modelli e Metodologie

Models and Methods

Talks

Questioni teoriche e problemi pratici di ecdotica computazionale: l'edizione dello Zibaldone Laurenziano di Giovanni Boccaccio

Roul Mordenti, Università di Roma "Tor Vergata", mordenti@uniroma2.it

La filologia informatica comporta due campi di problemi:

1. che la modalità digitale configura un testo di natura differente rispetto a quello del manoscritto o della stampa;
2. che occorra utilizzare le potenzialità della macchina informatica per determinare significativi incrementi conoscitivi (grazie alla codifica).

Gli inizi della filologia informatica dimostrano che un limite teorico di conservatività (punto 1) rappresenta il principale ostacolo a sviluppare le potenzialità della computazione (punto 2).

A partire da una distinzione concettuale fra Informatica Umanistica (HC = *Humanist Computing*) e Digitalizzazione degli oggetti testuali (DH = *Digital Humanities*), si propone l'esperienza di edizione dell'autografo *Zibaldone* di Boccaccio (BML, Plut. 29.8).

L'ipotesi è che una codifica adeguata dell'evoluzione della morfologia degli alfabemi boccacciani (già segnalata dalla linea filologica Barbi-Ricci-Branca) e del sistema paragrafematico possa permettere una datazione analitica dei 55 segmenti che compongono lo *Zibaldone* e contribuire ai problemi di attribuzione tuttora aperti. Tali acquisizioni possono rivestire un rilevante significato critico, sia sull'intertesto boccacciano sia considerando il riuso intratestuale operato da Boccaccio dei materiali dello *Zibaldone*.

La descrizione del lavoro svolto (compresa la creazione di una "tastiera dedicata") delineerà, sulla base dell'esperienza, le linee di una nuova procedura ecdotica incentrata sull'informatica, diversa da quella legata all'epistemologia, prima ancora che alla tecnologia, di Gutenberg.

Modelli di dialogo: testi virtuali e dubbi sull'epistemologia digitale

Luciano Longo, Università di Palermo

Chi nel presente libro cercasse una ricetta universale
per l'edizione critica, si troverebbe deluso.
(Pasquali, Storia, 21)

Se insegni, insegna anche a
dubitare di ciò che insegni
(O. y Gasset)

L'intervento proposto si interroga dapprima sulla possibile applicazione in ambito digitale della nozione di *testo unico* o di *testo definitivo* e sul concetto epistemologico di *ultima volontà d'autore*; successivamente la riflessione si concentrerà sull'edizione critica dal testo come *prodotto per la lettura* al testo come *processo creativo*. L'esemplificazione verterà su testi intesi come *mobili, dinamici, virtuali*, e come *atto di pensiero*; essi sono raggruppati in quattro tipologie. I testi oggetto dell'esemplificazione sono il prodotto di quattro progetti *in itinere*; di essi si portano i primi risultati raggiunti.

Prima tipologia: testi inediti in cui non è espressa *l'ultima volontà d'autore*, nei quali i processi compositivi risultano fluidi e instabili: è il caso di alcuni autografi inediti di Attilio Bertolucci risalenti agli anni '30-'40 del Novecento;[1]

Seconda tipologia: testi inediti che oltre a possedere le caratteristiche sopra citate pongono problemi di rappresentazione e di intelligibilità del testo, oltre a un problema di possibile *riscrittura coatta* a scopo difensivo: l'esempio preso a esame è il manoscritto autografo *2Qq E 29* (sec. XVII ex.) di Teresa di S. Geronimo che riporta il testo de *Il Castello dell'anima, diviso in tre libri per l'anime incipienti, proficienti e perfette, opera mistica di suor Teresa di S. Geronimo monaca terziaria de' carmelitani scalzi*[2];

Terza tipologia: testi preparatori di un progetto letterario non realizzato e documentato da diversi testimoni, uno dei quali viene pubblicato senza espressa *volontà dell'autore* ma che esprime la *volontà dell'editore*: in questo caso verrà presentato il cosiddetto «manoscritto di Populonia» (1959-1961) di Elio Vittorini[3];

[1]Luciano Longo, "Gli «scartafacci» di Attilio Bertolucci. Edizione critica digitale" (PhD project diss., Università degli Studi di Palermo, 2017).

[2]Suor Teresa de' Cermelitani Scalzi, Castello dell'anima. Edizione critica, ed. Rosa Casapullo and Luciano Longo et al., (Alessandria: Casa Editrice dell'Orso, 2015).

[3]Luciano Longo, "Molteplicità testuale e movimenti compositivi nel «ms. di Populonia»: ipotesi di un lavoro digitale", Convegno: "Vittorini nella città Politecnica", Milano 19-20 febbraio, Università degli Studi di Milano- Università Cattolica del Sacro Cuore Milano.

Quarta tipologia: testi editi che presentano testimoni che esibiscono in alcuni casi una testualità autonoma rispetto all'edizione definitiva a stampa, in altri una testualità nascosta volontariamente dall'autore e in altri casi ancora una testualità interrotta; l'attenzione si è concentrata su *I Viceré* di Federico De Roberto, testimoniato da due redazioni manoscritte (un ms. datato 1892 e un secondo datato 1893) e da tre edizioni a stampa: 1894, Vecchi Galli- Catania; 1920, F.lli Treves, Milano (T920); 1935, F.lli Treves- Milano[4].

I testi verranno presentati secondo diversi modelli di codifica in XML-TEI. che ha visto nella sua prima fase un lavoro sulla semantizzazione e sulla concettualizzazione dei *valori* degli attributi dei *tag* e poi successivamente sulla loro rappresentazione, e infine sulla potenziale dinamica ipotestuale. La difficoltà maggiore del modello di marcatura dei testi sopra citati nasce dal fatto che le redazioni manoscritte e/o dattiloscritte presentano una straordinaria molteplicità di interventi correttivi immediati a cui si sovrappongono interventi tardivi, varianti sovrapposte di tipo instaurativo, sostitutivo e ri-propositivo che rendono molto complessa la trasposizione in XML-TEI. A questo si deve aggiungere una diversità di inchiostrazione e una struttura compositiva dei *testimonia* che non è quasi mai speculare e rende complesso il lavoro di individuazione di strutture narrative contigue. I modelli di codifica proposti vogliono rappresentare l'atto di ricostruzione del testo come *processo*, o per meglio dire, come atto di rivitalizzazione di un percorso scrittorio che da *atto* si fa *attività* e da *attività* diviene *processo* per esaurirsi in *fatto*, in prodotto finito o interrotto. L'edizione digitale proposta conterrà tre diverse informazioni: 1. la codifica del documento elettronico e dei testimoni; 2. la codifica base del testo; 3. la codifica del sistema variantistico dei testimoni.

Le caratteristiche dei *testimonia* presi in esame conducono subito a una prima riflessione per la strutturazione della codifica: i documenti testuali possedendo molteplici stesure manoscritte con varie caratteristiche, i fenomeni correttori da marcare conducono ad affrontare seri problemi di *overlapping*. Inoltre, un'altra problematica affrontata si basa sulle correzioni che essendo multiformi non permetteno un agevole confronto tra le diverse stesure, e rendono difficoltosa la ricostruzione della tessitura testuale, sia a livello di lemma/variante sia a livello di segmentazione testuale. Per questo motivo si sceglie di codificare le campagne correttorie non registrando soltanto il fenomeno, ma cercando di codificarne le motivazioni e la stratificazione tramite l'uso di attributi. Il lavoro evidenzia il fenomeno correttorio e di conseguenza del suo strutturarsi in strati e testualità, su tre livelli: 1. codifica dei fenomeni correttori; 2. codifica delle campagne correttorie; 3. codifica interpretativa tramite l'uso di attributi per entrambi i fenomeni.

L'intervento non solo propone quindi un possibile modello di codifica e di rappresentazione, ma si pone in dialogo con l'ecdotica tradizionale cercando di interrogarsi su alcuni dubbi metodologici alla base di una edizione critica. La proposta, tramite i testi presi in esame, cercherà di riflettere sul peso dell'atto ricostruttivo rispetto all'atto costruito (*il testo definitivo*); sulla tipologia di testo che si edita, se bisogna intenderlo come finito o "interrotto"; e sulle potenzialità "concorrenziali" che può avere la procedura di codifica e l'utilizzo di *tool* e di *output* rispetto alle testualità presentate da una edizione a stampa.

[4]Luciano Longo, "The 'insane' manuscript of I Viceré by Federico De Roberto", Convegno: SIS Biennal Conference, Oxford 28-30 September, University of Oxford.

Bibliografia

Brambilla, S., e F. Maurizio, cur. 2009. *La filologia dei testi d'autore. Atti del seminario di studi (Università degli studi di Roma Tre, 3-4 Ottobre 2007)*. Firenze: Cesati.

Contini, G. 1948. «La critrica degli scartafacci». *Rassegna d'Italia* 3 (10–11): 1048–56, 1155–60.

— . 1986. *Breviario di ecdotica*. Milano-Napoli: Riccardo Ricciardi Editore.

Fiormonte, D. 1995. «Varianti elettroniche». *Italiano & oltre* 2:87–94.

Giunta, C. 1997. «Prestigio storico dei testimoni e ultima volontà dell'autore». *Anticomoderno* 3:169–198.

Isella, D. 1987. *Le carte mescolate. Esperienze di filologia d'autore*. Padova: Liviana.

Italia, P. 2005a. «L'ultima volontà del curatore. Considerazioni sull'edizione dei testi del Novecento I». *Per leggere* 5 (8): 191–233.

— . 2005b. «L'ultima volontà del curatore. Considerazioni sull'edizione dei testi del Novecento II». *Per leggere* 5 (9): 169–198.

McGann, J. 2006. *La letteratura dopo il www. Il testo letterario nell'era digitale*. Bologna: Bonomia University Press.

Mordenti, R. 2001. *Informatica e critica dei testi*. Roma: Bulzoni.

Pierazzo, E. 2015d. *Digital scholarly editing: Theories, models and methods*. England: Ashgate P. L.

Quale futuro per le edizioni digitali? Dall'haute couture al prêt-à-porter

Elena Pierazzo, Università Grenoble Alpes

Le edizioni digitali si stanno affermando sempre di più come prodotti scientifici di alto valore, ma la loro creazione rimane privilegio di pochi. Le edizioni digitali si presentano come oggetti complessi, curati tecnicamente, dalle multifunzionalità e che offrono molto ai loro utilizzatori. Sono oggetti unici, personalizzati, che sono in un certo senso condannati a essere innovativi e diversi di volta in volta, pena il non ottenimento di fondi di ricerca necessari alla loro realizzazione. Questa situazione da un lato ha consentito e ha cavalcato lo sviluppo tumultuoso delle tecnologie digitali, dall'altro sta impedendo una vera diffusione di tali metodologie e lo stabilirsi di una folta comunità di filologi digitali. Preparare un'edizione digitale non è semplice e richiede non solo delle competenze filologiche, ma anche competenze tecniche avanzate (XML, XSLT, RDF, manipolazione di immagini ...), nonché l'accesso a risorse di tipo infrastrutturale (server, domini, ecc.); a questo si aggiungono problemi legati al mantenimento delle risorse sul lungo periodo e di accettazione accademica dei prodotti della ricerca. Il risultato è che la produzione di edizioni digitali è condizionata dall'accesso a notevoli fondi di ricerca oppure (o in aggiunta) alla disponibilità in loco di centri di ricerca sul digitale, che sono da un lato ancora abbastanza rari e dall'altro condizionano spesso il loro supporto all'ottenimento di fondi. Ne consegue che larghe fasce di editori sono tagliate fuori dai più innovativi sviluppi dell'indagine ecdotica rappresentati dalle edizioni digitali; i più penalizzati in questa situazione sono in particolare i giovani, dottorandi e post-dottorandi, che pur avendo l'entusiasmo per il nuovo e il digitale, si trovano spesso nell'impossibilità ad accedere ad adeguati supporti di ricerca. Da molte parti si lamenta la mancanza di software e strumenti facili da usare e che possano limitare la necessità da parte degli editori di fare tutto da soli, dalla codifica, alla trasformazione in HTML, al web design, alla creazione e gestione dei metadati, delle API, alla gestione dei server, e in effetti ci si potrebbe chiedere come mai con quasi 30 anni di ricerca nel settore delle edizioni digitali ci siano a tutt'oggi un numero così limitato di strumenti di tale genere[1].

La verità è che tali strumenti non sono di facile realizzazione: differenze disciplinari e culturali rendono complicata la realizzazione di strumenti adatti per un numero economicamente rilevante di ricercatori[2], a cui si aggiungono questioni più pragmatiche come la difficoltà di ottenere fondi di ricerca per la creazione di strumenti informatici e ancora di più per garantirne il mantenimento e l'aggiornamento richiesto dal vorticoso cambiamento dei sistemi operativi e dei browser. È chiaro come questa situazione sia insostenibile: se le edizioni digitali debbono essere necessariamente innovative, "ground-breaking" e personalizzate, è chiaro che queste non potranno mai stabilirsi come il metodo normale di

[1] Fra cui si vedano per esempio: la Versioning Machine (http://v-machine.org); Juxta e Juxta commons

[2] Si veda Tara Andrews, *The Third Way. Philology and Critical Edition in the Digital Age*, "Variants", 2013 (10), pp. 61-76.

produzione editoriale con grave conseguenze per l'intera disciplina. Le soluzioni a questo problema cominciano fortunatamente a emergere e sono offerte in particolare dal settore bibliotecario e da quello dell'editoria. La biblioteca universitariadi Cambridge[3], per esempio, offre supporto tecnico e una piattaforma unificata per la pubblicazione di edizioni digitali molto diverse fra loro ma tutte accessibili a partire dallo stesso portale. Tale approccio consente di limitare il numero di conoscenze tecniche richieste al ricercatore, che può quindi concentrarsi sull'oggetto della sua ricerca: il testo. Una biblioteca ha inoltre una vocazione alla conservazione sul lungo periodo, il che contribuisce ad alleviare i problemi di durabilità dell'oggetto digitale. Tale soluzione impone però dei compromessi, in particolare quello di dover limitare le personalizzazioni e di accettare un'interfaccia standardizzata, oltre, e forse soprattutto, il rinunciare a priori alle potenzialità più sperimentali offerte dal digitale. Il progetto Fonte Gaia[4] si ispira in parte a tale modello. Approcci simili sono offerti dalla rivista digitale "Schoalrly Editing"[5] che pubblica una sezione di edizioni digitali all'interno di un'interfaccia comune e dalle funzionalità più o meno standardizzate. Simile in un certo senso è anche la proposta delle Presses Univeristaires de Caen[6] che offre la possibilità di creare edizioni digitali basate su XML-TEI partendo da file word grazie a un workflow innovativo. Dal canto suo, la TEI è a buon punto con l'elaborazione di un nuovo prodotto chiamato TEI Simple[7] che contiene al suo interno la possibilità di specificare un "processing model" e che quindi può essere utilizzato per la creazione rapida e standardizzata di prodotti digitali 'finiti'. Queste iniziative per quanto molto diverse fra loro hanno tutte in comune l'aspirazione alla standardizzazione dell'editoria scientifica digitale, ma quello che manca ancora è il contributo degli editori stessi che sono oggi interpellati a produrre quel modello scientifico di base che dovrebbe servire da base comune per tali standardizzazioni. Il rischio è infatti che si arrivi alla versione prêt-à-porter delle edizioni digitali basandola sull'esperienza di un numero limitato di studiosi, di discipline e di tradizioni culturali o anche basata su considerazioni pragmatiche e tecniche. Occorre quindi che gli editori e in particolare gli editori digitali assumano un ruolo propositivo all'interno di questo processo di standardizzazione che è diventato ormai urgente. Tale processo non dovrà necessariamente bloccare lo sviluppo di prodotti sperimentali e innovativi, ma anzi potrà (e dovrà) beneficiare della ricerca più innovativa; d'altra parte però tale ricerca beneficerà dall'essere sostenuta da un vivaio di giovani e meno giovani editori, formati alla ricerca digitale e quindi adatti anche a spingersi oltre i limiti, qualora ne avessero la voglia.

Bibliografia

Andrews, T. 2013a. «The Third Way. Philology and Critical Edition in the Digital Age». *Variants* 10:61–76.

[3] https://cudl.lib.cam.ac.uk
[4] http://fontegaia.hypotheses.org/projet-fonte-gaia-2
[5] http://scholarlyediting.org
[6] https://www.unicaen.fr/puc/html/index.html
[7] https://github.com/TEIC/TEI-Simple

Livelli di rappresentazione del testo nell'edizione del *De nomine* di Orso Beneventano

Paolo Monella, Università di Palermo

1 Introduzione

La comunicazione discute l'impianto metodologico e tecnologico dell'edizione critica digitale del *De nomine* di Orso Beneventano (http://www.unipa.it/paolo.monella/ursus/transcription.html), un trattatello grammaticale latino del IX secolo di ambiente beneventano (si vedano Morelli 1910 e Fioretti 2010).

Si tratta di un prototipo che intende mettere alla prova ed approfondire le innovazioni metodologiche proposte da Tito Orlandi, sistematizzate in Orlandi 2010.

2 Metodo

La codifica XML-TEI è strutturata intorno ad elementi <w>, al cui interno il markup relativo alla trascrizione di fonti primarie (in paricolare elementi come <abbr> e <expan>) subisce una specificazione semantica fondata sulla semiotica, in particolare la grafologia, e sui recenti sviluppi della paleografia digitale (Stokes 2011). L'obiettivo è codificare il testo del documento a più livelli testuali, semioticamente fondati. I livelli scelti per questa edizione sono:

1. *Livello grafematico*, le cui unità minime sono i grafemi specifici del sistema grafico del testimone, inclusi i segni paragrafematici e le brachilogie sistematiche. Tutti i caratteri Unicode contenuti in <w> rappresentano grafemi. Tutti i grafemi (glifi aventi valore distintivo) individuati dall'editore sono elencati e commentati nella "Graphemic Table of Signs" (GToS), un file CSV distinto dalla trascrizione XML-TEI ma che costituisce parte integrante dell'edizione. Tramite la GToS l'editore fornisce una sua descrizione del sistema grafematico del manoscritto. La GToS rappresenta un'altra innovazione metodologica, realizzazione dell'idea di "tabella dei segni" teorizzata da Orlandi 2010 e giustificata dal noto principio saussuriano per cui un segno si definisce solo all'interno di un sistema semiotico, in contrasto con ogni altro segno di quel sistema. Se un manoscritto M non ha una distinzione tra "u" e "v" (dunque ha un solo glifo per entrambi, "a forma di u") ed un manoscritto N *ha* tale distinzione (dunque ha due glifi separati), il glifo "u" in M non è lo stesso grafema del glifo "u" in N, anche se "corrisponde" ad esso, né il glifo "v" in M è del tutto un altro grafema, in quanto

anch'esso gli "corrisponde". Le pratiche attuali, ed anche le linee guida XML-TEI, si accontentano di usare lo stesso carattere Unicode "u" (U+0075) nella trascrizione grafematica del glifo "a forma di u" di M e di quello di N. Ciò è sufficiente per la visualizzazione a schermo e l'intuito del lettore, ma per ogni ulteriore elaborazione dei dati da parte del computer ciò implica fallacemente che "u" di M e "u" di N siano lo stesso grafema, e "u" di M e "v" di N siano grafemi diversi. La GToS serve dunque a "de-finire" ogni grafema all'interno del sistema grafico di uno specifico manoscritto. Appare però evidente che la collazione tra manoscritti diversi debba avvenire ad un livello di trascrizione più alto, cioè al livello alfabetico (se si vogliono confrontare le ortografie) o al livello linguistico (se interessano le sole varianti "significative").

2. *Livello alfabetico*, le cui unità minime sono lettere alfabetiche, qui chiamate "alfabemi" e distinte dai grafemi: i grafemi "significano" alfabemi. La GToS riporta, per ogni grafema, il suo significato alfabetico standard (ad es.: il significato standard del grafema codificato col carattere Unicode "u" è l'alfabema codificato col carattere Unicode "u", mentre il significato standard del grafema codificato con "¢" è la sequenza di alfabemi codificati rispettivamente con "q", "u" e "i"). La trascrizione al livello alfabetico non è codificata esplicitamente dall'editore per tutto il testo. Si danno infatti due casi: (A) Se un determinato grafema nel testo ha il suo significato alfabetico standard, indicato nella "Graphemic Table of Signs" (GToS), uno script JavaScript desume l'alfabema (o gli alfabemi) corrispondenti sulla base della trascrizione grafematica XML-TEI (file `casanatensis.xml`) e della GToS (file `GToS.csv`); (B) Se un determinato grafema non ha un significato alfabetico standard desumibile dalla GToS (ad es. l'abbreviazione iniziale in c̄fer per "confer", o il brevigrafo iniziale di "¢a" per "quia"), l'editore, in fase di codifica, marca il passaggio tramite un elemento `<choice>` con all'interno un `<abbr>`, che contiene una sequenza di grafemi, e un `<expan>` che contiene una sequenza di alfabemi, cioè il significato alfabetico di quei grafemi. Nel caso B, dunque, il software trova i due livelli (grafematico e alfabetico) già codificati esplicitamente. In sintesi, la codifica a livello alfabetico è fornita esplicitamente dall'editore all'interno di un elemento `<expan>` solo quando essa non sia desumibile dalla codifica grafematica tramite la tabella di corrispondenze fornita nella GToS.

3. *Livello linguistico*, le cui unità minime sono parole flesse, ma intese sinteticamente, non come sequenza di lettere alfabetiche. La parola flessa (es.: lat. "lupi") viene così identificata nell'edizione tramite la combinazione di un lemma ("lupus, -i") e di informazioni morfologiche (genitivo singolare). Nel codice XML-TEI, si sono usati a questo fine gli attributi dell'elemento `<w>` `@lemma` e `@ana` (per l'analisi morfologica). Nell'esempio di "lupi", dunque: `<w ana="11B---B1--1" lemma="lupus" n="lupi">`. Nella realizzazione pratica dell'edizione, si è partiti da una trascrizione "normalizzata" di ogni parola, su cui si è fatto girare il lemmatizzatore/PoS tagger *TreeTagger* (http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/), col parameter file *Latin-ITTB UD treebank*, scaricato il 15/01/2016 dal sito *Universal Dependencies* (http://universaldependencies.org/) e basato sull'*IT-TB –Index Thomisticus Treebank* (http://itreebank.marginalia.it/). Il risultato della lemmatizzazione è stato poi verificato e corretto dall'editore. Attraverso la distinzione, in fase di codifica, del "livello linguistico" dai livelli "grafematico" ed "alfabetico", la lin-

guistica computazionale torna ad essere, come è sempre stata la lingustica in ambito tradizionale, supporto fondamentale dell'attività ecdotica.

3 Discussione

Quale di queste trascrizioni è dunque il "testo da leggere"? Quale corrisponde all'edizione critica tradizionale?

Uno dei vantaggi dell'edizione critica digitale è, in realtà, proprio quello di integrare in un sistema unico l'edizione cosiddetta "diplomatica" e quella "critica" o interpretativa, superando tale dicotomia (Vanhoutte 2000, Bodard e Garcés 2009, Orlandi 2010, Pierazzo 2011, Brüning, Henzel e Pravida 2013, Pierazzo 2015a, *Vercelli Book Digitale* http://vbd.humnet.unipi.it/, *Menota –Medieval Nordic Text Archive* http://menota.org/).

Nel modello di edizione qui proposto, se l'edizione "diplomatica" è facilmente identificabile con il livello grafematico, le funzioni svolte dall'edizione "critica" cartacea vengono distribuite a più livelli.

Nel caso di una tradizione testuale unitestimoniale come quella del *De nomine* di Orso Beneventano, la funzione-lettura viene diffratta tra il livello alfabetico e quello linguistico.

Il lettore leggerà il livello alfabetico se vuole ignorare le abbreviazioni e la punteggiatura antica (livello grafematico), ma è interessato all'ortografia originale del testo, senza alcuna "normalizzazione" delle monottongazioni o dell'ortografia antica, né correzione dei *lapsus calami*.

Leggerà invece il livello linguistico se è interessato al "testo" astratto: vi troverà l'ortografia e la punteggiatura "normalizzate", moderne, e i *lapsus calami* corretti. A questo livello, infatti, viene trascritta *la parola latina* che, secondo l'editore, lo scriba *intendeva scrivere*.

Se invece, a parere dell'editore, lo scriba intendeva effettivamente scrivere una determinata parola flessa (ad es. l'accusativo di un sostantivo), ma l'editore vuole *emendare* il testo tràdito (ad es. in un dativo), questo tipo di intervento apparterrà anch'esso concettualmente al livello linguistico, ma andrà indicato tramite markup specifico. Per questa via, di fatto, si aggiunge un ulteriore livello di codifica del testo: il testo emendato dal filologo.

Un ulteriore livello di complessità sarà aggiunto, in futuri esperimenti di questo modello di edizione, nelle edizioni di testi con più testimoni.

In questi casi, ogni manoscritto avrà una sua edizione sui tre livelli citati, e la collazione in vista della definizione di un "testo dell'editore" confronterà le parole flesse al livello linguistico, evidenzando così solo le varianti dette "significative", ignorando invece le varianti "grafiche" (paleografiche, appartenenti al livello grafematico, o ortografiche, appartenenti al livello alfabetico).

Tale distinzione di livelli permette dunque di codificare univocamente la parola flessa al livello linguistico al di là del sistema grafico e persino dell'alfabeto utilizzati nei singoli testimoni, offrendo così un livello di codifica utile per la visualizzazione "normalizzata" (la tradizionale edizione "critica" o interpretativa), per la ricerca e l'analisi testuale, per l'interoperabilità dell'edizione al di là delle specificità paleografiche del manoscritto e, potenzialmente, per la generazione di apparati critici dinamici in tradizioni pluritestimoniali.

4 Sviluppo, licenze e riuso

L'edizione è *open source*: tutto il codice sorgente (markup e software) è stato scritto con l'editor Vim (http://www.vim.org/) ed è disponibile sul repository https://github.com/paolomonella/ursus sotto la GNU General Public License, insieme ad una ampia documentazione in inglese.

La scelta dei linguaggi utilizzati, e in particolare di XML-TEI, è mirata a consentire il riuso del codice sorgente. Allo stesso fine mira la documentazione dettagliata di ogni aspetto, metodologico e tecnologico, del sistema-edizione, e in particolare della semantica degli elementi XML-TEI utilizzati.

L'impianto teorico dell'edizione prevede che le riproduzioni digitali del *codex unicus* che contiene il testo (Casanatensis 1086) siano collegate alla trascrizione. Tuttavia, un accordo con la Biblioteca Casanatense di Roma, che ha fornito le riproduzioni, attualmente vincola l'editore a non rendere disponibili le immagini online.

Bibliografia

Bodard, G., e J. Garcés. 2009. «Open Source Critical Editions: A Rationale». In *Text Editing, Print, and the Digital World*, a cura di M. Deegan e K. Sutherland, 83–98. Digital Research in the Arts and Humanities. Aldershot: Ashgate.

Brüning, G., K. Henzel e D. Pravida. 2013. «Multiple encoding in genetic editions: the case of "Faust"». *Journal of the Text Encoding Initiative*, n. 4. doi:10.4000/jtei.697.

Fioretti, P. 2010. «L'eredità di un maestro. Genesi ed edizione della grammatica di Orso beneventano». In *Libri di scuola e pratiche didattiche. Dall'antichità al Rinascimento, Atti del Convegno Internazionale organizzato dall'Università degli Studi di Cassino (Cassino 7-10 maggio 2008)*, a cura di L. Del Corso e O. Pecere, 293–328. Cassino: Edizioni Università di Cassino.

Morelli, C. 1910. «I trattati di grammatica e retorica del cod. Casanatense 1086. Nota del dott. C. Morelli, presentata dal Socio G. Vitelli». *Rendiconti della Reale Accademia dei Lincei, Classe di scienze morali, storiche e filologiche* XIX:287–328.

Orlandi, T. 2010. *Informatica testuale. Teoria e prassi*. Roma: Laterza.

Pierazzo, E. 2011. «A rationale of digital documentary editions». *Literary and Linguistic Computing* 26 (4): 463–477. doi:10.1093/llc/fqr033.

Stokes, P. A. 2011. «Describing Handwriting, Part IV: Recapitulation and Formal Model». *DigiPal Blog* October 14. http://www.digipal.eu/blogs/blog/describing-handwriting-part-iv/.

Vanhoutte, E. 2000. «Where is the editor? Resistance in the Creation of an electronic critical edition». In *DRH 98. Selected papers from Digital resources for the Humanities 1998*, a cura di M. Deegan, J. Anderson e H. Short, 171–183. London: Office for Humanities Communication. http://etjanst.hb.se/bhs/ith//1-99/ev.htm.

Edizioni Digitali
Digital Editions
Talks

The Virtual Manuscript Room Collaborative Research Environment

Troy A. Griffitts, Akademie der Wissenschaften zu Göttingen, troy@crosswire.org
Ulrich B. Schmid, Akademie der Wissenschaften zu Göttingen, uschmid1@gwdg.de

1 Introduction

The Virtual Manuscript Room Collaborative Research Environment (VMR CRE)[1] brings community and a toolbox of powerful research components to support all stages of research and production of a digital edition. Beginning with the popular open-source portal, Liferay, the VMR CRE integrates 30 DH components to naturally support: Cataloging witnesses; Managing and displaying images; producing well-formed TEI transcriptions using a web-based WYSIWYM editor and storing those transcriptions to a versioned transcription repository; community volunteer task assignment management; automatic realtime collation of witnesses; regularization and apparatus editing; online publishing of the final results– as a traditional apparatus, or with interactive tools which let users choose different ways to visualize the data produced in the edition. Originally developed to facilitate the globally disparate teams editing the approximately 5700 Greek New Testament manuscript witnesses for the Editio Critica Maior[2], the VMR CRE has been adopted by a variety of other projects, including The Digital Edition and Translation of the Coptic-Sahidic Old Testament[3]The Canons of Apa Johannes the Archimandrite[4], and The Avestan Digital Archive[5]. Each of these adaptations has presented unique challenges, with the result that the software, which was originally highly focused on New Testament manuscript research, has evolved into a useful tool for general digital edition research and electronic publishing.

The VMR CRE is an ecosystem of components, enabling projects to choose only the tools pertinent to their research. Many teams have contributed to the ecosystem, including the European Union's COST initiative, Interedition[6], providing the CollateX collation engine[7] and initial work on variant graph display; University of Trier, providing a web-based WYSIWYM TEI editor for manuscript transcription[8]; and the Institute for Textual Scholarship and Electronic Editing (ITSEE) at the University of Birmingham, contributing the Apparatus Editor[9].

[1] http://vmrcre.org
[2] http://ntvmr.uni-muenster.de
[3] http://coptot.manuscriptroom.com
[4] coptot.manuscriptroom.com/web/apa-johannes
[5] http://www.avesta-archive.com
[6] http://www.interedition.eu
[7] http://collatex.net
[8] http://wfce-ote.sourceforge.net
[9] http://www.birmingham.ac.uk/research/activity/itsee

A walk through the workflow at the Institut für neutestamentliche Textforschung (INTF), in their efforts to edit the Editio Critica Maior (ECM[10]), provides opportunity to touch on many components available in the VMR CRE. Work can be dividing in 9 discrete stages, progressively: 1. witness cataloging; 2. witness selection; 3. image management; 4. indexing of folio content; 5. transcribing; 6. regularizing; 7. collating; 8. editing an apparatus; 9. genealogical analysis of the witness corpus.

2 Metadata and Feature Tagging

The VMR CRE stores with each manuscript a very limited set of descriptive data and reserves the primary metadata capture for a dynamic tagging facility called Feature Tagging. A Feature is any defined metadata information which might be captured for a manuscript or manuscript page. For example, an alternative catalog identifier, an external image repository, the canvas material, the ink type, the script type; these are all Features which might be captured for a manuscript; For individual pages, an illumination, a canon table, or even individual sample script characters might be tagged at Features. These Features must first be defined in the system, and the VMR CRE comes by default with a predefined set of Feature Definitions used at the INTF. A Feature Definition can specify that zero or more values should be captured with the Feature tag and what those values and value types should be. Once a Feature is defined, it can be used to tag manuscripts or manuscript pages, capturing individual Feature values for each tag, if necessary.

Every Feature Definition adds to the number of facets available in the catalog search facility. For example, one might search for all manuscript folio sides from Egypt which include Illuminations and any part of the Gospel of John. A Feature Tag can also include a region box, marking the area on a folio image where the Feature is present. If a region box is captured, a search query can specify to show the region box clips in the result. For example, one might choose to capture a set of representative letters for each manuscript and then perform a search for all double column manuscripts with a height of at least 20cm between the II and V centuries, and to ask the query results to show the representative α (alpha) clips.

3 Transcription and Reconciliation

Transcription work in the VMR CRE is done using a What You See Is What You Mean (WYSIWYM) web-based editor originally developed by the University of Trier in collaboration with the INTF and ITSEE in Birmingham. This transcription editor has been developed as a plugin for the popular TinyMCE[11] HTML editor component. The editor includes menus and dialogs to assist the researcher with composing a transcription, without asking the transcriber to learn special markup codes. The content may then be obtained as EpiDoc[12] influenced TEI.

The VMR CRE saves content in a versioned transcription repository backed by Git[13].

[10] http://egora.uni-muenster.de/intf/projekte/ecm_en.shtml
[11] https://www.tinymce.com
[12] http://www.stoa.org/epidoc/gl/latest
[13] https://git-scm.com

A user may have access to create and edit their own personal transcription, a project-wide transcription, or a site-wide (= published) transcription– each having version history.

The VMR CRE also includes a palaeography tool to assist a transcriber when encountering rare symbols, abbreviations, or ligatures. If a portion of the unknown text can be identified, the researcher can enter one or more letters and will be presented with images of text instances elsewhere which including these letters, offering possibilities. As more and more rare text items are tagged, the system grows more helpful.

Quality assurance for the ECM requires that a transcription for a manuscript be produced independently by two transcribers. The products are then compared to each other and differences are reviewed by a manager and reconciled to produce a final transcription. The VMR CRE provides tools to facilitate this reconciliation work.

4 Collation and Visualization

Collation is a key component to find differences in text witnesses when producing a critical edition. Collation facilities in the VMR CRE are performed by CollateX. Collation and regularization of uninteresting differences is an iterative cycle in digital editing and the VMR CRE ties these two actions together with an intuitive visual interface. Visualization of a collation, either during the editing process or for the reader, can be rendered as a variant graph, an alignment table, or as a traditional negative apparatus.

5 Web Services, Open Programmatic Access

The VMR Web Services API layer is primarily useful for exposing the functionality of the VMR to other research projects wishing to access the functionality or contribute to the dataset through their own systems and tools. The VMR CRE Web Services API generally uses noun/verb nomenclature organized by category. This means that the last 2 segments of an API URL will consist first of the type of object the call will affect, and second of the action to be performed on the object. Any path before the final two segments are merely for organizational purposes. This is different from a strict REST convention which confines the action to one of 6 HTTP verbs. The VMR CRE places no semantic meaning on the HTTP verb. Both GET and POST HTTP verbs are accepted as identical, relegating the verb, or semantic action to the final segment of the URL. This allows easy testing and examples for every action directly within a web browser. Parameters to a service request are passed as standard HTTP FORM POST parameters or as query string parameters. As an example, an API request to obtain a transcription of manuscript 30093 of the Apocalypse by Darius Mueller in TEI form with no stylesheet applied would be:

http://ntvmr.uni-muenster.de/community/vmr/api/transcript/get/?docID=30093&indexContent=Rev&fullPage=true&userName=darius.mueller&format=teiraw

Notice the final noun/verb segments of the URL: transcript/get

References

Calzolari, V., and M. E. Stone. 2014. *Armenian Philology in the Modern Era: From Manuscript to Digital Text*. BRILL.

Finneran, R. J. 1996. *The Literary Text in the Digital Age*. Ann Arbor, MI: University of Michigan Press.

Russell, J., and R. Cohn. 2012. *Epidoc*. Book on Demand.

Mapping Castiglione's Letters

Roberto Vetrugno, Nicolaus Copernicus University

Le 1779 lettere di Baldassar Castiglione, ora finalmente edite, offrono una quantità rilevante di dati ecdotici e linguistici che possono essere visualizzati e interrogati in modo non tradizionale, sviluppando in chiave digitale le varie indicizzazioni che un'edizione cartacea prevede.

Una prima elaborazione di questi dati è di natura semantica e riguarda i contenuti e il lessico: la corrispondenza del diplomatico mantovano si presenta a noi come un'affascinante narrazione della vita quotidiana di un gentiluomo del Rinascimento: gli interessi estetici per l'abbigliamento, l'oggettistica, la salute e la medicina, l'alimentazione, l'amministrazione, le monete, l'economia e l'amministrazione delle sue proprietà e dei domini gonzagheschi, il mestiere delle armi etc.

L'accesso di un lettore a un corpus epistolare è abitualmente dettato da un ordine cronologico e dagli indici dei nomi, dei luoghi e, talvolta, delle forme notevoli. Un nuovo modello di consultazione, ideato per le lettere di Castiglione e applicabile ad altri epistolari moderni, permette di leggere le lettere in maniera "trasversale" e tematica partendo dal lessico. Vediamo prima un campione di glossario tradizionalmente ordinato alfabeticamente:

camisa (camiscia) 'camicia', 146, 4; 174, 6; 196, 3; 203, 1; 209, 1; 225, 2; 226, 1; 343, 11; 619, 5; 716, 8; 728, 3; 740, 1; 1401, 11;

Cort III 32

-- di cortina lavorate d'oro, 1547, 7

-- di taffetà, 281, 3; 282, 3°; 284, 1

camisoletta, 36, 9.

Lo studioso può consultarlo cercando una parola prestabilita oppure sfogliando faticosamente l'intero glossario alfabetico per individuare le parole che lo interessano, poniamo ad esempio quelle relative alla medicina o al lessico delle armi. L'obiettivo è quindi trasformare un tradizionale indice delle forme notevoli (o glossario) in un nuovo sistema di consultazione che parta da un ordine tematico delle parole.

La struttura proposta è una sorta di mappatura lessicale della vita di Castiglione che possiamo "vedere" e consultare partendo dalle diverse categorie tematiche viste sopra (abbigliamento, oggettistica, medicina, alimentazione etc.). Un prototipo di visualizzazioni tematiche e lessicali per insiemi (o sfere) è rappresentato in Fig. 1 e Fig. 2.

Questo sistema è interrogabile attraverso l'accesso ai singoli insiemi: selezionando una delle sfere, compaiono le parole e cliccando su una singola forma si accederà alla lista delle occorrenze (cronologicamente distinte) e da qui, se possibile, al contesto (oppure si rimanderà all'edizione cartacea). Lo studioso, ad esempio di storia delle armi, potrà così "entrare" nella sfera intitolata "mestiere delle armi" e vedrà nell'insieme tutte le parole che afferiscono a questo ambito ottenendo così un quadro complessivo dei lemmi di suo interesse presenti nell'epistolario. Per Castiglione si potrà anche rimandare all'occorrenza della parola nel

Questo sistema è interrogabile attraverso l'accesso ai singoli insiemi: selezionando una delle sfere, compaiono le parole e cliccando su una singola forma si accederà alla lista delle occorrenze (cronologicamente distinte) e da ogni se possibile, al contesto (oppure si rimanderà all'edizione cartacea). Lo studioso, ad esempio di storia delle armi, potrà così "entrare" nella sfera intitolata "mestiere delle armi" e vedrà nell'insieme tutte le parole che afferiscono a questo ambito ottenendo così un quadro complessivo dei lemmi di suo interesse presenti nell'epistolario. Per Castiglione si potrà anche rimandare all'occorrenza della parola nel *Cortegiano*, così da visualizzare la stessa parola usata dal cortigiano reale nelle lettere e dal cortigiano ideale nel Dialogo:

Cortegiano, così da visualizzare la stessa parola usata dal cortigiano reale nelle lettere e dal cortigiano ideale nel Dialogo:

284. A Cristoforo Tirabosco
(Roma, 17 luglio 1514)

1 Christoforo. Ho recevuto alchune lettere vostre, et insieme ho hauto la **camisa** de tafetà. Non occorre altro, se non che quando vi occorre, scriviati a M.a mia matre, che ce mandi denari, perché gli è forza. [...]

284. A Cristoforo Tirabosco
(Roma, 17 luglio 1514)
1 Christoforo. Ho recevuto alchune lettere vostre, et insieme ho hauto la **camisa** de tafetà. Non occorre altro, se non che quando vi occorre, scriviati a M.a mia matre, che ce mandi denari, perché gli è forza. [...]

Cortegiano, III, 32

Chii; li quali non potendo contrastare, tolsero patto col giuppon solo e la **camiscia** uscir della città. Intendendo le donne così vituperoso accordo, si dolsero, rimproverandogli che, lassando l'arme, uscissero come ignudi tra' nemici.

Cortegiano, III, 32
Chii; li quali non potendo contrastare, tolsero patto col giuppon solo e la **camiscia** uscir della città. Intendendo le donne così vituperoso accordo, si dolsero, rimproverandogli che, lassando l'arme, uscissero come ignudi tra' nemici.

Il secondo sistema di interrogazione digitale riguarda la descrizione e l'ordinamento dei testimoni di lettere, manoscritte e a stampa, attraverso la messa in evidenza della loro natura testuale: si possono quantitativamente visualizzare le diverse tipologie di testimoni con cui vengono trasmessi i testi epistolari consentendo di acquisire informazioni complessive circa il carattere testuale delle lettere e la distribuzione negli anni; illustrare così la datazione, la quantità e le differenti tipologie di lettere: missive originali autografe, missive originali non autografe, minute autografe, copia su registri o copialettere, copia tarda (realizzata non simultaneamente all'originale). Tab. 1 mostra i tipi e alle quantità di lettere cronologicamente ordinate (si tratta delle lettere di Castiglione di due periodi: 1519-1524, quando è am).

Il secondo sistema di interrogazione digitale riguarda la descrizione e l'ordinamento dei testimoni di lettere, manoscritte e a stampa, attraverso la messa in evidenza della loro natura testuale: si possono quantitativamente visualizzare le diverse tipologie di testimoni con cui vengono trasmessi i testi epistolari consentendo di acquisire informazioni complessive circa il carattere testuale delle lettere e la distribuzione negli anni; illustrare così la datazione, la quantità e le differenti tipologie di lettere: missive originali autografe, missive originali non autografe, minute autografe, copia su registri o copialettere, copia tarda (realizzata non simultaneamente all'originale). Di seguito una tabella relativa ai tipi e alle quantità di lettere cronologicamente ordinate (si tratta delle lettere di Castiglione di due periodi: 1519-1524, quando è ambasciatore gonzaghesca a Roma; 1525-1529, quando è nunzio apostolico in Spagna):

anno	quantità	missiva autografa	missiva non autografa	minuta autografa	copia (cancelleresca, registri e "copialettere")	copia tarda	edizione (testimone unico a stampa)
1519	95	34	59			1	1
1520	72	57	13		1		1
1521	387		114		74	7	
1522	419	147	98	1	170	2	1
1523	115	53	18	1	41		
1524	215	115	86	3	10	1	
1525	85	13	7	22	36	3	4
1526	51	4	6	13			
1527	24	10	1	5		3	
1528	10	7	1			2	
1529	2	2					

Tabella 1: Tipi e quantità di lettere

La tabella quantifica i manoscritti secondo campi distinti per tipologia di testimone: per interrogarla possiamo selezionare una cella di nostro interesse per accedere ai testimoni distinti per tipi; quindi ad esempio interrogando la cella rossa relativa alle minute autografe del 1524, compariranno i dati relativi ai tre testimoni (destinatario, luogo di spedizione e di destinazione, sede di conservazione e collocazione, se possibile trascrizioni immagini dei mss.). In Fig. 3 e Fig. 4 si vedono i grafici relativi ai dati della tabella; si noti come per il periodo 1525-1529 il quadro dei testimoni delle lettere muta radicalmente a causa della perdita degli originali inviati dalla Spagna e per la criticità delle comunicazioni internazionali negli anni contigui al sacco di Roma:

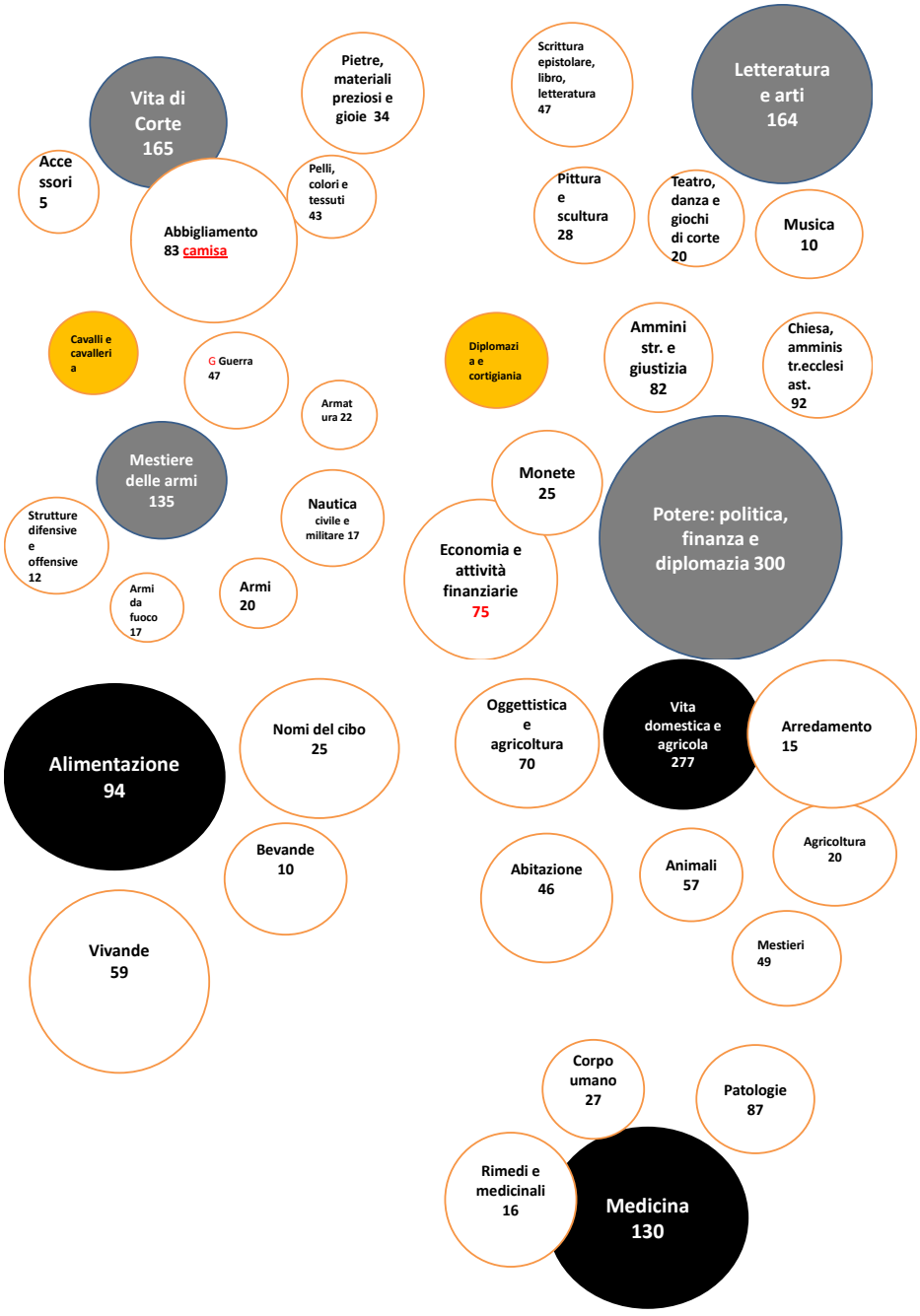


Figura 1: Visualizzazioni tematiche e lessicali per insiemi (a)

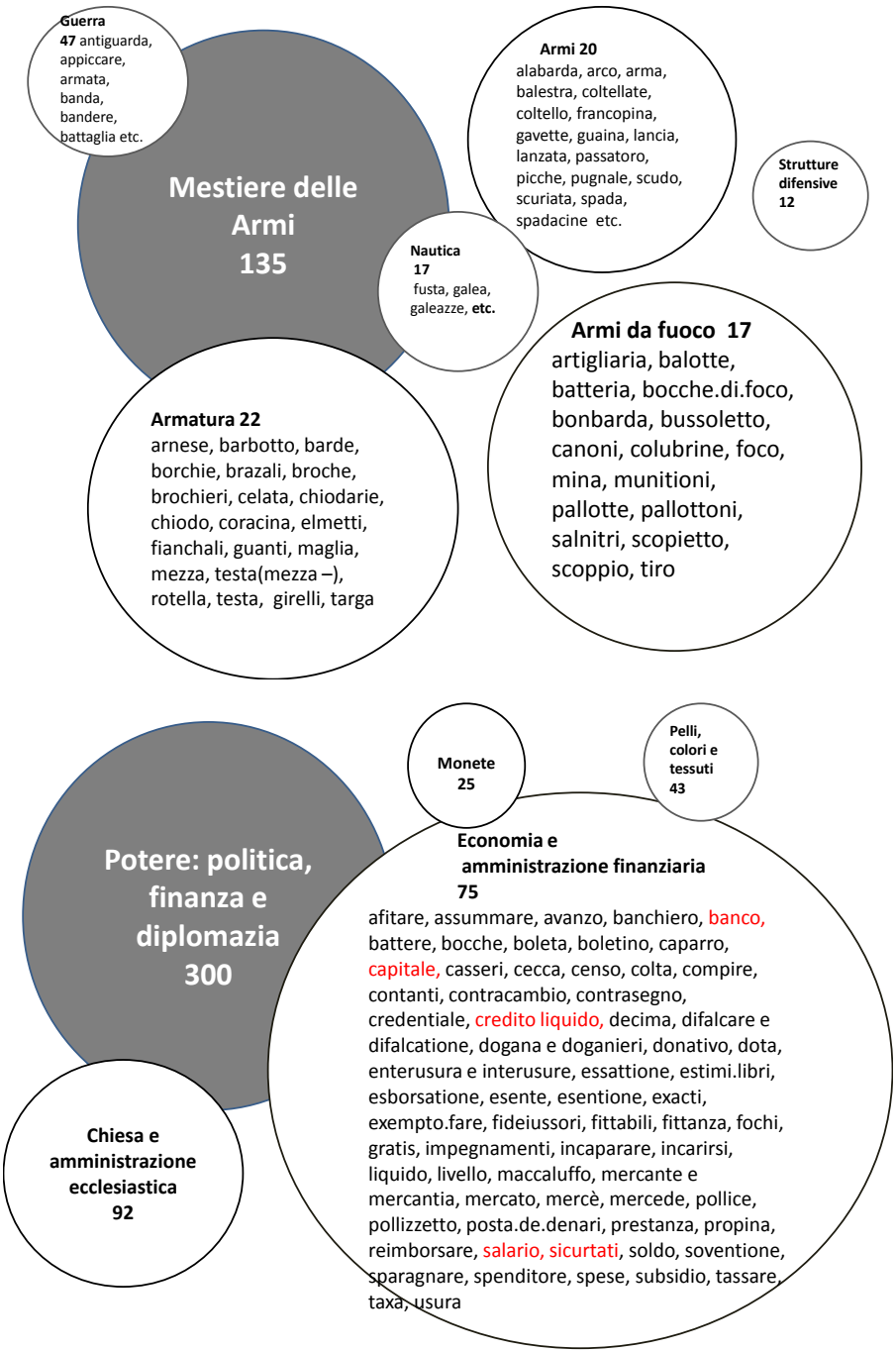


Figura 2: Visualizzazioni tematiche e lessicali per insiemi (b)

la cella rossa relativa alle minute autografe del 1524, compariranno i dati relativi ai tre testimoni (destinatario, luogo di spedizione e di destinazione, sede di conservazione e collocazione, se possibile trascrizioni immagini dei mss.).
Di seguito un grafico relativo ai dati della tabella; si noti come per il periodo 1525-1529 il quadro dei testimoni delle lettere muta radicalmente a causa della perdita degli originali inviati dalla Spagna e per la criticità delle comunicazioni internazionali negli anni contigui al sacco di Roma:

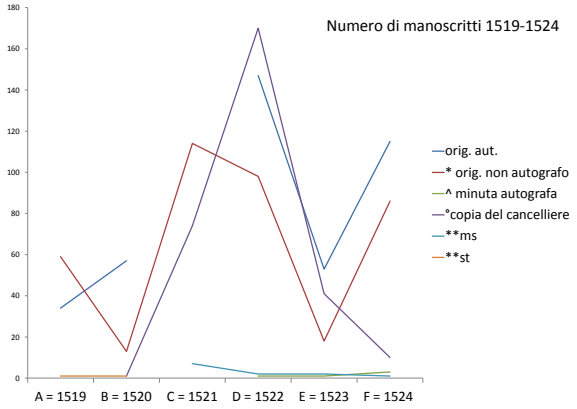


Figura 3: Variazioni nel numero di manoscritti (1519-1524)

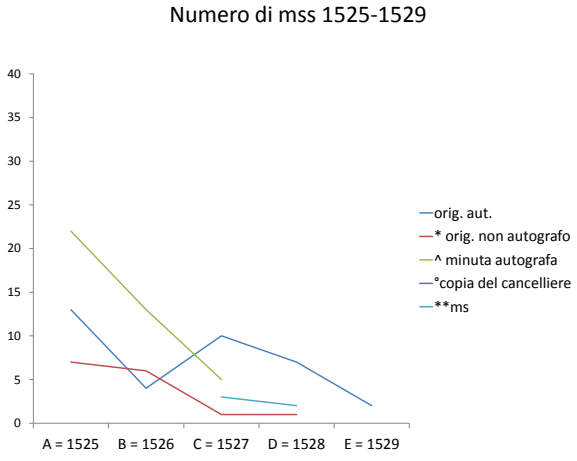


Figura 4: Variazioni nel numero di manoscritti (1519-1524)

Bibliografia e sitografia

B. Castiglione, *Lettere famigliari e diplomatiche*, a cura di A. Stella e U. Morando, Nota ai testi, indici e apparati a cura di R. Vetrugno, Torino, Einaudi, 2016.
R. Vetrugno, *La lingua di Baldassar Castiglione epistolografo*, Novara, Interlinea, 2010.
F. Moretti, *Graphs, Maps, Trees: Abstract Models for a Literary History*, Verso Books, 2007.

http://republicofletters.stanford.edu/
http://republicofletters.stanford.edu/publications.html
http://palladio.designhumanities.org/#/
http://hfroehli.ch/2014/05/11/intro-bibliography-corpus-linguistics/
http://stanfordnlp.github.io/CoreNLP/
http://programminghistorian.org/lessons/corpus-analysis-with-antconc

Bibliografia

Moretti, F. 2007a. *Graphs, Maps, Trees: Abstract Models for a Literary History*. Verso Books.

Stella, A., U. Morando e R. Vetrugno, cur. 2016. *Baldassar Castiglione, Lettere famigliari e diplomatiche*. Torino: Einaudi.

Vetrugno, R. 2010. *La lingua di Baldassar Castiglione epistolografo*. Novara: Interlinea.

Sitografia

http://republicofletters.stanford.edu
http://republicofletters.stanford.edu/publications.html
http://palladio.designhumanities.org/#
http://hfroehli.ch/2014/05/11/intro-bibliography-corpus-linguistics
http://stanfordnlp.github.io/CoreNLP
http://programminghistorian.org/lessons/corpus-analysis-with-antconc

Corrispondenze diplomatiche francesi del Seicento: le possibilità offerte dall'edizione digitale

Cecilia Russo, Università degli Studi di Torino

1 Introduzione

Con questo intervento intendo presentare alcune problematiche incontrate approntando l'edizione di una corrispondenza diplomatica francese del Seicento e spiegare come esse possano essere risolte, almeno in parte, grazie ad un'edizione digitale interattiva.

Si tratta delle lettere che Benoît Cise de Grésy, diplomatico al servizio dei Duchi di Savoia, indirizzò alla corte di Torino tra il 1621 e il 1684. Questi dispacci, conservati all'Archivio di Stato di Torino, hanno un indiscutibile valore storico ma sono preziosi anche da un punto di vista linguistico, poiché appartengono ad un genere testuale poco studiato e costituiscono una importante testimonianza di francese (pre)classico, in piena evoluzione a vari livelli (ortografico, morfosintattico...). La loro analisi può dunque contribuire a comprendere meglio il processo di elaborazione e fissazione delle norme che caratterizza il francese seicentesco.

Lo scopo di questa edizione quindi è duplice: da un lato rendere accessibile il testo ai lettori contemporanei modernizzandolo (e accompagnandolo da un imprescindibile apparato di note storiche) e dall'altro renderlo fruibile per analisi linguistiche conservando le peculiarità della lingua di Cise che è estremamente instabile (grafie variabili, scelte grammaticali e lessicali inconsuete...).

Come raggiungere questo obiettivo?

2 Stato dell'arte

Molte corrispondenze diplomatiche francesi del Seicento (*Correspondance du cardinal Mazarin*, E. T. Hamy, Monaco, 1904; *Correspondance du chevalier de Sévigné et de Christine de France, duchesse de Savoie*, J. Lemoine e F. Saulnier, Paris, 1911; e *Correspondance administrative sous le règne de Louis XIV*, G. B. Depping, Paris, 1850-1855) sono state realizzate da storici, con intenti prevalentemente contenutistici; prive di indicazioni filologiche precise sono inutilizzabili per indagini linguistiche.

Neppure le edizioni critiche più recenti (come ad esempio *La correspondance d'Albert Bailly*, 10 vol., Aosta, 1999-2010) che seguono i criteri e i mezzi della filologia tradizionale (formato cartaceo, una leggera modernizzazione della lingua, con interventi sempre segnalati, ricca annotazione storico-filologica), consentono agevoli interrogazioni linguistiche.

Negli ultimi anni sono state realizzate alcune edizioni digitali di corrispondenze cinqueseicentesche come la *Correspondance d'Antoine Du Bourg* (O. Poncet, 2011)[1], che mira a costruire un modello di interfaccia per tutte le corrispondenze non letterarie dell'epoca moderna, o la *Correspondance française de Guy Patin* (L. Capron, 2015)[2] la quale, pur avendo molti apparati di tipo contenutistico, non è interattiva.

Mancano, dunque esempi soddisfacenti di edizioni digitali interattive di corrispondenze francesi.

3 La corrispondenza di Cise e i vantaggi offerti dall'edizione digitale

Alcune caratteristiche specifiche delle lettere di Cise rendono difficilmente praticabile la scelta di un'edizione critica tradizionale su supporto cartaceo e mi portano a privilegiare l'edizione digitale interattiva:

- si tratta di un corpus vasto (oltre duemila lettere);
- si riscontra la presenza di alcune lettere cifrate, nelle quali l'autore si serve di simboli grafici che devono essere mostrati per rendere conto al lettore dei codici usati;
- la stessa disposizione spaziale della lettera trasmette indicazioni circa alcune convenzioni sociolinguistiche dell'epoca (spesso rivelando informazioni sul rapporto che intercorre tra mittente e destinatario) ed è bene che sia visualizzabile.

L'impiego di strumenti informatici si è quindi imposto per fornire la possibilità di fare ricerche specifiche all'interno del testo, facendo in modo che il sistema evidenzi, a seconda delle esigenze del lettore, aspetti di tipo contenutistico e/o aspetti di tipo filologico.

Inoltre, l'edizione digitale interattiva permette di superare il dilemma della scelta tra edizione diplomatica o interpretativa, con la visualizzazione di vari 'stadi' del testo (edizione diplomatica, interpretativa e versione modernizzata) e l'immagine del manoscritto. Infine il testo, taggato linguisticamente attraverso software o lemmatizzatori quali ad esempio TreeTagger, TXM o LGeRM potrà confluire in corpora linguistici digitali e interattivi.

Ho quindi iniziato, parallelamente ad un'edizione tradizionale, ad approntare un'edizione interattiva avvalendomi del software EVT, utilizzato con il visualizzatore Oxygen (a cui mi sono avvicinata grazie al seminario di filologia digitale AIUCD 2015 (tenutosi all'università di Torino nel novembre 2015).

Per poter effettuare analisi di tipo filologico ho cominciato ad annotare alcune lettere attraverso il programma TreeTagger, il quale, se utilizzato con testi-input dalla grammatica molto instabile, come le lettere di Cisa, presenta attualmente alcune criticità.

Nel corso della presentazione si evidenzierà quanto sia auspicabile l'utilizzo di un software che permetta al lettore di visualizzare, analizzare e confrontare diversi stadi del testo per studiare meglio alcune peculiarità linguistiche del documento. Verrà, inoltre, mostrata una prima demo di alcune lettere realizzate con software EVT.

[1] Disponibile online: http://elec.enc.sorbonne.fr/dubourg
[2] Disponibile online: http://www.biusante.parisdescartes.fr/patin

Una delle difficoltà più importanti con la quale ho dovuto confrontarmi, volendo realizzare un'edizione digitale, è stata la mancanza di conoscenze informatiche necessarie per utilizzare, modificare e sfruttare software o lemmatizzatori, per questo attualmente risulta ancora indispensabile una collaborazione tra il filologo tradizionale e l'informatico. Il primo si dedica all'interpretazione, alla trascrizione e all'annotazione del testo, tenendo conto delle esigenze del mezzo digitale; il secondo invece si occupa solitamente di fornire competenze digitali, codificando e implementando le informazioni nel processo di realizzazione dell'interfaccia. Si auspica, però, che in futuro si diffonda sempre più la figura dell'umanista digitale: un filologo che ha competenze digitali acquisite nel suo percorso di studi, che potrà realizzare autonomamente un'edizione di tipo digitale-interattiva.

4 Questioni aperte

Dopo aver illustrato i vantaggi e i risultati raggiungibili grazie all'edizione digitale interattiva, evocherò rapidamente due problematiche della filologia digitale ad oggi irrisolte, particolarmente delicate nel caso di un'edizione come la nostra, di tipo accademico (tesi di dottorato) che deve poter essere fruita dalla comunità scientifica e auspicabilmente diventare un modello di riferimento.

Così come il testo dell'edizione è in continua evoluzione, anche il mezzo digitale è soggetto a incessanti cambiamenti e, quindi, si dovrà provvedere ad associare ad ogni edizione un riferimento di identificazione permanente e univoco (per esempio DOI), per evitare la scomparsa dell'edizione in caso di cambiamenti di URL o chiusura del dominio che le ospita, o per scongiurare variazioni indesiderate in caso di modifiche al layout della pagina web che contiene l'edizione. Inoltre, qualora l'edizione sia destinata ad essere pubblicata, anche solo sul web, occorrerà identificarla con un codice ISBN, che pur non essendo obbligatorio, permetterà di agevolare la proprietà intellettuale oltre che le transazioni commerciali. Inoltre, gli eventuali diritti di utilizzazione economica dell'opera saranno subordinati all'iscrizione, da parte degli autori, alla SIAE.

Se poi l'editore modifica il software EVT egli dovrà tener presente che questo è soggetto a licenza di GNU General Public License version 2.0 (GPLv2), sarà quindi necessario ridistribuire sul web il software con le variazioni approntate.

Bibliografia

Ambrosio, A., S. Barret e G. Vogeler. 2014. *Digital diplomatics The computer as a tool for diplomatist?* Köln, Weimar, Wien: Böhlau Verlag.

Apollon, D., e C. Belisle. 2014. *Digital Critical Editions*. Adjust Address. Chicago, Springfield: University of Illinois Press.

Ciula, A., e F. Stella. 2007. *Digital Philology and Medieval Texts*. Pisa: Pacini Editore.

Driscoll, M. J., e E. Pierazzo. 2016. *Digital Scholarly Editing: Theories and Practices*. Cambridge, UK: Open Book Publishers. http://www.openbookpublishers.com/product/483/digital-scholarly-editing--theories-and-practices.

Fiormonte, D., T. Numerico e F. Tomasi. 2015. *The Digital Humanist: A Critical Inquiry*. New York: Paperback.

Guillot, C., et al. 2013. «La "philologie numérique" : tentative de définition d'un nouvel objet éditorial». In *Actes du XXVIIe Congrès international de linguistique et de philologie romanes (Nancy, 15-20 juillet 2013)*, a cura di R. Trachsler, F. Duval e L. Leonardi. https://halshs.archives-ouvertes.fr/halshs-00846767/document.

Robinson, P. 2013c. «Towards a theory of digital editions». *The Journal of the European Society for Textual Scholarship*:105–131. https://www.academia.edu/3233227/Towards%5C_a%5C_Theory%5C_of%5C_Digital%5C_Editions.

Rosselli Del Turco, R., et al. 2014a. «Edition Visualization Technology: A Simple Tool to Visualize TEI-Based Digital Editions». *Journal of the Text Encoding Initiative* 8. doi:10.4000/jtei.1077.

Shillingsburg, P. 2006a. *From Gutenberg to Google: electronic representations of literary texts*. Cambridge, UK: Cambridge University Press.

Trotter, D. 2015. *Manuel de la philologie de l'édition*. Berlin: De Gruyter.

La collazione semi-automatica dei testi medioevali tra linguistica e algoritmi

Elena Spadini

L'intervento affronta due aspetti della collazione semi-automatica applicata ai testi medioevali: 1. la possibilità di collazionare più testimoni senza utilizzare un testo di base e 2. la necessità di normalizzare i testi, data la frequente instabilità grafica, fonetica e morfologica[1].

A partire da questi due punti, si vuole riflettere sulla necessità di guardare dietro allo schermo e dentro al procedimento computazionale, per utilizzare consapevolmente gli strumenti per la collazione, individuarne caratteristiche e funzionalità rilevanti, e sfruttare il loro potenziale innovativo. L'innovazione è definita utilizzando l'approccio di Orlandi al campo dell'edizione digitale: la superiorità dell'edizione elettronica risiede nella possibilità di utilizzo del testo in sede elettronica stessa, più che nella presentazione al lettore come finalità ultima. Applicando tale approccio al nostro caso, vedremo come il potenziale innovativo della collazione semi-automatica (e in particolare dei due punti qui discussi) cresce esponenzialmente quando i risultati di tale operazione sono utilizzati in ulteriori procedimenti computazionali.

1. L'utilizzo di un manoscritto di base (o esemplare di collazione), con il quale comparare tutti gli altri, è un procedimento comune durante la collazione manuale ed è rimasto tale, fino a pochissimo tempo fa, anche negli algoritmi di allineamento che costituiscono il cuore dei programmi di collazione. Il confronto di ogni testimone all'esemplare di riferimento, però, poco si adatta ai fini della collazione di testi medioevali, che è spesso quello di stabilire i rapporti tra i manoscritti: riprendendo la metafora di Spencer e Howe, è difficile stabilire la distanza tra Frascati e Viterbo, se si conosce solo quella tra le due città e la capitale. Le applicazioni più recenti di algoritmi sviluppati in bio-informatica alla critica testuale prevedono l'uso di algoritmi di allineamento multiplo, che permettono la collazione tra più testi senza il ricorso ad un esemplare di riferimento e cercano di limitare la dipendenza dei risultati dall'ordine di inserimento dei dati. Queste tecniche, che richiedono lo sviluppo di algoritmi di grande complessità, forniscono risultati promettenti nel campo della critica testuale. Alcuni esempi mostreranno le differenze tra i risultati dell'allineamento a coppie e quelli dell'allineamento multiplo. Tali differenze, già rilevanti quando il filologo analizza i risultati e ne trae conseguenze, assumono conseguenze importanti se è la macchina ad usare i risultati per un ulteriore trattamento, come nel caso dell'analisi filogenetica. 2. Durante la collazione manuale di testi medioevali, il filologo riconosce e decide di marcare o meno le differenze formali tra i testi; per gestire la stessa situazione, non banale, un computer ha bisogno di informazioni precise.

Il procedimento di collazione automatica, come descritto nel modello di Gothenburg presto divenuto standard, prevede essenzialmente tre fasi: tokenizzazione, normalizzazio-

[1] Gli esempi e le risorse utilizzate si concentrano sull'antico-francese, ma sono potenzialmente applicabili ad altre lingue medioevali.

ne e allineamento, a cui aggiungere un eventuale feedback e la visualizzazione dei risultati. I tentativi messi in atto per superare il problema della variabilità grafica, fonetica e morfologica si sono concentrati nella fase di normalizzazione e di allineamento. Nel primo caso, sono stati creati manualmente dei lessici di forme non marcate, attraverso i quali 'normalizzare' i testi originali; nel secondo caso, per l'allineamento non vengono considerate solo le categorie 'uguale' e 'diverso', ma anche un abbinamento approssimativo (fuzzy match) basato sulla distanza di edit.

Gli sviluppi recenti nell'ambito del trattamento automatico delle lingue medioevali rendono possibile un nuovo approccio nella fase della normalizzazione. La nostra proposta, in fase di sviluppo, per la collazione di testi antico francesi, prevede la preparazione del testo tramite la lemmatizzazione automatica, che sostituisce la normalizzazione manuale. Una volta lemmatizzati, i testi sono collazionati.

La normalizzazione in questo caso non è un'operazione che avviene nella testa del filologo, ma un procedimento documentato, che, attraverso il ricorso alla lemmatizzazione automatica, può cosituire un fertile punto di incontro tra linguistica computazionale e filologia digitale, portando al contempo all'arricchimento dei corpora e al miglioramento dei risultati della collazione.

Bibliografia

Andrews, T. 2013b. «The Third Way. Philology and Critical Edition in the Digital Age». *Variants* 10:61–76.

Dekker, R., e G. Middell. 2016. «CollateX». Online. http://collatex.net.

Gottesman, R., e S. Bennett, cur. 1970. *Art and Error: Modern Textual Editing*. London: Indiana University Press.

Irigoing, J., e G. P. Zarri, cur. 1979. *La Pratique des ordinateurs dans la critique des textes: Paris, [Colloque international], 29-31 mars 1978*. Paris.

Orlandi, T. 2007. «Teoria e prassi di una edizione computazionale». In *Digital Philology and Medieval Texts*, a cura di A. Ciula e F. Stella. Pisa.

Robinson, P. 2004. «Rationale and Implementation of the Collation System». In *The Miller's Tale on CD-ROM. The Canterbury Tales Project*. Leicester.

Schmidt, D. A. 2009. «Merging Multi-Version Texts: a General Solution to the Overlap Problem». In *Proceedings of Balisage: The Markup Conference 2009, Montréal, Canada, August 11 - 14, 2009*, vol. 3. Montréal.

Spadini, E. 2016b. «Lemmatizzazione e collazione». Online. https://github.com/elespdn/lemm-coll.

Spencer, M., e C. J. Howe. 2004. «Collating Texts Using Progressive Multiple Alignment». *Computers and the Humanities* 38 (3): 253–270.

TRACEsoftTools: building and integrating text encoding and visualization tools

Alejandro Bia, Centro de Investigación Operativa Universidad Miguel Hernández, Elche, Spain, abia@umh.es

This paper shows the results of more than 15 years of experience building and integrating a wide collection of useful XML-TEI related tools, starting at the creation of the Miguel de Cervantes Digital Library in 1999 and finishing with to the more recent TRACE project which served to integrate and evolve these tools to make working with XML and TEI faster and easier, while favouring mobility and portability by offering them as online services.

This set of tools contains:

- Tools to graphically visualize and design markup vocabularies and XML document instances.
- Tools for automatic TEI markup from a Markdown-style lightweight markup language.
- Tools to validate, pretty-print, improve and transform XML documents.
- Tools to render XML documents

Some of the tools used in this project already exist, and the rest are of our own breed. Not all of them are currently available, neither as online services, nor for download. The tools that are available for download are not generally available as web services, so they need to be installed according to different requirements and be configured properly. For examples of some of these tools see: DiRT[1], Cover Pages[2], Garshol's XML Tools[3], TEI Tools[4]. In some cases, there was some re-engineering of the tools to adapt them for client-server operation. We realized we could do this with tools that performed in an extended filter fashion. Apart from tool integration as web services, a virtual desktop with file management capabilities has been developed to provide a complete integrated work environment.

The idea of having the tools adapted to run as web services, allows for anyone to use them from anywhere, anytime, without the need of installation. This is particularly useful for Digital Humanities hands-on courses, where one of the main problems is to get all the necessary tools installed before the course takes place. This set of online services will

[1]DiRT Directory, a registry of digital research tools for scholarly use: http://dirtdirectory.org

[2]Robin Cover, Cover Pages: Software and Markup Languages: http://xml.coverpages.org/software.html

[3]Garshol. L. M. *Free XML tools and software*: http://www.garshol.priv.no/download/xmltools

[4]TEI Consortium. *TEI Tools*: http://www.tei-c.org/Tools

also suffice for emerging or small digitization projects, and as a display and workbench for DH scholars looking for tools. Furthermore, some intelligent piping of the tools can be arranged to follow certain processing workflows required for digital libraries or DH document processing. The platform can serve as a playground for workflow and task pipelining experiments.

Following the Software as a Service delivery model (SaaS), the purpose was to set up a web-server able to run different types of tools, and to develop a user friendly front end to allow beginners to operate the tools through a web browser (thin client) without the need for installation and configuration of myriad computer programs. In this sense, the DH-Workbench would serve as a teaching aid for beginners, and as an entry level solution for emerging DH projects, allowing for savings in software and installation costs.

The implementation is based on the popular and reliable XAMPP platform (Apache HTTP Server, MySQL, PHP, Perl), plus the Java Runtime Environment (JRE) and eventually any other runtime processor required by the tools to run (Saxon, Xalan, Phyton, etc.).

The following is a non-exhaustive list of the services and tools included in the DH Workbench:

- dtd2xs[5]: Dtd2Xs allows conversion of complex, modularized XML DTDs and DTDs with namespaces to W3C XML Schemas.
- DTDinst[6]: Converts DTDs to XML-DTD, i.e. DTD structure represented in XML format. Useful for XML processing of DTDs.
- DTDexplorer[7]: DTD explorer is a java 1.1 applet and application that loads a DTD and displays the parent/child relationship, allowing you to quickly understand the possibilities of the grammar of a DTD.
- DTDprune: A DTD simplification tool (Bia and Carrasco 2001).
- DTD/Schema comparison: This service allows the user to compare two different DTDs or Schemas.
- JavaScript validator: Works as JavaScript on the client side (Windows only)
- Mindmap diagrams generated from XML document instances to visualize and analyze the document structure (Bia, Muñoz, and Gómez 2010).
- Mindmap diagrams generated from DTDs and Schemas to visualize and analyze the structuring rules (Bia, Muñoz, and Gómez 2010).
- Multilingual Markup Translator (previously called “Multilingual Markup Website”) (Bia, Malonda, and Gómez 2006)
- rxp[8]: RXP is a validating XML parser written in C.

[5]See: http://www.syntext.com/products/dtd2xs
[6]See: http://www.thaiopensource.com/relaxng/dtdinst
[7]See: http://www.activemath.org/~paul/DTDexplorer
[8]See: http://www.cogsci.ed.ac.uk/~richard/rxp.html

- Tidy[9]: HTML Tidy is a computer program and a library the purpose of which is to fix invalid HTML and to improve the layout and indent style of the resulting markup. It was developed by Dave Raggett of W3C, then passed on to become a Sourceforge project.
- TEIdown: automatic markup tool that converts a document with a very simple and easy to apply lightweight markup to a valid TEI document (Bia 2015).
- Trang[10]: Receives DTDs, Relax-NG and Relax-NC Schemas, and XML document instances as input, to produce DTDs, XML Schemas (XSD), and Relax-NG and NC Schemas. Trang can also infer a schema from one or more example XML documents.
- xmllint[11]: The xmllint program parses one or more XML files, specified on the command line as xmlfile. It prints various types of output, depending upon the options selected. It is useful for detecting errors in XML documents.
- XSLTdoc[12]: The XSLTDoc Application helps you to browse and understand XSLT stylesheets. It shows summaries of stylesheets and explains each XSLT instruction in detail.
- XSLT ready-made transformations: Several standard ready-made transformations (e.g. tei2html).
- XSLT online transformations: This is an online service providing transformations of XML document instances by user-provided XSLT scripts, using one of several parsers offered: csxslt, MSXSL, Saxon (versions 6, 7 and 8), Xalan, Xerces, xml4j, xsltproc and XT
- XSV[13]: XML Schema Validator is an open source (GPLed) work-in-progress attempt at a conformant schema-aware processor.

All these tools have been integrated using an interface (DH workbench) that includes a file management view (desktop) that resembles a conventional file explorer, and an editing view that resembles an online editor. The file management area allows operations like: file upload and download, copy, move, rename, delete and selection of files for editing or processing. The editing view, from which transformations and processing functions can be launched, offers editing facilities to work with up to three files in parallel folders (e.g., input, transformation and output files). The online application allows for three types of users: not registered (can use the online tools but cannot store documents in the cloud), registered users (have their own work area where documents can be left to be used in future work sessions), and administrators (can perform application management tasks).

We hope that this online platform will serve to test and share new tools in the future. Hence, easy maintenance and upgradeability were amongst the main design goals of the project.

[9] See: http://tidy.sourceforge.net
[10] See: http://www.thaiopensource.com/relaxng/trang.html
[11] See: http://xmlsoft.org/xmllint.html
[12] See: http://www.jenitennison.com/xslt/utilities
[13] See: http://www.ltg.ed.ac.uk/~ht/xsv-status.html

Although power tools like these can enhance digital humanities training and production activities, we have to agree with Schreibman and Hanlon that "*tool development is indeed considered a scholarly activity by developers, but recognition of this work and rewards for it lag behind rewards for traditional scholarly pursuits (such as journal articles and book publication)*" (Schreibman and Hanlon 2010). In this sense is not always easy to find support and recognition for this type of projects.

Some of these tools will be briefly showcased during the presentation, and will soon be available online at: http://dhw.umh.es

Acknowledgements

This work has been developed within the TRACEsofTools project: Software Tools for Contrastive Analysis of Texts in Parallel Bilingual Corpora, and has been financed with aid FFI2012-39012-C04-02 from the VI National Plan for Scientific Research, Development and Technological Innovation of the MINECO (Ministry of Economy and Competitiveness of Spain).

References

Bia, A. 2015. «Down to TEI: use of extended markdown to speed-up the creation of TEI documents». In *15TH TEI Conference and Member's Meeting, Université Lumière Lyon 2, Lyon (France), 28-31 October 2015*.

Bia, A., and R. C. Carrasco. 2001. «Automatic DTD simplification by examples». In *ACH/ALLC 12001. The Association for Computers and the Humanities, The Association for Literary and Linguistic Computing, The 2001 Joint International Conference, pages 7-91 New York University, New York City, 13-17 June 2001.*

Bia, A., J. Malonda, and J. Gómez. 2006. «The Multilingual Markup Website». In *Digital Humanities 2006: The First ADHO International Conference, pages 26-31, Universite Paris-Sorbonne: Centre de Recherche Cultures Anglophones et Technologies de l'Information, 5-9 July 2006.* Ed. by C. Sun, S. Menasri, and J. Ventura.

Bia, A., R. Muñoz, and J. Gómez. 2010. «Using Mind Maps to Model Semistructured Documents». In *Research and Advanced Technology for Digital Libraries: 14th European Conference, ECDL 2010, Glasgow, UK, September 6-10, 2010. Proceedings*, ed. by M. Lalmas et al., 421–424. Berlin-Heidelberg: Springer. http://dx.doi.org/10.1007/978-3-642-15464-5_47.

Schreibman, S., and A. M. Hanlon. 2010. «Determining Value for Digital Humanities Tools: Report on a Survey of Tool Developers». *Digital Humanities Quarterly* 4 (2). http://www.digitalhumanities.org/dhq/vol/4/2/000083/000083.html.

Edizioni Digitali
Digital Editions
Posters

Mauro Zanchetta: Per la *Cronaca dello Pseudo-Brunetto*: l'edizione digitale dell'autografo

Mauro Zanchetta

Il contributo proposto si concentra sull'edizione digitale di un testo cronistico italiano tardomedievale, noto come *Cronaca dello Pseudo Brunetto (Latini)*, o, secondo la sua edizione più affidabile, come *Cronica fiorentina compilata nel secolo XIII*. Si tratta di un'opera storiografica adespota e incompiuta, allo stato di abbozzo, redatta in volgare tra fine Duecento e inizio Trecento, che narra la storia dei papi e degli imperatori da Cristo e da Augusto fino a inizio Trecento, concentrandosi però progressivamente, nei suoi tratti finali, che sono anche i più interessanti, sulla storia degli ultimi due/tre secoli della città di Firenze. Questo assetto è il risultato di una trafila testuale complessa, essendo l'opera sostanzialmente un'elaborazione e un ampliamento della cosiddetta *Cronaca dello Pseudo Petrarca*, che è a sua volta un volgarizzamento, ampliato con inserzioni di storia fiorentina, del *Chronicon pontificum et imperatorum* di Martino Polono (pubblicato attorno al 1275; il volgarizzamento è immediatamente successivo). Quanto al testimoniale, la cronachetta è tràdita, oltre che da un apografo quattrocentesco completo (ms. Gaddi 77 della Biblioteca Medicea Laurenziana) e da alcuni frammenti di Età Moderna, da un autografo acefalo di 15 cc., contenute nel ms. II.IV.323 della biblioteca Nazionale Centrale di Firenze. Quest'ultimo è in realtà una copia di lavoro, un esemplare di servizio, non una copia in pulito, e pertanto pone di fronte al filologo delicati problemi di carattere ecdotico. L'importanza del testo in esame, anche per la sua parziale autografia, non è secondaria, tanto dal punto di vista storiografico, quanto linguistico, letterario e culturale; tuttavia di esso non è disponibile un'edizione che soddisfaccia alle più moderne e complesse esigenze ecdotiche.

La lacuna bibliografica sta per essere colmata, almeno parzialmente, grazie all'edizione digitale dell'autografo della *Cronaca*, ormai in via di compimento: l'edizione è basata sulla codifica TEI XML e da me allestita tramite il software EVT 1.0, sviluppato da un gruppo di ricerca coordinato dal professor Roberto Rosselli Del Turco dell'Università di Torino. Questo software permette, una volta completata la codifica XML del testo da pubblicare: la visualizzazione a computer e la fruibilità *online* di riproduzione fotografica, trascrizione diplomatica ed edizione interpretativa (critico-interpretativa) del manoscritto su *frames* separati, e quindi il loro confronto; la possibilità di *zoom in/out* per le immagini del codice; il collegamento testo-immagine; la presenza di *hotspots* per i dettagli più intreressanti del manoscritto; la ricerca, attraverso un'apposita maschera, di parole o combinazioni più piccole di caratteri e l'individuazione precisa delle loro occorrenze nell'opera; la creazione, grazie al *markup* semantico, di liste di antroponimi, di toponimi e di dati cronologici; l'inserimento, infine, da parte del curatore, di prolegomena al testo e di note esplicative di qualsiasi tipo: filologico-testuale, linguistico, storiografico, ecc.

L'allestimento di un'edizione tramite EVT 1.0 di un'opera come quella in esame in particolare permette:

- un accertamento del testo che sia presumibilmente più vicino alla volontà dell'autore rispetto a quello delle pur benemerite edizioni di fine Ottocento e della prima metà del Novecento, perché costantemente riscontrato sul codice autografo, base dell'edizione;
- la fruizione agevole e vasta dell'edizione (grazie alla pervasività del mezzo informatico e del web, e alla distribuzione *open-source*) da parte di un pubblico differenziato, formato ad esempio da: 1. codicologi e paleografi, verosimilmente interessati alla riproduzione fotografica e alla trascrizione diplomatica del manoscritto; 2. storici della lingua e della grafia, interessati, oltre che alla lettura diretta del manoscritto e della trascrizione del testo, alla ricerca di combinazioni grafiche, di lessemi, di forme e costrutti individuabili grazie alla maschera di ricerca; 3. storici medievali *tout court*, ai quali è messa a disposizione un'importante fonte, corredata da note di commento, da rinvii alle fonti, da un apparato formato da liste di nomi, di luoghi e di date; 4. italianisti, e filologi in particolare, che, oltre a disporre di un testo ecdoticamente più solido, hanno la rara opportunità di leggere un esemplare di lavoro di veneranda antichità e quindi di entrare, per così dire, nel laboratorio di uno scrittore delle Origini e di vedere il testo nel suo progressivo divenire; dal punto di vista filologico, inoltre, il confronto continuo del testo 'definitivo' con la riproduzione fotografica dell'autografo permette forse di superare un limite, da più parti avvertito, della filologia d'autore, cioè l'eccessiva formalizzazione e complessità, insomma la difficile leggibilità, anche da parte degli addetti ai lavori, degli apparati diacronici delle varianti; 5. ultimi, ma non meno importanti, studenti delle facoltà umanistiche, 'allievi filologi', che possono confrontare diverse modalità di trascrizione di un testo e vedere concretamente esemplificate le operazioni che conducono dal testo manoscritto alla sua edizione scientifica.

I temi di riflessione sollecitati dal *call for papers* del convegno AIUCD 2016 hanno certo a che fare con l'ottica in cui si iscrive il mio lavoro, segnatamente:

- il dialogo attorno al testo da parte di comunità scientifiche diverse;
- il superamento della dicotomia tradizionale tra edizione diplomatica ed edizione critica (e, aggiungerei, edizione commentata), grazie all'elasticità e alla fruibilità per livelli differenziati permesse dall'edizione digitale di un testo;
- (strettamente connesso col punto precedente) l'avvicinamento tra la comunità delle Digital Humanities e la comunità delle discipline umanistiche tradizionali.

Bibliografia

Martino Polono e le compilazioni storiche universali tardomedievali – Ed. di riferimento

Weiland, L., cur. 1872. *Martini Oppaviensis Chronicon Pontificum et Imperatorum*, MGH, Scriptores 22.

Martino Polono e le compilazioni storiche universali tardomedievali

Brincken, D. von den. 1986. «Inter spinas principum terrenorum. Annotazioni sulle summe e sui compendi storici dei Mendicanti». In *Aspetti della letteratura latina del secolo XIII . Atti del primo Convegno internazionale di studi dell'AMUL, Perugia 3-5 ottobre 1983*, a cura di C. Leonardi e G. Orlandi, 77–103. Firenze-Perugia.

Guenée, B. 1986. «Lo storico e la compilazione nel XIII secolo», a cura di C. Leonardi e G. Orlandi, 57–76. Firenze-Perugia.

La storiografia volgare delle Origini

Del Monte, A. 1950a. «La storiografia fiorentina dei secoli XII e XIII». *Bullettino dell'Istituto Storico Italiano per il Medio Evo e Archivio Muratoriano* 62:175–282.

Santini, P. 1903. *Quesiti e ricerche di storiografia fiorentina*. Firenze.

Villari, P., cur. 1894. *I primi due secoli della storia di Firenze*. 1:37–58. Firenze.

Zabbia, M. 2012. «Prima del Villani. Nota sulle cronache universali a Firenze tra l'ultimo quarto del Duecento e i primi anni del Trecento». In *Le scritture della storia*, a cura di F. Delle Donne e G. Pesiri, 138–162. Roma.

Storia della Firenze bassomedievale

Davidsohn, R. 1966–1978. *Storia di Firenze*. Firenze.

Ottokar, N. 1962. *Il comune di Firenze alla fine del Dugento*. Torino.

Raveggi, S. 1978. *Ghibellini, guelfi e popolo grasso*. Firenze.

La *Cronaca* dello pseudo-Brunetto – Ed. di riferimento

Schiaffini, A., cur. 1954 (1926). «Cronica fiorentina compilata nel secolo XIIII». In *Testi fiorentini del Dugento e dei primi del Trecento, […]* 82–150. Firenze.

La *Cronaca* dello pseudo-Brunetto

D'Agostino, A. «Itinerari e forme della prosa». In *Storia della letteratura italiana diretta da E. Malato*, 1:527–630. Vedere in particolare pp. 588–591. Roma.

Del Monte, A. 1950b. «La storiografia fiorentina dei secoli XII e XIII». *Bullettino dell'Istituto Storico Italiano per il Medio Evo e Archivio Muratoriano* 62:186–187.

Faini, E. 2009. «Il convito fiorentino del 1216». In *Conflitti, paci e vendette nell'Italia comunale*, a cura di A. Zorzi, 105–130. Firenze.

La filologia d'autore

Battaglia Ricci, L. 2010. «Edizioni d'autore, copie di lavoro, interventi di autoesegesi: testimonianze trecentesche,» in *Di mano propria: gli autografi dei letterati italiani. Atti del convegno internazionale di Forlì, 24-27 novembre 2008*, a cura di G. Baldassarri, 123–157. Roma.

Italia, P., e G. Raboni. 2010. *Che cos'è la filologia d'autore*. Roma.

Gli autografi medievali

Brunetti, G., M. Fiorilla e M. Petoletti, cur. 2013. *Gli autografi dei letterati italiani. Dalle Origini al Trecento,* vol. 1. Dir. M. Motolese, P. Procaccioli ed E. Russo. Roma.

Chiesa, P., e L. Pinelli, cur. 1994. *Gli autografi medievali, problemi paleografici e filologici. Atti del convegno di studio della Fondazione Elio Franceschini, Erice, 25 settembre-2 ottobre 1990*. Premessa di C. Leonardi (37–60). Firenze.

Text Encoding Initiative

TEI Consortium. 2016. *TEI P5: Guidelines for Electronic Text Encoding and Interchange, version 3.0.0, last updated on 29th March 2016*. http://www.tei-c.org/Guidelines/P5.

Edition Visualization Technology

Capochiani, F., C. Leoni e R. Rosselli Del Turco. 2014. «Codifica, pubblicazione e interrogazione sul web di corpora diplomatici per mezzo di strumenti open source,» in *Digital diplomatics. The computer as a tool for diplomatists?,* a cura di A. Ambrosio, S. Barret e G. Vogeler, 31–60. Köln-Weimar-Wien. http://bit.ly/2cevT6f.

Fiorentini, F. 2007–2008. «EVT: Edition Visualization Technology; Progettazione e sviluppo un software per la consultazione di edizioni digitali». Tesi di laurea mag., Università degli Studi di Pisa. http://etd.adm.unipi.it/theses/available/etd-09132008-133733.

Rosselli Del Turco, R. 2011. «After the editing is done: designing a Graphic User Interface for Digital Editions,» *Digital Medievalist Journal* 7. http://digitalmedievalist.org/journal/7/rosselliDelTurco.

Rosselli Del Turco, R., et al. 2014–2015. «Edition Visualization Technology: A Simple Tool to Visualize TEI-based Digital Editions». *Journal of the Text Encoding Initiative* 7. http://jtei.revues.org/1077.

Team EVT. «EVT 1.0». Online (software download). https://sourceforge.net/projects/evt-project.

Il Progetto "I libri dei Patriarchi 2.0. Un percorso multimediale nella cultura scritta del Friuli medievale": obiettivi, scelte metodologiche, elementi di innovatività ed originalità

Stefano Allegrezza, Università degli Studi di Udine, stefano.allegrezza@uniud.it
Nicola Di Matteo, Altavisio srl – Algoweb, Bologna, nicoladimatteo@gmail.com
Sandro Piussi, Ufficio Beni Culturali, Diocesi di Udine, direzione@archiviodiocesano.it
Cesare Scalon, Istituto Pio Paschini per la storia della Chiesa in Friuli (Udine), cesare.scalon@uniud.it
Egidio Screm, Istituto Pio Paschini per la storia della Chiesa in Friuli (Udine), egidio.screm@teletu.it

1 Introduzione

Il progetto "*I libri dei Patriarchi 2.0. Un percorso multimediale nella cultura scritta del Friuli medievale*" si inserisce nell'ambito di un percorso di ricerca e di approfondimento frutto di un decennio di intensa investigazione scientifica, volto a ricostruire la storia culturale del Friuli medievale valorizzando e rendendo disponibile il patrimonio codicologico antico, di ineguagliabile valenza storica, documentaria e artistica, conservato sia presso prestigiose istituzioni del Friuli Venezia Giulia – quali la Biblioteca Civica Guarneriana di San Daniele del Friuli, la Biblioteca Capitolare e la Biblioteca Patriarcale di Udine, la Biblioteca del Seminario Teologico di Gorizia, la Biblioteca del Museo Archeologico Nazionale di Cividale del Friuli – sia presso alcune fra le principali biblioteche europee ed americane[1], stimolando e sollecitando un pubblico molto più ampio rispetto a quello rappresentato dalla ristretta cerchia degli studiosi e degli specialisti della materia e che ha come scopo precipuo l'attivazione di dinamiche finalizzate alla promozione del turismo culturale, al coinvolgimento delle scuole e alla divulgazione scientifica, grazie ad un sapiente utilizzo delle tecnologie dell'informazione e ad alcune scelte metodologiche lungimiranti ed innovative. Tutto è partito da un progetto che si proponeva di offrire una rilettura in forma multimediale dei preziosi materiali raccolti nel volume "*I libri dei patriarchi. Un percorso nella cultura scritta del Friuli medievale*" (edito nel 2014 dalla Deputazione di Storia Patria per il Friuli e dall'Istituto

[1]Su un centinaio di libri interi o frammentari presentati nel volume a stampa "I libri dei patriarchi. Un percorso nella cultura scritta del Friuli medievale" da cui il portale, circa un terzo proviene da biblioteche straniere e rappresenta una selezione di codici di provenienza friulana individuati o microfilmati nel corso della ricerca.

Pio Paschini, a cura di Cesare Scalon) allo scopo di "promuovere la divulgazione scientifica nelle scuole e il turismo culturale"; ben presto, tuttavia, ci si è resi conto della necessità di far "esplodere" tanta massa documentaria in un percorso che agevoli nell'utente la ricerca e la "navigazione" all'interno delle diverse sezioni che fra loro si intersecano e si compenetrano, e che è già evidente nella veste editoriale cartacea, ha trovato una sua concretizzazione nella realizzazione di un portale web (`http://www.libridei patriarchi.it`) capace di enfatizzare e ampliare e rendere ancor più stimolante l'esplorazione dei documenti messi, finalmente, a disposizione di un pubblico virtualmente senza confini (cfr. Fig. 1).

2 Gli obiettivi del progetto

Il progetto è stato pensato avendo bene in mente tre obiettivi principali: la promozione del turismo culturale, il coinvolgimento dei docenti e delle scuole superiori, la divulgazione scientifica. Grazie alla sua veste grafica estremamente accattivante e di facile consultazione, agevolato da traduzioni in varie lingue (inglese innanzitutto, ma anche tedesco, friulano, sloveno, etc.), il portale è pensato come una straordinaria opportunità rivolta al turismo culturale e potrà invogliare i turisti europei (e non solo) a visitare la regione Friuli Venezia Giulia iniziando proprio da quel circuito virtuoso rappresentato dalle città di Udine, San Daniele del Friuli, Cividale del Friuli, che nelle biblioteche Patriarcale e Capitolare, Civica Guarneriana, del Museo Archeologico Nazionale conservano i preziosi tesori resi finalmente accessibili attraverso il portale.

Figura 1: La home page del portale "I libri dei patriarchi".

La scuola è un altro interlocutore privilegiato del progetto. Gli studenti delle scuole avranno per la prima volta l'opportunità di giovarsi di materiali documentari utilissimi per un approfondimento della storia regionale - altrimenti difficilmente reperibili - e per di più pensati entro strutture di consultazione elettroniche rese appetibili da un'interfaccia grafica accattivante e capace di stimolare la comunicazione e la divulgazione dei contenuti, in virtù di una serie di codici mediatici fra loro intrecciati (testi, immagini, video, musica, animazioni, apparati iconografici). Gli studenti potranno così giovarsi di un repertorio di

fonti rilette nella loro straordinaria ed eterogenea molteplicità, accedendo ad approfondimenti ipermediali in cui gli esperti potranno loro illustrare, con un linguaggio semplice e accattivante, i temi di volta in volta affrontati durante la "navigazione". Finalmente, in virtù di questo progetto, il repertorio storiografico del Medioevo friulano potrà entrare nei programmi della scuola quale sussidio ai tradizionali corsi di studio e potrà essere utilizzato come laboratorio di ricerca in cui andare a verificare le abilità acquisite nell'ambito dei curricula scolastici tradizionali.

Infine, grazie al portale, i repertori artisticamente più belli saranno per la prima volta messi a disposizione di un pubblico molto più vasto di quello costituito dagli studiosi e dagli esperti, gli unici, fino ad ora, ammessi alla consultazione di tanta bellezza. Il patrimonio codicologico antico del Friuli Venezia Giulia (cfr. Fig. 2) potrà essere portato alla conoscenza di chiunque: non più solo di una ristretta cerchia di pochi iniziati ed esperti, ma, più in generale, di un pubblico mediamente colto. Inoltre sarà possibile "sfogliare" i codici esattamente come se ci si trovasse fisicamente nell'istituto culturale presso il quale sono custoditi, aprendo, in questo modo la loro fruizione ad un pubblico molto vasto ed in maniera indipendente dalla sua collocazione geografica (cfr. Fig. 3).

3 Le scelte metodologiche: elementi di innovatività ed originalità

Il progetto è innovativo sia per i suoi contenuti che per le sue forme: per i contenuti, perché i repertori artisticamente più belli sono stati per la prima volta messi a disposizione di un pubblico molto più vasto di quello costituito dagli esperti e dagli studiosi; per le forme, perché la realizzazione informatica, pensata per agevolare al massimo la fruibilità e la leggibilità dei materiali, si è concretizzata in una veste grafica accattivante forse mai utilizzata prima in Italia e ben lontana dalle consuete interfacce di navigazione, molto scarne e assolutamente prive di qualsiasi elemento di attrattività, che caratterizzano le realizzazioni di progetti simili.

Le risorse digitali del portale (non solo immagini, ma anche contenuti audio e video, percorsi geografici e cronologici, etc.) sono state strutturate in maniera tale da rendere disponibili, rispetto all'edizione cartacea, percorsi differenziati ma tutti fra loro intercorrelati; in questo modo i diversi codici di lettura (quello iconografico, quello linguistico, quello musicale, quello visivo, etc.) hanno potuto convergere per una rivisitazione virtuale dei contenuti, senz'altro godibilissima e per certi aspetti anche ludica, di cui mai prima d'ora è stato possibile giovarsi. Ciò ha richiesto una intensa attività redazionale necessaria per rivedere integralmente i contenuti e "riprogettarli" ai fini della fruizione sul web.

Conclusioni. Il portale è stato presentato pubblicamente il 4 dicembre 2015 e da allora ha riscosso un notevole successo, testimoniato non solo dalle molteplici manifestazioni di apprezzamento ricevute, ma anche dall'elevato numero di accessi sia da parte di visitatori italiani che esteri. Infatti, come si può osservare dalla Figura 4, che riporta graficamente il numero degli accessi al portale (il numero di visitatori è direttamente proporzionale all'intensità del colore azzurro), vi sono visite non solo da Paesi europei come l'Austria, la Slovenia, la Croazia, la Germania, la Francia ma anche da Paesi extra-europei come gli Stati Uniti, il Canada, il Brasile, l'Argentina, il Sudafrica, l'Australia, la Russia, la Cina e perfino

artisticamente più belli sono stati per la prima volta messi a disposizione di un pubblico molto più vasto di quello costituito dagli esperti e dagli studiosi; per le forme, perché la realizzazione informatica, pensata per agevolare al massimo la fruibilità e la leggibilità dei materiali, si è concretizzata in una veste grafica accattivante forse mai utilizzata prima in Italia e ben lontana dalle consuete interfacce di navigazione, molto scarne e assolutamente prive di qualsiasi elemento di attrattività, che caratterizzano le realizzazioni di progetti simili.

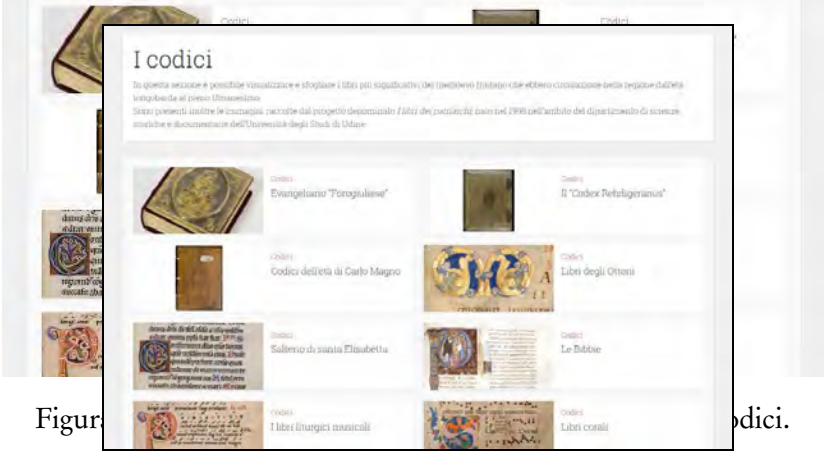


Figura 2. La sezione del portale che consente l'accesso ai codici.

Figura 3. Il Salterio di S. Elisabetta interamente "sfogliabile" on-line

Le risorse digitali del portale (non solo immagini, ma anche contenuti audio e video, percorsi geografici e cronologici, etc.) sono state strutturate in maniera tale da rendere disponibili, rispetto all'edizione cartacea, percorsi differenziati ma tutti fra loro intercorrelati; in questo modo i diversi codici di lettura (quello iconografico, quello linguistico, quello musicale, quello visivo, etc.) hanno potuto coesistere nel rispetto della rivisitazione virtuale dei contenuti, senza alcuna possibilissima e per certi aspetti anche ludica, di cui mai prima d'ora è stato possibile giovarsi. Ciò ha richiesto una intensa

l'Indonesia. Anche il posizionamento su motori di ricerca è eccellente, considerati i pochi mesi trascorsi dal lancio del portale; basti pensare che il suo indirizzo web viene restituito da Google come primo risultato quando si inseriscono come chiavi di ricerca le parole "libri" e "patriarchi" (questo risultato di tutto rilievo non è assolutamente scontato, come ben sa chi si occupa di SEO, *Search Engine Optimization*); il portale risulta ai primissimi posti anche quando vengono inserite chiavi di ricerca come "salterio di Egberto" o "summa artis notariae" (terza posizione), "beato Bertrando" (quarta posizione), "patriarcato di Grado" o "la patria" (quinta posizione), "la patria del Friuli" (sesta posizione) e così via. Si tratta di risultati notevoli raggiunti in poco tempo, frutto di una attenta progettazione e realizzazione del portale, e questo non fa che confermare la bontà del progetto.

Figura 4: La ripartizione su base geografica dei visitatori del portale

Bibliografia

Scalon, C., cur. 2014. *I Libri dei patriarchi. Un percorso nella cultura scritta del Friuli medievale*. Udine: Deputazione di Storia patria per il Friuli – Istituto Pio Paschini per la Storia della Chiesa in Friuli.

Progetto Mambrino: Metodologie e prospettive di un'edizione digitale

Stefano Bazzaco, Università degli Studi di Verona, stefano.bazzaco.1@gmail.com
Tiziana Mancinelli, Cologne Centre of eHumanities, tiziana.mancinelli@uni-koeln.de

L'obiettivo di questo poster è quello di presentare le fasi di realizzazione di un progetto di edizione digitale di un copioso numero di testi curato dal "Progetto Mambrino", un gruppo di ricerca nato nel 2003 su iniziativa di Anna Bognolo (Università di Verona) che si occupa dello studio dei romanzi cavallereschi italiani di derivazione spagnola prodotti a Venezia a metà del Cinquecento (1540-1630).

Il gruppo si propone di studiare l'esteso corpus di traduzioni, imitazioni e continuazioni originali italiane di cicli spagnoli di enorme successo che si moltiplicarono per più di un secolo grazie all'attività imprenditoriale di editori e poligrafi veneziani, grandi interpreti del mercato librario rinascimentale che contribuirono alla diffusione del romanzo cavalleresco in tutta Europa.

L'indagine del ricco panorama testuale, che comprende circa 350 edizioni, si è inizialmente concentrata sulla ricerca bibliografica e sulla localizzazione degli esemplari dei romanzi cavallereschi italiani d'ispirazione spagnola, giungendo alla realizzazione di un Catalogo pubblicato sul sito del Progetto Mambrino (www.mambrino.it) a cura di Stefano Neri, in corso di continuo aggiornamento. Il ciclo più conosciuto e influente, quello di *Amadís de Gaula*, è stato studiato a fondo nel volume Bognolo, Cara, Neri, *Repertorio delle continuazioni italiane ai romanzi cavallereschi spagnoli. Ciclo di Amadis di Gaula* (2013): l'opera contiene uno studio preliminare sulla fortuna del romanzo cavalleresco in Italia, un censimento degli esemplari conservati e il riassunto e l'indice dei personaggi per ciascuno dei tredici libri che compongono il ciclo.

Grazie alla collaborazione con biblioteche locali che conservano una quantità significativa di esemplari appartenenti al corpus, ha preso avvio l'iniziativa di restituire a specialisti e lettori una versione consultabile in rete dei romanzi cavallereschi studiati, di cui ancora non esiste un'edizione moderna. Un primo passo in questa direzione è stata la creazione di un archivio digitale di libera consultazione contenente le riproduzioni fotografiche in alta risoluzione dei volumi appartenenti al ciclo amadisiano conservati presso la Biblioteca Civica di Verona, già pubblicato sul sito del Progetto, cui successivamente si aggiungeranno riproduzioni di volumi provenienti dai fondi della Biblioteca Bertoliana e la Biblioteca internazionale La Vigna di Vicenza.

A partire dal 2015 ha preso corpo l'idea di realizzare edizioni scientifiche digitali di questi romanzi, Digital Scholarly Edition (DSE). L'obiettivo è mettere a disposizione di ricercatori e lettori un testo affidabile e facilmente accessibile, affiancando al facsimile una trascrizione dei testi rispondente a criteri filologici condivisi, consultabile in rete in modo interattivo.

Sarà innanzitutto resa disponibile una versione interrogabile e liberamente accessibile di quello che è stato scelto come testo pilota: *Le prodezze di don Florarlano* di Mambrino Roseo da Fabriano, per arrivare in seguito alla creazione di una collezione completa del ciclo di *Amadís* (all'incirca venti opere se si considerano le traduzioni e le continuazioni originali italiane).

Questo poster provvederà a evidenziare le problematiche più significative incontrate durante la realizzazione di un'edizione digitale con lo scopo di presentare le possibili soluzioni e le metodologie attuabili. Il modello di codifica scelto, conforme agli standard TEI (*Text Encoding Initiative*) e inizialmente vincolato ad un sistema complesso di descrizione e annotazione del testo, è stato successivamente rivisto tenendo in considerazione solo un *subset* dello schema principale, consentendo in questo modo la pubblicazione immediata dei materiali. Il modello di marcatura attualmente predisposto consentirà di includere una descrizione dettagliata degli aspetti materiali della cinquecentina e di distinguere i livelli strutturali del testo (capitoli, paragrafi, etc.), cui successivamente si integreranno dei collegamenti all'indice dei nomi e ai riassunti di ogni capitolo.

In seguito alla fase di modellizzazione e rappresentazione formale dei dati, il lavoro è stato caratterizzato dalla scelta di un *tool* per la visualizzazione. Ci si è affidati quindi a EVT (*Edition Visualization Technology*), un software di trasformazione basato sul collegamento e la visualizzazione doppia di testo e immagine. Sebbene in linea con gli obiettivi del progetto, EVT è uno strumento progettato nell'ambito dell'edizione di manoscritti medievali; si ricorrerà pertanto alla modifica di alcuni parametri al fine di mettere a disposizione degli studiosi una versione concepita specificatamente per la rappresentazione digitale di "early printed books".

Il Progetto Mambrino contribuirà inoltre al dibattito scientifico allegando una riflessione che metta in luce i vantaggi e gli svantaggi relativi all'utilizzo di infrastrutture e di *tools* nell'ambito delle pubblicazioni digitali.

Bibliografia

Bognolo, A., G. Cara e S. Neri. 2013. *Repertorio delle continuazioni italiane ai romanzi cavallereschi spagnoli. Ciclo di Amadis di Gaula*. Roma: Bulzoni.

Bognolo, A., e S. Neri. 2014. «Progetto Mambrino. Resultados y Perspectivas». *Janus. Estudios Sobre El Siglo De Oro* 3.

Buzzetti, D., e J. McGann. 2006. «Electronic Textual Editing: Critical Editing in a Digital Horizon». In *Electronic Textual Editing*, a cura di L. Burnard, K. O'Brien O'Keeffe e J. Unsworth. New York: Modern Language Association of America. http://www.tei-c.org/About/Archive_new/ETE/Preview/mcgann.xml.

Pierazzo, E. 2014. *Digital Scholarly Editing: Theories, Models and Methods*. hal-01182162. https://hal.inria.fr/hal-01182162/document.

Renear, A. H. 2004. «Text Encoding». In *A Companion to Digital Humanities*, a cura di S. Schreibman, R. Siemens e J. Unsworth. Oxford: Blackwell.

Rosselli Del Turco, R. 2009a. «Il progetto Vercelli Book Digitale: codifica e visualizzazione di un'edizione diplomatica grazie alle norme TEI P5». In *Medieval texts - contemporary media: The art and science of editing in the digital age*, 131–152. Como – Pavia.

Rosselli Del Turco, R., et al. 2015. «Edition Visualization Technology: A Simple Tool to Visualize TEIBased Digital Editions». *Journal of the Text Encoding Initiative* 8.

Rydberg-Cox, J. A. 2009a. «Digitizing Latin Incunabula: Challenges, Methods, and Possibilities». *Digital Humanities Quarterly* 3 (1). http://www.digitalhumanities.org/dhq/vol/3/1/000027/000027.html.

Sperberg-McQueen, C., e L. Burnard. 2007. *Guidelines for Electronic Text Encoding and Interchange*. Oxford-Providence-Charlottesville-Nancy.

Problemi ecdotici dell'opera di Goffredo da Viterbo. Speculum Regum e il Paris NAL 299, il testimone più antico conservato: proposta di un'edizione diplomatica digitale

Antonio Corvino, Università degli Studi Suor Orsola Benincasa di Napoli, corvino.antonio@ymail.com
Francesco Pacia, Università degli Studi di Salerno, fpacia@unisa.it - aslan88fpa@gmail.com

1 Problemi ecdotici dell'opera di Goffredo da Viterbo e il testimone più antico dello Speculum Regum [F. Pacia]

Nel 1872 Georg Waitz pubblicò per i *Monumenta Germaniae Historica* l'opera completa di Goffredo da Viterbo, notaio e cappellano alla corte del Barbarossa, dedicandogli gran parte del tomo XXII della sezione *Scriptores*. I *Gesta Friderici* e i *Gesta Heinrici VI* erano già stati pubblicati nel 1870 dallo stesso Waitz e ripubblicati con gli altri tre testi di Goffredo: lo *Speculum regum*, la *Memoria seculorum* e il *Pantheon*, opere di storiografia universale.

L'edizione, per quanto utilissima, mostra ormai molti limiti: l'editore non pubblicò interamente le opere omettendo, come segnalava lui stesso, le parti non storiche. Data, poi, la complessa elaborazione delle opere — si tratterebbe di un'unica opera continuamente rielaborata, a partire dalla materia dello *Speculum* ampliata e risistemata nella *Memoria* e poi nel *Pantheon* –, Waitz compì una scelta ecdotica di semplificazione, editando solo la prefazione della *Memoria* e collazionando per il testo del *Pantheon* cinque redazioni (due della *Memoria* e tre del *Pantheon*) diverse per contenuto, dedica e finalità, e che ebbero una tradizione propria indipendentemente da una pubblicazione dell'autore.

Waitz, inoltre, ignorava una quindicina di manoscritti, tra cui alcuni molto vicini all'autore cronologicamente e geograficamente. Essi possono integrare il lavoro di Waitz e hanno permesso di conoscere altri testi di Goffredo, come il manoscritto Paris, BNF, NAL 299, finora il più antico testimone conosciuto dello *Speculum regum*, costituito da 26 fogli di pergamena, 250x165. Il codice, redatto su due colonne da mani italiane, è di area viterbese e risale alla prima metà del XIII secolo. Al suo interno si trovano, oltre allo *Speculum regum* (ff. 1r-13r; 14v-15v), la *Denominatio regnorum imperio subiectorum* (ff. 13r-14v), bozza di un'altra opera di Goffredo non altrimenti conosciuta, e frammenti di altri testi goffrediani (ff. 15v-22v), circostanza che fa pensare che esso sia apografo di una copia di lavoro di Goffredo.

I limiti dell'edizione di Waitz oggi possono essere opportunamente ripensati e risanati con la filologia digitale: essa permette, per l'assenza di vincoli '*spaziali*', di dar conto di tutti gli stadi -– dallo *Speculum regum* al *Pantheon* – della scrittura di Goffredo e soprattutto dei manoscritti che li tramandano, confrontando i vari testi e le varie redazioni più direttamente, cogliendone somiglianze e differenze, in vista della ricostruzione dell'ultima volontà dell'autore.

2 Paris NAL 299: proposta di edizione diplomatica digitale [A. Corvino]

Il progetto di una moderna edizione di Goffredo è attualmente partito col primo testo del *corpus*, lo *Speculum regum*, e coll'edizione diplomatica digitale del suo manoscritto più antico.

La presenza di più opere all'interno del codice BNF NAL 299, intersecate tra loro, si prestano perfettamente alle esigenze delle edizioni diplomatiche digitali: un'*image-based digital edition* non come un'edizione diplomatica ipertestuale, ma un testo avente come supplemento le immagini del manoscritto. Una serie di funzionalità quali una *virtual magnifying glass*, il restauro virtuale ed un motore di ricerca testuale. Il supporto digitale è uno dei migliori strumenti attraverso cui è innestabile il processo di edizione diplomatica *tecnologico*: esso consente allo studioso di avere un'alternativa alla consultazione diretta del manoscritto, avvalendosi del supporto bifronte testo / immagine. Con alla base il linguaggio XML per la codifica dei testi letterari, secondo gli standard definiti dalla TEI (*Text Encoding Initiative*).

La TEI fornisce *guidelines* necessarie per la codifica dei testi, creando modelli con precise caratteristiche e regole, basandosi sul sistema di codifica XML, un metalinguaggio per la definizione di linguaggi di *markup*, ovvero un linguaggio marcatore basato su un meccanismo sintattico che consente di definire e controllare il significato degli elementi contenuti in un documento o in un testo. Essi devono, infatti, essere:

- sufficienti per la rappresentazioni delle caratteristiche testuali dei documenti;
- indipendenti dai software;

ed avere:

- compatibilità con standard esistenti ed emergenti;
- ricchezza del markup definita dall'utente;
- documentazione del testo e della sua codifica.

EVT (*Edition Visualization Editing*) è concepito come visualizzatore di file codificati in XML-TEI P5: un *front-end* basato sui tag presenti nel file XML-TEI, progettato dal prof. Roberto Rosselli Del Turco e dal suo staff. L'editore può concentrarsi effettivamente sulla trascrizione e la codifica dei vari elementi del testo allestendo un unico file, avendo garantito la perfetta coerenza e logica dell'edizione, facilitando il processo di revisione e di *update*.

D'altronde, scegliere una codifica XML-TEI P5 significa usare uno standard per le edizioni digitali, creando un'interazione con i progetti presenti in altre parti del mondo. È ben noto come la durata delle soluzioni *home-made* è estremamente ridotta, poggiando sull'energia delle persone che le hanno create; mentre l'azione dei singoli editori e sviluppatori viene potenziata dall'esistenza di una comunità consolidata come la comunità TEI.

Vi sarà quindi una codifica descrittiva attraverso il linguaggio XML-TEI, che marca la parola secondo taggature differenti, markup indicato dalla TEI per le edizioni diplomatiche digitali. Lo sguardo sarà rivolto, sin da subito, ad una successiva edizione critica attraverso lo sviluppo dello stesso EVT. Soluzione questa di certo innovativa che potrebbe, attraverso le disponibilità offerteci dai nuovi linguaggi di programmazione che superano i vecchi limiti, che vanno dalla visualizzazione all'inserimento di annotazioni e/o dati finora esclusi dalle edizioni critiche tradizionali in quanto non rientranti nei layout predefiniti. Questo *software*, per ora, consentirà di sfogliare virtualmente il manoscritto in diversi modi:

- esaminando i singoli fogli che lo compongono, utilizzando strumenti per l'ingrandimento di particolari dettagli, come i filtri per l'elaborazione grafica dell'immagine digitalizzata;
- esaminando la singola immagine con accanto l'edizione diplomatica del manoscritto, con annessa la possibilità di collegare il testo dell'edizione al punto corrispondente dell'immagine e viceversa;
- confrontando tra loro due pagine di testo, senza mai rinunciare alla possibilità di richiamare rapidamente l'immagine corrispondente, effettuando ricerche testuali complesse sul testo del manoscritto, sia a livello paleografico, sia a livello filologico-letterario.

Edizione diplomatica, questa, che è parte di un più grande progetto di edizione critica digitale dell'intero *corpus* di Goffredo da Viterbo, che consta di una quarantina di testimoni manoscritti. Sempre attraverso il lavoro ottenuto dalla mediazione tra codifica testuale e visualizzazione grafica delle codifiche XML-TEI va ad innestarsi l'evoluzione di EVT che, attraverso il supplemento di elementi innovativi quale ad esempio il *framework* AngularJS, va a coniare la perfetta sinergia tra edizione diplomatica attuale e quella critica futura.

Bibliografia

Buzzoni, M. 2005. «Le edizioni elettroniche dei testi medievali fra tradizione e innovazione: applicazioni teoriche ed empiriche all'ambito germanico». *Annali di Ca' Foscari: Rivista della Facoltà di Lingue e Letterature Straniere dell'Università Ca' Foscari di Venezia* 44 (1–2): 41–58.

Buzzoni, M., e R. Rosselli Del Turco. 2016. «Evolution or Revolution? Digital Philology and Medieval Texts: History of the Discipline and a Survey of Some Italian Projects». In *Mittelalterphilologien heute / Medieval Philologies Today*, forthcoming. Würzburg: Königshausen und Neumann.

Delisle, L. 1891. *Manuscrits latins et français ajoutés aux fonds des nouvelles acquisitions pendant les années 1875-1891. Inventaire alphabétique.* Vol. 1. Paris: H. Champion Libraire.

Dipper, S. 2005. «XML-based Stand-off Representation and Exploitation of Multi-Level Linguistic Annotation». In *Proceedings of Berliner XML Tage 2005*. Berlin.

Dorninger, M. 1997. *Gottfried von Viterbo. Ein Autor in der Umgebung der frühen Staufer*. Stuttgart: Verlag Hans-Dieter Heinz - Akademischer Verlag Stuttgart.

Foerster, T. 2015. *Godfrey of Viterbo and his Readers: Imperial Tradition and Universal History in Late Medieval Europe*. Farnham: Ashgate.

McGann, J. 2004. «Marking Texts of Many Dimensions». In *A Companion to Digital Humanities*, a cura di S. Schreibman, R. Siemens e J. Unsworth, 198–217. doi:http://doi.wiley.com/10.1002/9780470999875.ch16. http://www.digitalhumanities.org/companion.

Monella, P. 2006. «Edizioni critiche digitali, XML e letterature classiche». http://www1.unipa.it/paolo.monella/cattolica/2006/index.html#b1.

Rosselli Del Turco, R., et al. 2014b. «Edition Visualization Technology: A Simple Tool to Visualize TEI-based Digital Editions». *Journal of the Text Encoding Initiative*. http://jtei.revues.org/1077.

Viterbo, G. da. 1872. *Opera*. A cura di G. Waitz. Monumenta Germaniae Historica, Scriptores 22. Hannover: Impensis Bibliopolii Aulici Hahnian.

La Bibliotheca bibliothecarum manuscriptorum nova electronica (IRHT-CNRS, Paris)

Jérémy Delmulle, IRHT-CNRS, jeremy.delmulle@gmail.com
Bénédicte Giffard, IRHT-CNRS, be.giffard@gmail.com
Frédéric Duplessis, IRHT-CNRS, fredericduplessis@hotmail.fr

La *Bibliotheca bibliothecarum manuscriptorum nova*, pubblicata nel 1738 da Bernard de Montfaucon, benedettino di Saint-Maur, offriva per la prima volta al pubblico il testo di oltre 250 cataloghi ed inventari di biblioteche, soprattutto francesi e italiane, che menzionavano decine di migliaia di volumi. Questa pubblicazione imponente poggiava su circa 1000 documenti che sono appunti e memorie di eruditi, raccolti a Saint-Germain-des-Prés nel XVII e XVIII secolo. La maggior parte di questi documenti sono inediti.

Il progetto di edizione digitale della BBMN e di una parte delle sue fonti manoscritte è stato inaugurato nel 2014 dall'Equipex Biblissima. L'obiettivo è quello di fornire, per ogni elenco di libri, un'introduzione scientifica, un'edizione critica e un apparato di note che si rivolgono agli storici delle biblioteche e anche agli specialisti di manoscritti.

L'uso della TEI, che si è rivelato indispensabile per rispondere ai bisogni di un corpus così importante (già più di 88000 items), ha permesso un approccio generale, collaborativo, evolutivo e sostenibile di un materiale altrimenti impossibile da gestire, in tempi ragionevoli, da parte di un singolo ricercatore. L'obiettivo, a lungo termine, è di fare in modo che in ogni inventario gli autori, le opere, le date, i luoghi, le addizioni e le espunzioni siano codificati in linguaggio TEI. Inoltre, il ricorso a tale linguaggio permette di creare dei collegamenti tra gli items di diversi inventari che descrivono uno stesso manoscritto e di compararli facilmente.

Data l'ampiezza e l'eterogeneità dei dati, abbiamo deciso di ricorrere a una codifica a geometria variabile: ciò permette di dare accesso, in maniera relativamente rapida, a dei sotto-corpus senza perdere la coerenza della codifica dell'insieme dei sotto-corpus. I livelli di marcatura saranno tre: 1. un primo livello corrispondente ad una trascrizione del documento senza nessuna annotazione scientifica; 2. un secondo livello che presenta una marcatura parziale (ad esempio, tutti gli elementi 'title' e/o 'name', ma senza la fase ulteriore d'indicizzazione); 3. infine, un terzo livello di marcatura corrispondente a quello di un'edizione critica conforme alle esigenze editoriali più diffuse in ambito specialistico.

Questa scelta tecnica fa emergere diverse questioni, il cui interesse oltrepassa quello del presente progetto (trattamento semiautomatico dei dati, eterogeneità del corpus, posizionamento, ecc.). Per esempio, questo progetto editoriale partecipa all'arrichimento di thesauri collettivi contenenti dati sugli autori e sulle opere citati negli inventari (tutti i thesauri forniscono, per ogni lemma, delle informazioni scientifiche nonché gli identificanti unici e i collegamenti internet con i database di VIAF e della BNF).

Questo progetto è costruito su un dialogo permanente tra filologia e informatica: i bisogni scientifici modellano lo strumento digitale (ambiente XML_Mind) che, dal canto suo, permette al ricercatore di trattare con maggiore finezza il suo corpus, facilitando la navigazione in seno ai differenti elementi di questo corpus.

Vorremmo presentare, con un poster, l'interfaccia TEI elaborata per l'edizione digitale e le diverse tappe di questo progetto, dalla redazione originale (manoscritta) degli inventari fino alla loro trascrizione e alla loro stampa (spesso incompleta) nei volumi della *Bibliotheca bibliothecarum*, e infine la loro edizione digitale, accompagnata da apparato critico, pubblicata sul sito Inventaires mauristes, nella *Collection des inventaires anciens*, che sarà disponibile a partire da settembre 2016.

Attraverso l'edizione digitale, i diversi collegamenti ipertestuali fanno di questo sito un portale verso altre risorse disponibili online (digitalizzazioni, cataloghi, articoli scientifici).

Bibliografia

Delmulle, J. 2017. «Pour la reconstitution de la bibliothèque de l'abbaye de Saint-Martin de Sées. Des archives mauristes aux manuscrits retrouvés». In *Paper presented at the Journées d'études "Autour de la Bibliotheca bibliothecarum manuscriptorum nova. Bernard de Montfaucon, les mauristes et les bibliothèques de manuscrits médiévaux", Paris, January 14-15 (to be published).*

Duplessis, F. 2017. «Les méthodes de travail de dom Julien Bellaise. Étude des catalogues du ms. Paris, BnF, lat. 13073». In *Paper presented at the Journées d'études "Autour de la Bibliotheca bibliothecarum manuscriptorum nova. Bernard de Montfaucon, les mauristes et les bibliothèques de manuscrits médiévaux", Paris, January 14-15 (to be published).*

Gasnault, P. 1998. «Bernard de Montfaucon codicologue. La Bibliotheca bibliothecarum manuscriptorum nova». In *Dom Bernard de Montfaucon. Actes du colloque de Carcassonne (octobre 1996)*, a cura di D.-O. Hurel e R. Rogé, 2:1–21. Reprint in Gasnault, Pierre. 1999. 243–263. Carcassonne–Saint-Wandrille: Edition de Fontenelle.

— . 1999. *L'érudition mauriste à Saint-Germain-des-Prés*. Paris: Institut d'Études Augustiniennes.

Giffard, B. 2017. «'Lost in transcription' : les silences de la BBMN.» In *Paper presented at the Journées d'études "Autour de la Bibliotheca bibliothecarum manuscriptorum nova. Bernard de Montfaucon, les mauristes et les bibliothèques de manuscrits médiévaux", Paris, January 14-15 (to be published).*

Montfaucon, B. de. 1739. *Bibliotheca bibliothecarum manuscriptorum nova : Ubi, quae innumeris pene Manuscriptorum Bibliothecis continentur, ad quodvis Literaturae genus spectantia et notatu digna, describuntur et indicantur*. Paris: Briasson.

Petitmengin, P. 1998. «Montfaucon, dom Le Maître et la Bibliotheca Bibliothecarum». In *Du copiste au collectionneur. Mélanges d'histoire des textes et des bibliothèques en l'honneur d'André Vernet*, a cura di D. Nebbiai-Dalla Guarda e J.-F. Genest, 537–58. Turnhout: Brepols.

Edizione diplomatica digitale di un corpus quadrilingue: il progetto ANR TRANSSCRIPT

Laura Gili-Thébaudeau, Université de Lorraine, laura.gili@univ-lorraine.fr

1 Il progetto ANR TRANSSCRIPT Writing and Governance. Cultural Transfers between France and the Empire (13th-16th centuries)

Il progetto TRANSSCRIPT - http://www.agence-nationale-recherche.fr/?Projet=ANR-14-CE31-0022, che è finanziato dall' "Agence Nationale de la Recherche" (ANR) e dal suo omologo lussemburghese il "Fonds National de la Recherche" (FNR), è guidato da Isabelle Guyot Bachy e Michel Margue.

TRANSSCRIPT nasce dalla collaborazione di due laboratori di ricerca universitaria: l'*Atelier Diplomatique* del CRULH (*Centre de Recherche universitaire lorrain d'Histoire*) dell'Università della Lorena, erede del *Centre de Médiévistique Jean-Schneider*, centro rinomato per la ricerca nelle Digital Humanities, e l'*Institut d'histoire* dell'Università del Lussemburgo, che ha importanti competenze nell'analisi storica della Lotaringia medioevale, la diplomatica e la cartografia.

L'équipe dell'Università della Lorena è composta da Christelle Balouzat-Loubet, da Jean-Christophe Blanchard e dalla redattrice di queste note. Dell'équipe dell'Università di Lussemburgo fanno parte Michel Pauly, Hérold Pettiau, Timothy Salemme e il dottorando Jean-Daniel Mugeot. Le due équipe possono contare sull'aiuto di alcuni studenti universitari (come Denis Remy e Francis Carton). Ultimamente (fine 2016 – inizio 2017), all'équipe lussemburghese si sono aggiunti al progetto i ricercatori Anna Jagosova e David Kirt.

Questo progetto vorrebbe avvalersi delle possibilità offerte dall'informatica applicata alle fonti storiche e di uno strumento affidabile e duraturo (perenne, come si usa dire) per mettere in rete la propria produzione scientifica, soprattutto quella relativa alla trascrizione dei documenti (*Urkunde*, secondo la diplomatica tedesca) dei secoli XIII-XIV.

Al di là e grazie alla messa in rete sul Web, il progetto mira a studiare le pratiche della scrittura amministrativa della tarda epoca medioevale attraverso l'unione di due campi di ricerca: la gestione politica degli «stati regionali medioevali» e i transferts culturali (da cui deriva la prima parte dell'acronimo TRANSSCRIPT, TRANS) fra regioni che si trovano sulle due parti di una frontiera linguistica.

Internet permette, da una parte, di comunicare più rapidamente fra le due componenti dell'équipe del progetto TRANSSCRIPT e, dall'altra, di condividere con la comunità

scientifica alcuni dei risultati raggiunti. I documenti trascritti saranno interrogabili on line in due modi:

1. ricerca rapida: grazie ad una ricerca di dati nel database,

2. ricerca analitica: grazie alla visualizzazione della trascrizione integrale del documento.

La sfida di questo progetto è di trascrivere e pubblicare un numero abbastanza elevato di documenti (a oggi ne stimiamo circa tremila) e di permettere una vasta ricerca sui personaggi storici e sui luoghi, anche se citati in documenti di lingue diverse. Le lingue dei documenti di questo progetto "transfrontaliero" sono, infatti, quattro: latino, francese, tedesco e olandese.

All'inizio si era pensato di utilizzare l'XML/TEI facendo una codifica grazie all'editore di testi Oxygen®(https://www.oxygenxml.com/xml_editor.html). Si era pensato anche di far marcare le trascrizioni direttamente dai ricercatori impegnati nella ricerca scientifica. La riflessione di questi ultimi avrebbe permesso di far emergere meglio i dati per la ricerca sul Web. Man mano che si procede nella marcatura, per esempio, il ricercatore, dato che utilizza lui stesso lo strumento informatico, può scoprire tutte le possibilità per mettere in luce alcuni aspetti dei documenti, come, per esempio, la presenza di note dorsali o delle note di cancelleria. Questo principio, inoltre, è di grande utilità anche per fare evolvere, secondo le proprie necessità, gli elementi TEI (http://www.tei-c.org/index.xml).

L'idea era quella di seguire l'esempio dei lavori in rete dell'*École nationale des chartes* (Énc) – in particolare, l'edizione del *Cartulaire blanc* di *Saint-Denis*, sotto la direzione scientifica d'Olivier Guyotjeannin e di quella tecnica di Florence Clavaud: http://saint-denis.enc.sorbonne.fr e quella del registro del notaio Pierre Christofle a cura di Kouki Fianu dell'Università d'Ottawa la quale si è avvalsa, anche lei, della perizia di Florence Clavaud: http://elec.enc.sorbonne.fr/christofle). In questi lavori, infatti, il connubio fra informatici e storici è quasi perfetto, dato che gli storici e gli informatici lavorano insieme o che gli "storici" sono anche dei specialisti di XML/TEI: la codifica in XML/TEI è spesso effettuata da studiosi di storia, studenti di livello Master o anche di livello dottorato (la redattrice di queste note ha partecipato a tre progetti on-line dell'École nationale des chartes).

Ecco un esempio di codifica utilizzando l'editore Oxygen®per TRANSSCRIPT:

Dopo un anno di riflessione, l'équipe di TRANSSCRIPT ha rinunciato all'idea di una siffatta codifica in XML/TEI, perché la mole di documentazione non avrebbe permesso di arrivare a un risultato soddisfacente in soli quattro anni (la durata del finanziamento dell'ANR).

2 La piattaforma TELMA (Traitement électronique des manuscrits et des archives)

L'équipe di TRANSSCRIPT si è dunque rivolta al direttore scientifico di TELMA - http://www.cn-telma.fr, Paul Bertrand.

```
<p><lb n="01"/> Nos <persName id="Agnès, comtesse de Deux-
Ponts">Agnes</persName>, comitissa <placeName>Gemini Pontis</placeName>,
collateralis nobilis viri domini <persName id="Evrard, comte de Deux-
Ponts">Everhardi</persName>, comitis <placeName>Gemi<expan>ni</expan>
Pontis</placeName>, ad <lb n="02"/> noticiam omnium volumus pervenire quod nos,
nulla vi metus coacta, renunciavimus et renu<expan>n</expan>ciamus libere et
expresse <lb n="03"/> omni juri quod nobis in <placeName>castro de
Morispech</placeName> et in suis appendiciis universis r<expan>ati</expan>one
dotis seu dotalicii sive donationis <lb n="04"/> propter nuptias facte nobis ab
antedicto<ref type="note" n="1" target="#A-n-1"/> marito nostro de consensu illustris
domini nostri <persName id="Ferry III, duc de Lorraine">Friderici</persName>, ducis
<placeName id="Lorraine">Lothoringie</placeName> et marchi-<lb n="05"/>-onis,
cum <term key="fief">feodum</term> sunt ipsa bona competebat seu attingebat.[...] </p>
```

Figura 1: Nancy, Archives départementales de Meurthe-et-Moselle, B 568 n° 11

Figura 2: Home page del futuro database del progetto TRANSSCRIPT sul portale TELMA

L'obiettivo di questa infrastruttura corrisponde perfettamente ai fini del progetto: TELMA permette la pubblicazione elettronica e la messa a disposizione per la comunità scientifica dei corpora di fonti primarie e di strumenti di ricerca necessari alla loro gestione. È, inoltre, una piattaforma di servizi (dall'aiuto e consigli fino alla realizzazione di un intero corpus) e di divulgazione (seguendo il rispetto delle norme e degli standard, l'interoperabilità fra i corpora e l'archiviazione perenne dei dati).

Questa piattaforma edita dall'IRHT (*Institut de recherche et d'histoire des textes*) - CNRS, e sostenuta dal TGE Adonis (`http://renatis.cnrs.fr/spip.php?article219`), ospitando già altri progetti di edizione di atti medievali (come Chartae Burgundiae Medii Aevi - CBMA, *Chartes originales antérieures à 1121 conservées en France, Chartes originales (1121-1200) conservées en France, Scripta, Chartae Galliae*), ha dei notevoli vantaggi per l'équipe

TRANSSCRIPT e per la condivisione e gli sviluppi della ricerca:

- La rappresentazione dei dati. La piattaforma prevede un modo di ricerca abbastanza semplice per l'utente e, da un database all'altro, l'uniformità dei differenti database lascia il tempo all'utente di concentrarsi sulla ricerca, più che sull'aspetto tecnico (conoscenza del sito a priori).
- La ricerca sui corpora. A breve è prevista la possibilità di fare una ricerca su tutti i corpora inseriti sulla piattaforma e non solo una ricerca per singolo corpus.
- La perennità. La conservazione dei dati dovrebbe essere a lungo termine, perché il sovvenzionamento è pubblico.
- Il diritto d'autore. Questo diritto è ancora molto discusso, ma le comunità scientifiche francesi non cessano di pubblicare sul Web con licenze aperte. In questo modo, i ricercatori hanno visibilità per le loro ricerche e possono ottenere più facilmente fondi per proseguirle.

Occorre ricordare che l'IRHT-CNRS sviluppa dei progetti di edizione o di ricerca e che, parallelamente, migliora gli strumenti informatici. L'istituto, nelle sue produzioni informatiche, accoglie altresì, come detto, altri studi.

La raccolta di più basi di edizione, permette all'Istituto di ricerca del CNRS in storia di avere un gran numero di database da cui attingere informazioni storiche importanti.

Fra le strutture che collaborano con il CNRS e che permettono di condividere i dati e di avere una reale possibilità di interoperabilità, ricordiamo in questa sede il consorzio COSME (https://cosme.hypotheses.org/a-propos) in associazione con il TGIR Huma-Num del CNRS (http://www.huma-num.fr). Questo consorzio, che è nato nel 2013, ha per vocazione di recensire i differenti database o le edizioni elettroniche e di metterli in relazione fra loro.

3 Stato dell'arte

Alla fine del primo anno del progetto, abbiamo effettuato, tra l'Università della Lorena e quella del Lussemburgo, un migliaio di trascrizioni e, contemporaneamente, continuiamo l'indagine negli archivi per individuare i documenti da pubblicare. Per quel che concerne l'Università della Lorena, studi e progetti precedenti hanno permesso di avere già un buon numero di foto e, quindi, di trascrizioni. L'Università del Lussemburgo, invece, a parte la riutilizzazione di un buon numero di edizioni cartacee di documenti, sta indagando approfonditamente negli archivi.

Per quanto riguarda la messa in rete della documentazione, grazie a TELMA, se è vero che siamo stati obbligati a rinunciare a utilizzare Oxygen®, non siamo stati obbligati a rinunciare alla codifica in XML/TEI degli atti. Grazie al formulario progettato (gestito, al momento, da Cyril Masset), infatti, l'XML/TEI è alla base del lavoro di inserimento dei dati e delle trascrizioni anche se non si vedono i codici. Questo permette a una persona non esperta di XML di inserire agevolmente tutti i dati che saranno così disponibili per il Web.

Nella seconda parte del progetto, stiamo valutando la possibilità d'indicizzare i nomi propri o di luogo grazie all'uso di tags TEI come <persName> e <placeName> da inserire nel testo del formulario in uso (dunque, ancora una volta, senza passare per una marcatura XML/TEI completa dell'atto).

Negli elementi TEI sarà inserito un attributo @id per identificare rapidamente tutti i nomi di persona o di luogo quale che sia la lingua utilizzata. Infatti, il nome (di persona o di luogo) sarà direttamente nella lingua d'uso corrente. Per esempio, se il luogo si trova oggi in Francia, nell'indice il nome sarà in francese di uso contemporaneo, anziché in latino, anche se nel documento la lingua utilizzata è quest'ultima.

TELMA, inoltre, in collaborazione con l'équipe di BIBLISSIMA (http://www.biblissima-condorcet.fr) intende sfruttare l'interoperabilità delle fonti che sono già pubblicate e che lo saranno sulla piattaforma.
La tecnica di dare un numero univoco (*identifiant*, in francese) a personaggi storici e a luoghi, permette di rendere possibile questa interoperabilità. Per i nomi di persona sarà possibile far ricorso alla base VIAF (https://viaf.org) e per i luoghi geografici alla base Geonames (http://www.geonames.org), per esempio.

Sembra opportuno sottolineare che nel database di TRANSSCRIPT - TELMA, saranno altresì presenti immagini (grazie alle convenzioni stipulate con gli archivi dipartimentali in Francia e con gli archivi nazionali del Lussemburgo e del Belgio) e link verso un database di sigilli (SIGILLA: http://www.sigilla.org).

4 Conclusioni preliminari

Dalla ricerca e messa a disposizione di dati del progetto TRANSSCRIPT, ci aspettiamo di rinnovare l'analisi storica dello sviluppo dei principati transfrontalieri come il ducato di Lorena e la contea e il ducato del Lussemburgo.

Non è, peraltro, un caso che questa ricerca abbia luogo in questo momento: il mezzo informatico permette, finalmente, di incrociare tutti i dati in maniera facile e sicura ottimizzando i tempi di ricerca e di pubblicazione dei dati.

Bibliografia

Ansani, M. 2003. «Panorama de l'utilisation de l'informatique par les diplomatistes italiens». *Le médiéviste et l'ordinateur* 42. http://lemo.irht.cnrs.fr/42/mo42_10.htm.

Berra, A. 2012. «Faire des humanités numériques». In *READ/WRITE BOOK 2: Une introduction aux humanités numériques*, a cura di P. Mounier, 25–43. Marseille: Open Editions. doi:10.4000/books.oep.226. http://books.openedition.org/oep/238.

Burnard, L. 2012. «Du literary and linguistic computing aux digital humanities: retour sur 40 ans de relations entre sciences humaines et informatique». In *READ/WRITE BOOK 2: Une introduction aux humanités numériques*, a cura di P. Mounier, 45–58. OpenEdition Press. http://books.openedition.org/oep/242.

Muzerelle, D. 2011. «À la recherche d'algorithmes experts en écritures médiévales». *Gazette du livre médiéval* numero speciale, 56-57:5–20.

Orlandi, T. 2012. «Lo statuto dell'informatica umanistica. Per una storia dell'informatica umanistica». In *Dall'Informatica umanistica alle culture digitali. Atti del convegno di studi in Memoria di Giuseppe Gigliozzi, Roma, 27-28 ottobre 2011*, a cura di F. Ciotti e G. Crupi. Roma: DigiLab. http://digilab-epub.uniroma1.it/index.php/Quaderni_DigiLab/article/view/18/16.

Pratesi, A. 1977. «Limiti e difficoltà dell'uso dell'informatica per lo studio della forma diplomatica e giuridica dei documenti medievali». In *Informatique et histoire médiévale. Atti del convegno di studi, Roma, 20-22 maggio 1975*, a cura di L. Fossier, A. Vauchez e C. Violante, 187–190. Roma: École Française de Rome.

Rosselli Del Turco, R. 2009b. «Il progetto Vercelli Book digitale: codifica e visualizzazione di un'edizione diplomatica grazie alle norme TEI P5». In *Medieval texts – Contemporary media: The art and science of editing in the digital age*, a cura di M. G. Saibene e M. Buzzoni. Pavia: Ibis. http://www.academia.edu/11202544/Medieval_texts_-_Contemporary_media._The_art_and_science_of_editing_in_the_digital_age.

Vogeler, G. 2005. «Towards a standard of encoding medieval charters with XML». In *Literary and linguistic computing. Atti del congresso, Newcastle upon Tyne, Gran Bretagna 5 settembre 2004*, 20:269–280. 3. Oxford University Press. https://www.researchgate.net/publication/31201936_Towards_a_Standard_of_Encoding_Medieval_Charters_with_XML.

Edizione digitale del codice Napoli, Biblioteca Nazionale, MS xiii.b.29.

Omar Khalaf, Università degli Studi dell'Insubria, omar.hashem@uninsubria.it
Raffaele Cioffi, Università degli Studi di Torino, raffaelecioffi08@gmail.it

Il poster mira a promuovere il progetto di edizione digitale in formato XML-TEI del manoscritto Napoli, Biblioteca Nazionale, MS XIII.B.29, manufatto cartaceo di contenuto miscellaneo, risalente agli anni cinquanta del secolo XV.

Il codice, preziosissimo in quanto unico esempio di letteratura inglese tardomedievale conservato in Italia, contiene romanzi cavallereschi (*Sir Beuys of Hamptoun*, *Libeaus Desconus*, *Sir Isumbras*), un testimone del racconto di Griselda contenuto nei *Canterbury Tales* e un poema agiografico (*Seint Alex of Rome*) che conobbero ampia circolazione nell'Inghilterra del quattordicesimo e quindicesimo secolo, oltre a 137 ricette mediche. Fino ad oggi, l'interesse dimostrato dagli studiosi verso questo manoscritto è stato piuttosto scarso. Mentre i testimoni del *Seint Alex of Rome* e del *Lybeaus Desconus* sono stati infatti oggetto di edizioni specifiche (rispettivamente: Andreani 2009 e Mills 1969), il resto dei testi non sono stati dovutamente considerati dalla critica. Le ricette mediche e la *Griselda* sono state infatti edite in facsimile da Vallese (1940) ed accompagnate unicamenteda una breve introduzione e da un glossario. La versione acefala della *Griselda* è considerata anche nell'edizione di riferimento dei *Canterbury Tales* pubblicata da Manly e Rickert (1940), ma al solo fine della ricostruzione del testo. L'apparato critico di tale edizione non registra tutte quelle varianti che distinguono il testimone napoletano, ma solo quelle lezioni che hanno permesso ai due editori di stabilirne la tradizione testuale. Per quanto concerne poi i testimoni napoletani del *Sir Beuys* e del *Sir Isumbras*, va messo in evidenza come questi presentino interessanti varianti formali e contenutistiche, che sono state praticamente ignorate fino ad ora. Prendendo in considerazione il caso del Sir Beuys, le edizioni finora pubblicate, a partire da Kölbing (1885-1894) per arrivare a Burnley e Wiggins (2016) sono basate sul testo tradito nel manoscritto Auchinleck (Edinburgh, National Library of Scotland Adv MS 19.2.1), generalmente ritenuto il più autorevole. Gli stessi editori hanno messo in evidenza l'estrema variazione che caratterizza la trasmissione testuale del poema: l'apparato critico dell'edizione di Kölbing, peraltro, dimostra chiaramente le numerose peculiarità e varianti del testimone napoletano anche in confronto ai testi geneticamente più prossimi. Una situazione simile è rappresentata dal testimone del Sir Isumbras. Delle numerose edizioni del poema pubblicate finora, solamente quella di Schleich (1901), basata sul testimone tradito nel manoscritto Cambridge, Gonville and Caius College MS 175, conserva nell'apparato critico anche le varianti presenti nel testo napoletano. Secondo la sua ricostruzione, il testimone che l'editore ha preso a riferimento e il testimone napoletano apparterrebbero allo stesso ramo della tradizione, sebbene la variazione a livello testuale si evidenzi fin dall'incipit, dove il classico invito all'ascolto tipico dei romanzi cavallereschi viene sostituito da un'invocazione religiosa.

Di rilevanza non trascurabile appare anche la ricchezza di informazioni che possono esser tratte dal codice come oggetto, così come dal suo ricco apparato di marginalia, quasi certamente inseriti successivamente alla trascrizione dei testi: la presenza di varie iscrizioni in italiano da parte di alcuni membri della famiglia napoletana dei De Leonardis (ivi compreso un brevissimo componimento poetico finora inedito sull'imperatore Costantino) e, soprattutto, di un'illustrazione e di una nota che testimonierebbero la sua appartenenza a Tommaso Campanella, pongono il codice in una posizione di sicuro interesse storico. Tali elementi risultano di fondamentale pregnanza per la contestualizzazione della fruizione del codice nell'ambiente culturale napoletano del sedicesimo e diciassettesimo secolo: l'analisi del ricco apparato iconografico, delle note marginali, così come quello delle iscrizioni che si nascondono al di sotto di alcuni di essi, potrebbe inoltre potrebbe rivelare elementi importanti riguardo alla fruizione del manoscritto e, possibilmente, fornire alcuni indizi sui proprietari che si avvicendarono prima del suo arrivo alla Biblioteca Nazionale di Napoli.

L'obiettivo principale del progetto consisterà, dunque, nell'edizione digitale di un manoscritto finora praticamente quasi sconosciuto, ma di grande importanza storica e letteraria. Il progetto di tale edizione nasce dalla necessità di rendere conto sia degli elementi extratestuali contenuti nel codice (al fine di una sua possibile ricostruzione storica), sia delle caratteristiche di ciascun testo ivi contenuto: l'obiettivo sarà quello di fornirne una prima edizione di quel corpo di testi fino a questo momento trascurati dagli studi specialistici (come le ricette mediche), o di evidenziare i tratti di variazione che parte di essi denotano nei confronti del resto della tradizione.

L'edizione sarà image-based, così da fornire all'utente una riproduzione quanto più fedele possibile del manoscritto, delle sue peculiarità paleografiche e di quelle codicologiche. La codifica dei testi, condotta attraverso l'uso delle norme TEI-P5, sarà progettata in modo da organizzare l'edizione su differenti livelli.

La marcatura dei testi si articolerà poi su diversi livelli di codifica, in rapporto agli obiettivi ecdotici esposti sopra: alla trascrizione diplomatica in cui si riprodurranno tutte le caratteristiche strutturali e grafematiche, si aggiungerà una marcatura di livello interpretativo che segnalerà l'espansione di abbreviazioni, l'emendazione degli errori più evidenti compiuti dal copista e la presenza di varianti rispetto al testo standard proposto nelle principali edizioni esistenti.

Inoltre, l'utente avrà accesso al manoscritto attraverso un set di immagini digitali in alta definizione collegate al testo dell'edizione. Tale tipo di edizione fornirà non solo uno strumento completo di analisi sia filologica e letteraria del patrimonio tramandato dal codice, sia del manoscritto stesso come oggetto.

Di pregnanza primaria sarà la realizzazione di specifici fogli di stile per ognuno dei livelli di edizione del progetto. Tale operazione procederà in parallelo con quella di progressiva marcatura del contenuto del codice napoletano, in quanto vincolata non solo alle differenti necessità legate alla visualizzazione dell'edizione, ma anche alle possibili variazioni nell'uso dei vari marcatori intervenute nel corso della realizzazione dei file XML che compongono l'edizione digitale.

Per quanto concerne la visualizzazione e la navigazione all'interno delle differenti sezioni dell'edizione digitale, ci si affiderà al software EVT (Edition Visualization Technology), applicativo open source progettato e sviluppato in seno al progetto Vercelli Book Digitale e di cui è stata recentemente rilasciata la versione 1.0 (Rosselli Del Turco 2016): le funzio-

nalità di tale applicativo appaiono, infatti, del tutto rispondenti alle esigenze del progetto, oltre che implementabili in funzione delle diverse necessità che si dovessero presentare in corso d'opera.

Questo progetto di edizione digitale del codice napoletano si ripropone dunque di facilitare lo studio dei testi in esso contenuti nella loro specificità storica, in un ideale equilibrio tra lo studio del contesto culturale che diede vita al codice, l'indagine del manoscritto nella sua funzione di manufatto storico, l'analisi critica dei testi (in riferimento al resto della tradizione testuale) nel rispetto delle caratteristiche peculiari di ognuno di essi e, infine, la ricostruzione dell'uso che si fece del manoscritto stesso nel corso della storia. Uno degli obiettivi ultimi di tale ricerca, perciò, sarà quello di fornire tutti gli strumenti necessari per uno studio approfondito del codice non solo in quanto "contenitore di testi", ma anche come prodotto culturale di una determinata epoca e veicolo di materiale testuale che (caso unico nella letteratura inglese tardomedievale) entra nella penisola italiana e viene fruito da un pubblico certamente molto diverso da quello originario.

In definitiva, è nostra intenzione perseguire la realizzazione di una edizione che, associando vari livelli di codifica alla visualizzazione digitale del manoscritto, si rivolga ad un'ampia gamma di pubblico: dallo specialista interessato agli aspetti filologici e codicologici, allo studente universitario, all'appassionato di manoscritti e di letteratura medievale.

Il poster sarà organizzato sia attraverso il classico supporto cartaceo, sia con l'utilizzo di materiale informatico, utile alla visualizzazione del manoscritto e ad alcuni esempi preliminari di codifica e di visualizzazione. L'obiettivo è quello di attirare l'interesse degli studiosi presenti al convegno e di raccogliere consigli concernenti non solo le varie fasi della produzione dell'edizione digitale, ma anche il reperimento di fondi per lo sviluppo del progetto.

Bibliografia

Andreani, A. 2009. «Of Seint Alex of Rome. A Middle English Version of the Life of the Saint». *Linguistica e Filologia* 28:29–56.

Burnley, D., e A. Wiggins, cur. 2016. *The Auchinleck Manuscript*. Online Facsimile / Edition; site version: 1.1. Edinburgh: National Library of Scotland. http://auchinleck.nls.uk.

Kölbing, E., cur. 1885-1894. *Beues of Hamtoun*. London: Trübner.

Manly, J. M., e E. Rickert, cur. 1940. *The Text of the Canterbury Tales: Studied on the Basis of All Known Manuscripts*. Chicago: University of Chicago Press.

Mills, M., cur. 1969. *Libeaus Desconus*. London: Oxford University Press.

Rosselli Del Turco, R. 2016. «Edition Visualization Technology. Digital edition visualization software». Online. https://sourceforge.net/projects/evt-project.

Rosselli Del Turco, R., et al. 2014c. «Edition Visualization Technology: A Simple Tool to Visualize TEI-based Digital Editions». *Journal of the Text Encoding Initiative* 8. doi:10.4000/jtei.1077. http://jtei.revues.org/1077.

Schleich, G., cur. 1901. *Sir Ysumbras. Eine englische Romanze des 14. Jahrhunderts*. Berlin: Mayer & Müller.

Vallese, T., cur. 1939. *La Novella del chierico di Oxford: da un codice inglese inedito del 15. secolo di G. Chaucer, testo originale con trascrizione a fronte, varianti e introduzione*. Napoli: A.G.D.A.

— , cur. 1940. *Un ignoto ricettario medico inglese del 14. Secolo. Testo originale, trascrizione a fronte, introduzione, note e glossario*. Napoli: A.G.D.A.

Un secolo di scritture fuori dal canone. La bibliografia come modello informatico per la ricerca

Alessia Scacchi, University of Rome “Sapienza”, alessia.scacchi@uniroma1.it

1 Da SBN alla ricerca

La ricerca scientifica finalizzata alla costruzione di una bibliografia italiana delle scritture fuori dal canone del Novecento che si avvalga degli strumenti informatici, consta di un sistema di tabelle che ricostruiscono un quadro di presenze atipiche nel panorama letterario del Novecento. Le fonti utilizzate per tale ricerca sono molteplici: SBN, InternetCulturale, AIB, risorse di Istituti di Ricerca internazionali, oltre a Google-books per rintracciare informazioni o disambiguare dati controversi.

Seguendo lo schema della “scheda di autorità” proposto da SBN, per ogni autore / autrice il data base bibliografico contiene di prassi la dichiarazione delle fonti utilizzate sotto forma di link attivi, strumento essenziale per validare l’attendibilità delle fonti. La ricostruzione della bibliografia delle opere ha implicato inoltre la selezione non solamente della prima attestazione a stampa di ogni opera, ma delle nuove edizioni successive.

Le motivazioni profonde della costruzione e del web-design di una bibliografia che si concentri sulla ricostruzione di presenze letterarie fuori dal canone cela la tenace volontà di un gruppo di studiosi che da molti anni lavorano alla ricostruzione scientifica di un percorso letterario come quello del noto «doppio itinerario della scrittura» (Zancan 1998).

2 La base di dati

Il data base appare, quindi, la forma precipua di catalogazione e sistematizzazione di materiali raccolti nel tempo. Mentre mutavano i sistemi operativi e le architetture del web – trasformatosi in web 2.0 – la possibilità di creare ordine nel disordine creativo della rete restava una chimera. Oggi non lo è più. In un anno di ricerca, la bibliografia dei soggetti fuori dal canone può dirsi a buon punto, grazie alla sistematizzazione delle informazioni bibliografiche in tabelle poi utilizzate come base di dati per la ricerca.

In tal senso il lavoro di raccolta e riorganizzazione delle informazioni stanno facendo emergere un quadro digitale, una rete di senso che oltrepassa le differenze di genere, per proporre quella che Weinberger chiama una «eterogeneità interconnessa» di significati. La forma tabulare, inoltre, trasforma i dati in una «infrastruttura di significato» creando flussi di informazioni, di spiegazioni e di ipotesi (Weinberger 2010).

Lo scavo nella documentazione bibliografica su materiale che non è sistematizzato e che non presuppone la differenza di genere come valore definente e di distinzione si configura, quindi, come un lavoro ricostruttivo; un restauro con nuovi materiali recuperando ciò che di vecchio e dimenticato si aveva negli archivi nazionali e internazionali. Si è dato vita, perciò ad un'archeologia del documento librario che riscrive la mappa dell'editoria dall'Unità ad oggi, ridefinendo campi d'interesse e di forza disegnati non dalla differenza di genere come pregiudizio, quanto piuttosto dalla molteplicità dell'offerta editoriale del secolo breve, nella sua densità di significati.

3 La ricerca e le DH

La speranza è che il lavoro avviato da questa ricerca faccia emergere la bibliografia ottenuta, come un fenomeno di significato essa stessa, mentre evidenzi nel disordine della rete la coerenza ed unitarietà – oltre che scientificità – della modellizzazione informatica dei dati testuali. La costruzione della bibliografia in formato elettronico conduce a ripensare, infatti, la strutturazione dei dati bibliografici, al fine di ottenere il massimo grado di leggibilità e di chiarezza della cospicua mole catalografica, ma anche di garantire la massima permeabilità ed adattamento dei dati grezzi immessi nel database.

Lo strumento è in fase di elaborazione quindi si presenta come un caso di studio per la scelta della tecnologia più utile alla trasmissione di dati così rilevanti per la ricerca in ambito letterario. Gli studiosi e le studiose di Letteratura Italiana – sia in Italia, sia all'estero – potranno giovarsi della bibliografia come strumento di studio in maniera trasversale, ovvero attraversando la letteratura del Novecento con il filtro desiderato, dal punto di vista necessario per la ricerca di tipo: tematico, diacronico, sincronico, per parole chiave. Se lo strumento bibliografico è molto utile in ambito accademico, esso è una risorsa direi fondamentale per gli/le studiosi/e ed i/le docenti che lavorano alla predisposizione dei materiali per l'insegnamento della Letteratura Italiana negl'Istituti d'Istruzione superiore.

A questo proposito i dati saranno rielaborati e messi a disposizione della comunità scientifica internazionale per il tramite di un sito web dedicato all'interno del quale sarà possibile visionare i dati e fare ricerca. È per questo motivo che il lavoro che presentiamo implica il ripensamento del concetto stesso di bio-bibliografia delle opere del Novecento, attualizzandone scopi e architettura, riflettendo infine sulle specificità connesse alle Digital Humanities.

Bibliografia

Bolasco, S. 2013. *L'analisi automatica dei testi. Fare ricerca con il text mining*. Roma: Carocci.

Burdick, A. 2014. *Umanistica digitale*. Milano: Mondadori.

Ciotti, F. 2007. *Il testo e l'automa: Saggi di teoria e critica computazionale dei testi letterari*. Roma: Aracne.

Colombo, C. 2012. *Il web per le scienze umanistiche: strumenti e risorse*. Bologna: Universitas Studiorum.

Del Bono, G. 2000. *La bibliografia. Un'introduzione*. Roma: Carocci.

Fabrini, P. 2002. *La rete per gli antichisti. Strumenti e strategie per la ricerca on-line*. Pisa: Servizio Editoriale Universitario di Pisa.

Foucault, M. 2009. *La prosa del mondo*. Milano: BUR.

Gigliozzi, G. 2003. «L'informatica, la didattica e il grillo parlante». In *Informatica umanistica. Dalla ricerca all'insegnamento*, a cura di D. Fiormonte, 111–118. Roma: Bulzoni.

Laudon, K. C., e J. P. Laudon. 2006. *Management dei sistemi informativi*. Milano: Pearson Italia.

McGann, J. 2014. *A New Republic of Letters: Memory and Scholarship in the Age of Digital Reproduction*. Cambridge, MA: Harvard University Press.

Metitieri, F., e R. Ridi. 2005. *Biblioteche in rete: Istruzioni per l'uso*. Bari: Laterza. http://www.laterza.it/bibliotecheinrete/index.htm.

Protetti, C. 2009. *La giosta multimediale 2.0. Editoria e nuovi media nell'era dei Social Network*. Roma: Aracne.

Roncaglia, G. 2010. *La quarta rivoluzione: sei lezioni sul futuro del libro*. Bari: Laterza.

Vivarelli, M. 2013. *Le dimensioni della bibliografia: Scrivere di libri al tempo della rete*. Roma: Carocci.

Weinberger, D. 2010. *Elogio del disordine: Le regole del nuovo mondo digitale*. Milano: BUR.

Whittaker, K. 2002. *Metodi e fonti per la valutazione sistematica dei documenti*. Manziana: Vecchiarelli.

Zancan, M. 1998. *Il doppio itinerario della scrittura. La donna nella tradizione letteraria italiana*. Torino: Einaudi.

Didattica e Disseminazione
Didactics and Dissemination
Talks

Nuove frontiere delle Digital Humanities in classe: esperienze dal campo

Mariantonietta Rizzetto, docente del Liceo Classico “Marco Polo” di Venezia, totirizzetto@gmail.com
Antonella Trevisiol, docente del Liceo Classico Marco Polo di Venezia, a.trevisiol@teletu.it
Dario Falcone, studente del Liceo Classico Marco Polo di Venezia, dariofalcone00@libero.it
Nicoletta Pilon, studentessa del Liceo Classico Marco Polo di Venezia, nicoletta.pilon@gmail.com
Paola Tomè, “MarieCurie” Fellow, University of Oxford, paola.tome64@gmail.com
Federico Boschetti, CNR-ILC, federico.boschetti@ilc.cnr.it
Edoardo Bighin, WikiMedia, edowiki@gmail.com
Uwe Springmann, Ludwig-Maximilians-Universität München Centrum für Informations-und Sprachverarbeitung, uwe@springmann.net

1 Introduzione generale

Quello che si propone di seguito è il progetto relativo alla dissemination che la REA (Research European Agency) richiede a ogni “Marie Curie” fellow. È parso opportuno, vista la lunga esperienza di insegnamento al liceo classico in Italia della dott.ssa Tomè, responsabile del progetto, portare in Europa la voce degli studenti e dei docenti di questo indirizzo di studio, ancora vivo nel nostro Paese, ma la cui sopravvivenza è per vari aspetti piuttosto compromessa. Il progetto di ricerca verte sul ritorno dello studio del Greco in Occidente nel XV secolo e si basa su fonti manoscritte e a stampa, tenendo come punti di riferimento Roma sotto il pontificato di Niccolò V e Venezia negli anni ’70 del XV secolo. Si colloca in un lasso cronologico più limitato di quello abbracciato all’interno del progetto di dissemination in sè stesso, contemplando tuttavia, nella sua sezione introduttiva, una presentazione generale del panorama degli studi di Greco in Italia e in Europa nel secolo XV.

Tra il 2016 e il 2017, con alcune scuole del veneto e in collaborazione con il CNR, varie istituzioni accademiche e biblioteche del territorio, sono in corso di svolgimento attività didattiche e di ricerca fra loro integrate allo scopo di mettere i ragazzi in contatto diretto con le antiche edizioni a stampa per lo studio del Greco e con le nuove frontiere dello studio delle lingue classiche con risorse e strumenti digitali.

In cosa consiste tutto questo? Anzitutto si tratta di una sorta di adozione virtuale di testi antichi a stampa in classe, approfondendo i dati concreti relativi alla circolazione a stampa

delle prime grammatiche, lessici greco-latini e manuali per lo studio del greco andati a stampa in Europa tra la metà del sec. XV e la morte di Andrea Asolano, episodio conclusivo dell'aurea parentesi manuziana a Venezia e in Europa.

In secondo luogo, con un rapido salto in avanti, si permette ai ragazzi di sperimentare quali siano le nuove frontiere in cui la filologia digitale si sta spingendo per quanto riguarda lo studio delle lingue classiche, con la collaborazione dell'AIUCD (Associazione Italiana di Informatica Umanistica e Cultura Digitale), di Musisque Deoque (gruppo di lavoro condotto da Paolo Mastandrea e Luigi Tessarolo in Italia, presso l'Università Ca' Foscari di Venezia), dell'Open Philology Project guidato da Gregory Crane a Lipsia, di Wikimedia Italia, del Ludwig-Maximilians-Universität München Centrum für Informations- und Sprachverarbeitung, e infine col fondamentale apporto del Laboratorio di Filologia Collaborativa e Cooperativa (CoPhiLab) dell'Isttituto di Linguistica Computazionale "A. Zampolli" del CNR di Pisa (CNR-ILC).

2 Illustrazione della scansione del progetto

Le scuole e i docenti interessati sono liberi di coinvolgersi nel progetto in quattro diversi livelli di profondità:

- **primo livello**: work-shop introduttivo (13 Novembre 2015, in collaborazione con la Biblioteca Nazionale Marciana) in cui sono state presentate in linee generali il significato del progetto (studiare il ritorno dello studio del greco in Occidente nel XV secolo) con la partecipazione di studiosi ed esperti del settore della filologia classica, della storia del libro a stampa e delle digital humanities;

- **secondo livello**: tra il 2016 e il 2017 ogni collega interessato coi ragazzi ha creato e continua a crea dei brevi testi, correlati tra di loro, su singoli concetti, parole-chiave, eventi, personaggi, opere, collegati al progetto, in lingua italiana e inglese (con l'aiuto dei colleghi di lingue);

- **terzo livello**: sempre tra 2016 e 2017: workshop nelle biblioteche: gruppi classe o singoli ragazzi più interessati si sono recati e si recano nelle biblioteche vicine (Marciana a Venezia, Universitaria a Padova, Comunale a Treviso) per prendere contatto diretto coi testi; in una fase successiva potranno inoltre curare la trascrizione e la traduzione in lingua inglese di alcuni documenti prefatori, da cui ricavare informazioni sulla selezione dei testi dati a stampa e su come / perché si studiasse il greco all'epoca;

- **quarto livello**: workshops con la collaborazione del mondo accademico e degli Istituti di Ricerca summenzionati in cui gruppi selezionati di ragazzi delle scuole coinvolte entrano in contatto con le nuove frontiere della Digital Philology e fanno piccole esperienze di digitazione e di annotazione morfosintattica di porzioni testuali tratte dalle antiche edizioni a stampa (per le classi del triennio) o dai manuali di greco e latino (per tutte le classi).

  In particolare quest'ultima attività (l'annotazione morfosintattica dei testi) costituisce uno strumento di potenziamento nell'apprendimento linguistico estremamente utile nella fase di avviamento degli studi per tutto il biennio, spendibile inoltre anche

nelle classi avanzate per i corsi di recupero. Su questo argomento è stato organizzato un corso di aggiornamento specifico, per permettere ai docenti di utilizzare in classe la piattaforma elaborata dal gruppo di ricercatori dell'Open Philology Project facenti capo a Gregory Crane, ora docente presso l'Università di Lipsia;

- **workshop finale**: è previsto il workshop conclusivo (maggio / giugno 2017) per presentare i lavori prodotti dalle classi e dai docenti, alla presenza di esperti di filologia classico-umanistica e di digital humanities.

3 Corso di aggiornamento per docenti sul treebanking

In collaborazione con il Perseus Project (Tufts University, Boston) e l'Open Philology Project (Università di Lipsia), è stato organizzato un workshop con i docenti per spiegare i principi della dependency grammar e le tecniche di annotazione sintattica di testi greci e latini tramite la piattaforma Perseids.

Dopo un inquadramento pedagogico-didattico per spiegare il differente paradigma di analisi sintattica rispetto alle metodologie tradizionali, si è chiesto ai docenti di annotare sintatticamente alcune favole di esopo e i loro risultati sono stati discussi con esperti dell'Università di Lipsia in videoconferenza e con esperti del CNR di Pisa in presenza.

Gli esperti hanno dato particolare spazio al dibattito con i docenti per individuare le ricadute sull'apprendimento della sintassi delle lingue antiche attraverso l'annotazione di risorse digitali. Si è giunti alla conclusione che questo tipo di attività promuove il ragionamento deduttivo in fase di creazione delle risorse annotate, grazie all'applicazione di regole a casi specifici, e il ragionamento induttivo in fase di interrogazione delle risorse create, grazie alla scoperta di pattern sintattici presenti nei testi, quali ad esempio l'ordine delle parole o la preferenza di un autore per certi stilemi rispetto ad altri.

4 OCR STORICO: Digitalizzare testi antichi a scuola: diario di un'esperienza didattica

Si ha l'intenzione di illustrare gli aspetti metodologici e discutere i primi risultati di un'attività didattica che ha coinvolto docenti e studenti di licei classici del Veneto per la digitalizzazione dell'Orthographia di Tortelli, un incunabulum del 1471.

L'obiettivo principale è far partecipare gli studenti, proporzionalmente alle loro capacità, alle fasi più avanzate del processo editoriale digitale, che parte dalle immagini digitali della fonte primaria, passa attraverso il riconoscimento ottico dei caratteri e arriva alla correzione manuale tramite piattaforma web. Inoltre, l'esposizione degli studenti al latino umanistico ha il duplice vantaggio di affrontare l'aspetto diacronico (latino del Quattrocento vs latino classico) e l'aspetto metalinguistico (uso del latino per descrivere fenomeni della lingua greca), che abitualmente sono lasciati in secondo piano, se non addirittura trascurati, nei programmi scolastici.

La fase di preparazione dei materiali ha richiesto il coordinamento fra l'Università di Oxford (scelta dei materiali), l'Università di Monaco di Baviera (digitalizzazione del testo

tramite OCR), il CNR di Pisa (conversione dei formati) e WikiMedia Italia (caricamento sulla piattaforma di WikiSource, destinata alle attività di correzione manuale dell'OCR).

Gli studenti hanno svolto l'attività in un laboratorio informatico dell'Università "Ca' Foscari" di Venezia, sotto la guida di un ricercatore in linguistica computazionale del CNR e di un docente di latino membro di WikiSource. Il fatto di contribuire al miglioramento di una risorsa digitale utile alla ricerca scientifica in ambito filologico, ha motivato gli studenti ad impegnarsi intensamente nelle operazioni di correzione. Dopo l'iniziale spaesamento dovuto alla natura del testo (mai visto in precedenza; non disponibile in traduzione) e alle consuetudini editoriali (uso di abbreviazioni ereditate dai manoscritti; segni di punteggiatura desueti), gli studenti hanno cominciato a prendere confidenza con l'attività di emendamento dell'OCR e di scioglimento delle abbreviazioni.

A ciascun allievo è stata affidata una pagina diversa da correggere e, successivamente, una pagina di un compagno da rivedere. Secondo le regole di WikiSource, infatti, le operazioni di correzione e di revisione delle correzioni non possono essere eseguite dalla stessa persona, per aumentare la possibilità di individuare sviste e per promuovere la discussione nei casi di difficile lettura o interpretazione, tramite il forum presente sulla piattaforma.

Grazie agli strumenti di allineamento in dotazione sulla piattaforma di WikiSource fra le diverse revisioni del testo, gli studenti hanno potuto controllare il lavoro svolto e osservare pattern d'errore ricorrenti sfuggiti alle fasi di correzione automatica, nel post-processing dell'OCR.

Il risultato materiale consiste nella correzione di più di un decimo del volume in circa due ore di attività. Tuttavia riteniamo che il risultato più importante sia il raggiungimento di un importante obiettivo educativo: far comprendere agli studenti come il loro piccolo sforzo legato ad una attività didattica non sia fine a se stesso ma contribuisca alla ricerca e come esista un mondo, come quello wiki, in cui una passione personale, sia durante che dopo il corso degli studi, può interagire ancora una volta con la ricerca.

Bibliografia

Celano, G. 2016. «Ancient Greek Dependency Treebank – Guidelines 2.0». Online. https://sites.google.com/site/giuseppegacelano/treebanking/agguidelines.

Piotrowski, M. 2012. *Natural Language Processing for Historical Texts*. 5:1–157. Synthesis Lectures on Human Language Technologies 2. Morgan & Claypool Publishers.

Rydberg-Cox, J. A. 2009b. «Digitizing Latin Incunabula: Challenges, Methods, and Possibilities». *Digital Humanities Quarterly* 3 (1).

Springmann, U., et al. 2014. «OCR of Historical Printings of Latin Texts: Problems, Prospects, Progress». In *Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage*, 71–75. DATeCH '14. Madrid, Spain: ACM. ISBN: 978-1-4503-2588-2. doi:10.1145/2595188.2595205. http://doi.acm.org/10.1145/2595188.2595205.

Tomè, P. 2012. «L'"Orthographia" di Giovanni Tortelli: studio dell'opera e delle fonti». Tesi di dott., Università "Ca' Foscari".

Un'esperienza di alternanza scuola-lavoro per una edizione digitale del corpus epigrafico del Museo Civico Castello Ursino

Serena Agodi, Liceo Artistico M.M. Lazzaro, Catania, serena.agodi@istruzione.it
Valentina Noto, Direzione Cultura e Turismo, Comune di Catania,
valentina.noto@comune.catania.it
Jonathan Prag, Faculty of Classics, University of Oxford,
jonathan.prag@merton.ox.ac.uk
Daria Spampinato, CNR - Istituto di Scienze e Tecnologie della Cognizione,
daria.spampinato@cnr.it

1 Introduzione

Questo contributo illustra una proficua esperienza di collaborazione tra diverse istituzioni e differenti professionalità, coinvolte nell'ambito dell'alternanza scuola-lavoro ai fini della digitalizzazione e valorizzazione del patrimonio epigrafico del Museo Civico Castello Ursino del Comune di Catania.

Il Museo possiede una prestigiosa collezione epigrafica, costituita da due raccolte settecentesche catanesi. Alcune iscrizioni sono esposte secondo i vecchi criteri di fruizione museale; la maggior parte è custodita in deposito. Uno dei principali obiettivi dell'esperienza è dunque far conoscere al pubblico fruitore ed alla comunità degli studiosi l'ingente patrimonio epigrafico del Castello attualmente non esposto al pubblico. L'idea si colloca nell'ambito della promozione del museo tramite il suo inserimento nel circuito di mostre internazionali, le iniziative di comunicazione web e l'implementazione delle modalità di fruizione integrata delle collezioni permanenti (attività di didattica museale; apertura del museo ad iniziative culturali, teatrali e musicali; l'adesione alla piattaforma multicanale open di audioguide izi.TRAVEL; la partecipazione al Manifesto e alle iniziative di Invasioni Digitali[1]).

L'attività di alternanza scuola-lavoro del liceo artistico "M.M. Lazzaro"[2] di Catania pres-

[1] Izi.TRAVEL https://izi.travel/it;Invasionidigitalihttp://www.invasionidigitali.it (visitati il 27 luglio 2016).

[2] Si ringrazia il Dirigente Scolastico prof. G. Sciuto per il sostegno fornito alle docenti referenti per l'alternanza D. Musumeci e C. Leonardi, promotrici con S. Agodi del progetto, e alle docenti tutor S. Agodi, C. De Grandi e P. Cantone.

so il Museo Civico, in applicazione della L. 107/2015[3], è stata incentrata sulla partecipazione diretta degli studenti alle attività di ricerca dei progetti EpiCUM (EPIgraphs of Castello Ursino Museum) e I.Sicily[4]. Il progetto di alternanza, con finalità orientativo-formative, è andato infatti oltre la semplice esperienza di didattica museale, per esplorare da vicino la complessità gestionale, amministrativa e soprattutto scientifica del museo ed avvicinare gli studenti ai diversi profili professionali museali, con particolare attenzione per l'ambito della catalogazione digitale.

La scuola dunque non è stata semplicemente interrogata sui suoi bisogni per adeguare ad essi la risposta del museo; in questo caso la scuola è stata coinvolta in maniera fattiva e propositiva, anche alla luce delle proficue esperienze di collaborazione scuola-museo in forma di tirocinio, svolte nei precedenti anni scolastici.

2 EpiCUM e I.Sicily

Lo scopo del progetto EpiCUM, cui gli studenti sono stati introdotti, è rendere disponibili in un unico portale web tutte le iscrizioni del Museo Civico di Catania con la codifica in EpiDoc TEI XML / Unicode[5], con la possibilità di essere interoperabili col progetto collaborativo EAGLE[6]. I dati vengono modellati e rilasciati in formato aperto, nel rispetto della normativa nazionale vigente, e in modalità Linked Data, per essere accessibili da qualsiasi applicazione indipendentemente da linguaggi di programmazione e tecnologie. Quasi tutto l'insieme delle epigrafi del Castello Ursino è stato oggetto dell'analisi dello studioso Kalle Korhonen; alcune iscrizioni, che risultano ancora inedite o edite solo localmente, saranno pubblicate secondo le specifiche del progetto I.Sicily, cui il lavoro di EpiCUM è correlato. EpiCUM si propone di presentare e rendere fruibile tutto il corpus epigrafico del Museo, completo di immagini, delle informazioni archeologiche ed epigrafiche sia sul testo che sul reperto e dell'apparato critico. I concetti presenti nei diversi archivi saranno collegati semanticamente tra di loro e disponibili in un unico contenitore omogeneo disponibile sul web secondo standard internazionali consolidati e con diverse modalità di accesso, visualizzazione e servizi e fruibile da diversi stakeholder.

Il progetto I.Sicily ha lo scopo di creare l'edizione digitale di tutte le iscrizioni su pietra della Sicilia antica (VII a.C. - VII d.C), secondo standard e vocabolari controllati internazionali (ad es. Pleiades[7] per posizioni geografiche e i lessici controllati di EAGLE[8] per i tipi di iscrizioni, i materiali e i supporti) al fine di consentire la creazione di Linked Open Data e con tutti i requisiti di una pubblicazione scientifica, tra cui l'editorial board. L'edizione prevede le trascrizioni del testo di ogni iscrizione e la raccolta delle informazioni complete sulla provenienza, ubicazione attuale e materiale del reperto epigrafico, insieme con imma-

[3]LEGGE 13 luglio 2015, n. 107 *Riforma del sistema nazionale di istruzione e formazione e delega per il riordino delle disposizioni legislative vigenti*.

[4]I.Sicily https://isicily.wordpress.com/su_isicily (visitato il 27 luglio 2016).

[5]EpiDoc - Epigraphic Documents in TEI XML http://epidoc.sourceforge.net (visitato il 27 luglio 2016).

[6]EAGLE - The Europeana network of Ancient Greek and Latin Epigraphy http://www.eagle-network.eu (visitato il 27 luglio 2016).

[7]Pleiades http://pleiades.stoa.org (visitato il 27 luglio 2016).

[8]Lessici controllati EAGLE http://www.eagle-network.eu/resources/vocabularies (visitato il 27 luglio 2016).

gini di alta qualità e con le traduzioni del testo in inglese e in italiano. Il progetto si avvale delle collaborazioni con i Musei e le Soprintendenze per i Beni Archeologici siciliani, per identificare e registrare tutte le iscrizioni conosciute e propone funzionalità di modifica online, per consentire a chiunque di contribuire con dati nuovi o rivisti, registrandone tutte le responsabilità autoriali. Le nuove edizioni e revisioni sono soggette a redazione peer review prima della pubblicazione. Anche il progetto I.Sicily utilizza EpiDoc TEI XML; tutti i dati vengono inseriti online con una licenza CC-BY-SA e possono essere scaricati gratuitamente in formato XML o CSV. I file saranno ospitati sin un eXsist database; file di immagine sono presentati utilizzando il visualizzatore IIP. Ogni singola iscrizione è collegata ad un identificativo univoco (URI) e il progetto collabora col progetto Trismegistos[9] per mantenere queste identificazioni uniche.

Entrambi i progetti di ricerca dunque si inseriscono programmaticamente all'interno dei più recenti orientamenti della comunità internazionale degli epigrafisti digitali, che ha adottato lo standard EpiDoc TEI XML per la codifica di documenti interoperabili.

3 Metodologie e organizzazione delle attività

All'interno dell'esperienza di alternanza scuola-lavoro le diverse professionalità coinvolte hanno contribuito ad integrare le varie metodologie di lavoro tramite diversi momenti di formazione: in aula con i docenti per la preparazione teorica anche tramite un uso flessibile delle metodologie della *flipped classroom*; a scuola con interventi di formazione specialistica con i responsabili dei progetti scientifici I.Sicily ed EpiCUM, infine *on the job*, seguiti sia dai tutor della scuola che da quelli dell'ente ospitante, attraverso una inusuale modalità di fruizione del museo, a contatto con gli uffici, negli ambienti della biblioteca e del magazzino, condividendone i ritmi lavorativi. Tutte le fasi sono state documentate e sono stati sfruttati i *social media* e in particolare Facebook, utilizzando la pagina del Museo Civico Castello Ursino[10] per la promozione e la divulgazione sia delle collezioni epigrafiche, sia dell'esperienza di alternanza scuola-lavoro.

Il percorso pratico dell'alternanza si è dispiegato nelle seguenti attività: riscontro inventariale guidato; pulitura a secco delle epigrafi; esame autoptico e verifica dei dati pubblicati da Korhonen; verifica ed integrazione dei dati metrici; documentazione grafica; documentazione fotografica; schedatura informatica del reperto, con campi raggruppati in 5 aree (*Location, Physical description, Epigraphic description, Apparatus and bibliography, General information*); documentazione giornaliera dell'attività (report, foto, post su facebook). Due classi dell'indirizzo Arti Figurative si sono occupate di tutte le attività ad esclusione della documentazione fotografica, curata da una terza classe della sezione di Video e Multimedialità, avviata alle tecniche della fotografia professionale mediante l'uso delle attrezzature in possesso del Liceo.

Come spesso accade in una esperienza di alternanza scuola-lavoro le metodologie utilizzate durante l'attività pratica sono state quelle tipiche della didattica laboratoriale, caratterizzata dal *cooperative learning*. Le classi sono state suddivise in gruppi secondo una organizzazione partecipata, che ha previsto la rotazione dei ruoli affinché ciascuno studen-

[9] Trismegistos (TM) http://www.trismegistos.org (visitato il 27 luglio 2016).

[10] Museo Civico Castello Ursino https://www.facebook.com/MuseoCivicoCastelloUrsino (visitato il 27 luglio 2016).

te potesse fare esperienza di ogni attività da svolgere e per favorire il *peer tutoring*. Alle attività sono stati intervallati momenti di approfondimento metodologico-tecnico sulle tematiche digitali e sullo specifico dell'epigrafia latina, sentiti come una esigenza emergente dalle domande degli studenti.

Si è sollecitato infine un quotidiano processo di autovalutazione del lavoro, svolto tramite schede strutturate, e di riflessione conclusiva sui punti di forza e di debolezza dell'esperienza, attraverso la somministrazione di questionari.

4 Obiettivi raggiunti

Rispetto a progetti pensati per Licei Classici e Linguistici la proposta di far avvicinare alle epigrafi di età romana gli studenti del Liceo Artistico, che non studiano la lingua latina, è una sperimentazione originale, che è apparsa fruttuosa per l'approccio legato alla sensibilità degli allievi verso l'oggetto, il prodotto artigianale-artistico, l'attenzione verso la cultura materiale e verso l'aspetto tecnico-creativo, legato alle competenze degli studenti del terzo anno dell'indirizzo artistico. Il contatto con il materiale archeologico ha fornito, nel percorso fin qui svolto, non solo una concreta esemplificazione della funzione delle fonti epigrafiche per la conoscenza del rapporto tra macrostoria e microstoria, ma soprattutto ha sviluppato negli studenti la consapevolezza di radici filogenetiche linguistiche e culturali dal valore identitario.

L'esperienza ha consentito un accostamento concreto alle diverse professionalità legate ai Beni Culturali, favorendo l'avvio di un orientamento consapevole, obiettivo primario dell'alternanza scuola-lavoro nei Licei. Il progetto ha inoltre dimostrato un notevole valore formativo, sia contribuendo a fornire nuove competenze e conoscenze specifiche nell'ambito della catalogazione digitalizzata, della fotografia scientifica, dell'archeologia e dell'epigrafia latina, sia favorendo, grazie alla messa in atto di una didattica partecipata, il potenziamento delle competenze trasversali cui la scuola mira innanzi tutto: la capacità di decodifica dei testi e di selezione dei dati; il *problem solving*; la capacità di lavorare in gruppo e di comunicare; lo spirito di collaborazione e di iniziativa; l'apprendimento tra pari; l'autonomia decisionale ed organizzativa; l'adattamento a luoghi e ritmi diversi; la capacità di concentrazione finalizzata all'apprendimento.

Il contributo degli studenti ha inoltre portato benefici alla ricerca quali: l'avvio del progetto EpiCUM di digitalizzazione del corpus epigrafico del Castello Ursino; la divulgazione delle tematiche di ricerca presso i giovanissimi; l'apertura del dibattito scientifico anche al di fuori della comunità di riferimento, per sollecitare condivisione e collaborazione; il potenziamento della continuità formativa tra scuola e università per contribuire all'orientamento consapevole dei giovani, sollecitando le competenze (grafiche, fotografiche, catalografiche) della formazione artistica; l'elaborazione di idee per una valorizzazione del patrimonio culturale, anche attraverso i nuovi strumenti di comunicazione (*social media*, video, immagini).

Infine, risultato non meno importante, il rapporto tra studenti, territorio e mondo della ricerca scientifica, della conservazione e della divulgazione dei Beni Culturali è apparso consolidato. Ma soprattutto l'analisi dell'autovalutazione degli studenti ha evidenziato, quale punto di forza dell'esperienza, una nuova consapevolezza del valore del patrimonio

culturale, di cui si sono sentiti parte attiva come studiosi e operatori, tanto da immaginare una nuova idea di fruizione anche performativa del museo.

Questa esperienza rappresenta un modo mirato di sperimentare l'alternanza scuola-lavoro, poiché favorisce un dialogo costruttivo tra il punto di vista scientifico e quello didattico. Infatti gli studenti hanno partecipato e contribuito direttamente alla pubblicazione di un catalogo nativo digitale: questo ha permesso loro non solo di vedere i risultati del lavoro svolto, ma di avere possibilità di interagire con i reperti, sia toccandoli con mano, sia analizzandone il contenuto, sia promuovendoli sui *social media*, sia nei modi che verranno esplorati con loro nelle prossime tappe del progetto.

5 Prossime tappe del progetto

Il progetto, di durata triennale, ha concluso quest'anno la prima tappa del suo percorso, che intende coinvolgere tutti gli indirizzi del Liceo, cioè non solo Arti Figurative e Video e Multimedialità, ad oggi interessate, ma anche Grafica, Architettura e Ambiente, Scenografia e Design.

La collaborazione tra il Museo, gli enti di ricerca e la scuola prevede infatti come ulteriore obiettivo, finalizzato alla valorizzazione delle raccolte museali e alla divulgazione dell'esperienza, la realizzazione di una mostra temporanea delle epigrafi studiate, che consenta una modalità espositiva incentrata sulla sensibilizzazione dell'utenza giovane alla fruizione consapevole del patrimonio culturale museale civico e proiettata verso la comunicazione integrata tramite l'uso della comunicazione web, del digitale e degli *open data*. La mostra presenterà diversi aspetti digitali, sia come mostra virtuale pubblicata sul sito web EpiCUM, sia attraverso l'utilizzo di strumenti digitali per ricostruzioni 3D di monumenti archeologici della città e di reperti, per video e *podcasting*, sia attraverso *digital storytelling*.

Bibliografia

Comoglio, M., cur. 1999. *Il cooperative learning. Strategie di sperimentazione*. Quaderni di Animazione e Formazione, Animazione Sociale. Torino: Edizioni Gruppo Abele.

Cummings, J., J. Prag e J. Chartrand. 2016. «Creating An EpiDoc Corpus for Ancient Sicily». In *Digital Humanities 2016: Conference Abstracts. Jagiellonian University and Pedagogical University, Kraków*, 165–167.

Korhonen, K. 2004. *Le Iscrizioni del Museo Civico di Catania*. Helsinki: Societas Scientiarum Fennica.

Prag, J., J. Chartrand e J. Cummings. 2016. «I.Sicily: an epidoc corpus for ancient Sicily.» In *Digital and Traditional Epigraphy in Context. Proceedings of the Second EAGLE International Conference. Rome 27-29 January 2016*, a cura di S. Orlandi et al., 39–52. Rome: Sapienza Università Editrice.

Reali, M., e G. Turazza. 2015. «Parole di pietra: epigrafia e didattica del Latino». In *Prospettive per l'insegnamento del latino*, a cura di A. Balbo e M. Ricucci. I Quaderni della ricerca 16. Torino.

Spampinato, D. 2014. «BILG - Inscriptiones Graecae et Latinae Bruttiorum. A digital corpus in EpiDoc of Roman Law inscriptions». In *Posters, Projects, Demos Track at EAGLE 2014 International Conference, 29-30 September / 1 October 2014, Paris, France*. http://www.eagle-network.eu/about/events/eagle2014/digital-poster-exhibition/?pid=43.

Homeric Greek WordNet: costruire una risorsa lessico-semantica fra ricerca e didattica

Antonio Stanzione, Lab. di Antropologia del Mondo Antico (LAMA) dell'Università di Pisa
Giulia Re, Lab. di Antropologia del Mondo Antico (LAMA) dell'Università di Pisa
Gloria Mugelli, Lab. di Antropologia del Mondo Antico (LAMA) dell'Università di Pisa
Andrea Taddei, Lab. di Antropologia del Mondo Antico (LAMA) dell'Università di Pisa
Federico Boschetti, CNR-ILC, federico.boschetti@ilc.cnr.it
Riccardo Del Gratta, CNR-ILC, riccardo.delgratta@ilc.cnr.it

1 Introduzione

Questo lavoro illustra i primi risultati dei tirocini curricolari concordati fra il Laboratorio di Filologia Collaborativa e Cooperativa (CoPhiLab) del CNR-ILC di Pisa e il Laboratorio di Antropologia del Mondo Antico (LAMA) dell'Università di Pisa.

Da alcuni anni il CoPhiLab, insieme all'Alpheios Project, cura la risorsa lessico-semantica AncientGreekWordNet (AGWN), descritta in Bizzoni et al. (2014). Le wordnets multilingui, derivate da Princeton WordNet (Fellbaum 1998), permettono di raggruppare in un insieme (detto synset) parole con lo stesso senso attorno a nodi concettuali descritti da una glossa (ad es. [bat, chiropteran]: nocturnal mouselike mammal...) e consentono di distribuire i diversi sensi di una stessa parola polisemica fra diversi synsets (ad es. bat[1]: nocturnal mouselike mammal..., bat[2]: an implement used for hitting the ball in various games).

Mentre Latin WordNet, realizzata da Stefano Minozzi presso l'Università di Verona in collaborazione con l'FBK è una risorsa matura e validata manualmente, AGWN è ancora ad uno stadio aurorale e necessita di validazione. Per questo motivo si è deciso di iniziare la correzione manuale da un sottoinsieme coerente di AGWN costituito dai synsets correlati al lessico omerico, Homeric Greek WordNet (HGWN).

In effetti, una risorsa digitale sulla semantica omerica (e in particolare della sinonimia), può avere molteplici impieghi in ambito filologico, ad esempio per la valutazione di varianti e per lo studio dei sistemi e dei tipi formulari (mentre per lo studio delle formule omeriche in senso basta combinare l'analisi metrica con la lemmatizzazione: cf. Pavese e Boschetti 2005). La creazione di HGWN permette inoltre di gettare le basi per una risorsa lessico-semantica diacronica, dove sia possibile seguire l'evoluzione dei signifcati dei termini da Omero agli stadi successivi della lingua greca.

2 Metodo

Ciascuno studente che partecipa allo stage concorda con il suo docente universitario del LAMA e con i ricercatori del CoPhiLab un campo semantico sufficientemente rappresentato dal lessico omerico, su cui basare la propria attività di 225 ore presso il CNR-ILC. Dopo la prima fase di ricerca bibliografica, gli stagisti procedono alla correzione dei synsets tramite l'interfaccia web di HGWN. Dalla vista sintetica sui sensi pertinenti al testo omerico dei termini da validare (si veda Fig. 1), si accede alla scheda di validazione (Fig. 2). Oltre a dizionari bilingui online, gli studenti hanno a disposizione lessici specifici (Ebeling 1880–1885; Cunliffe 2012; etc.), monografie di carattere generale (ad es. Hernandez 1997) e opere su campi semantici specifici.

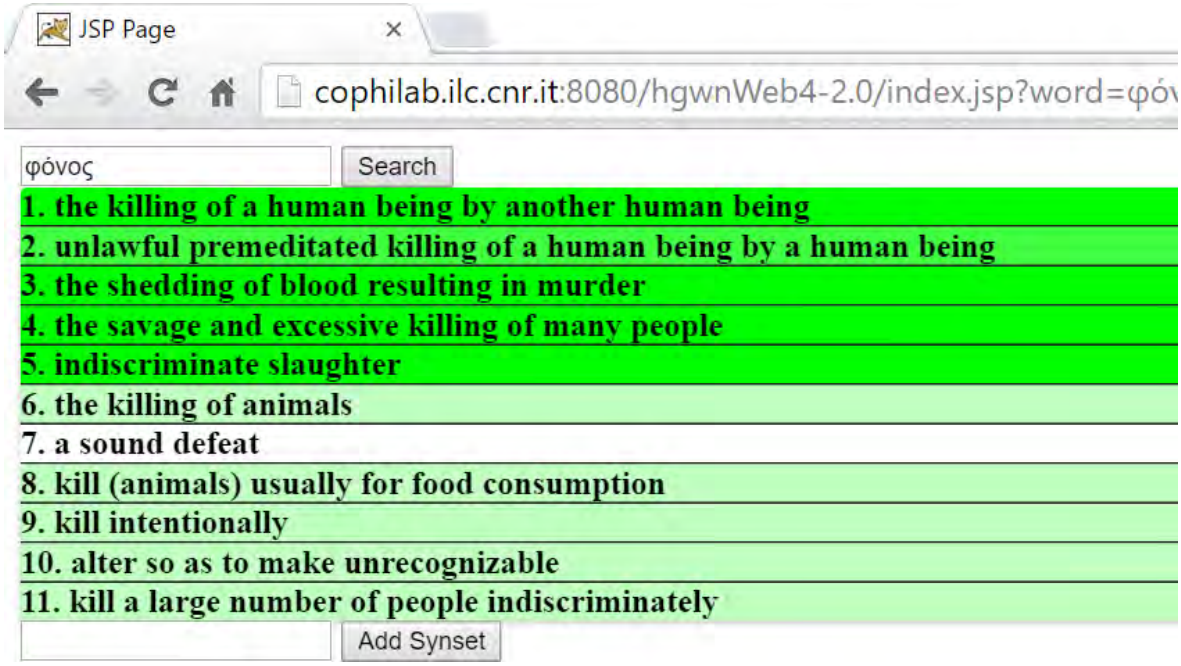


Figura 1: φόνος (phónos: omicidio / strage / spargimento di sangue). Legenda: verde brillante: senso omerico; verde pallido: senso attestati in Greco, ma non in omero; bianco: senso non attestato in Greco

I nostri primi stagisti hanno scelto di lavorare su campi semantici lontani fra di loro: quelli della morte, del sangue e della guerra da un lato e quello della percezione visiva e del colore, dall'altro.

La scelta del campo semantico della morte, del sangue e della guerra (operata seguendo alcuni dei raggruppamenti sinonimici di Paraskevaides 1984 e tenendo presente anche altre monografie, come Di Mauro Battilana 1985 o Mawet 1981) è stata dettata dalla ricchezza (per ovvie ragioni tematiche) di termini omerici di pertinenza. Termini come φόνος (phónos, cfr. Fig. 2), "omicidio" ma anche "sangue versato", mostrano una caratteristica del lessico omerico, secondo cui il concreto e l'astratto spesso si fondono insieme (e che HGWN separa in synsets diversi).

La scelta del campo semantico ottico parte dall'assunto che "i personaggi di Omero agiscono e reagisono soprattutto in base alle loro percezioni visive" (Mugler 1964). Di

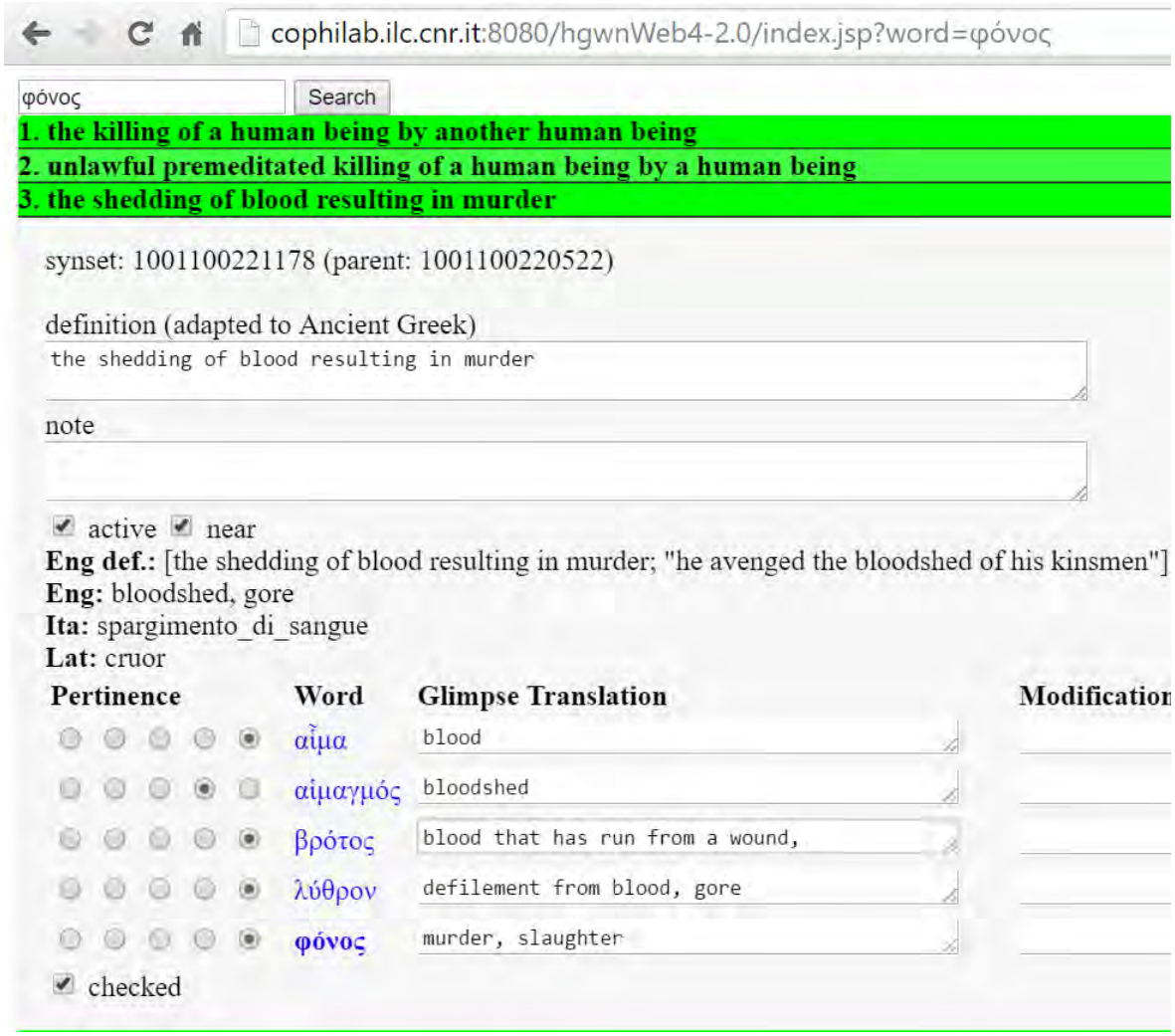


Figura 2: Scheda di validazione. Legenda dei livelli di pertinenza: 1. non pertinente; 2. legame semantico largo; 3. legame semantico stretto; 4. pertinente in Greco post-omerico; 5 pertinente in Omero

grande interesse è l'evoluzione di termini relativi alla vista transitati in ambito filosofico-speculativo, il cui uso è già omerico. Lo studio di tale ambito consente di isolare eventuali sinonimi omerici di termini filosofici successivi, alla ricerca di connessioni semantiche e lessicali (un termine come εἶδος, reso celebre dall'uso platonico come "idea", aveva in Omero un senso molto più concreto, simile a quello di "corpo" / "figura"): vale la pena notare che nei synsets di HGWN sono presenti non solo i termini attestati in Omero, ma anche i termini non omerici ad essi sinonimi e marcati con un livello di pertinenza più basso.

Il campo della percezione visiva è inoltre legato alla sfera dei sentimenti, spesso descritta attraverso metafore o associazioni luminose: in Omero αἴγλη connota il bagliore di una luce riflessa, ma subisce un ulteriore sviluppo semantico figurato ad esempio in Pindaro, in cui indica la gloria acquisita, la magnificenza dell'eroe o dell'atleta.

Le glosse dei synsets sono ereditate da Princeton WordNet, ma gli stagisti sono invitati a modficare la glossa (e stabilire la relazione di near equivalence con Princeton WordNet)

nei casi in cui i termini facciano riferimento a paradigmi epistemici differenti:si pensi alla tassonomia degli animali, alla definizione dei corpi celesti o delle malattie, etc.

3 Risultati ottenuti finora e risultati attesi

La dimensione del lessico omerico è di circa 9000 lemmi, di cui oltre il 50% è contenuto nella AGWN attuale. I circa 50 termini, presi in considerazione dai primi due stagisti nei tre mesi di operatività del progetto, hanno già portato alla validazione di un totale di circa 1000 synsets, (data una media di 20 synsets per parola). Mano a mano che il lavoro progredisce, sarà possibile validare sempre più termini, grazie al fatto di ritrovarli in synsets già validati contenenti sinonimi dei termini ancora da validare.

4 Conclusioni e sviluppi futuri

Considerate le esperienze svolte e i risultati raggiunti, ad altri studenti che intendono lavorare nel gruppo del LAMA sarà proposto di sfruttare la possibilità di uno stage presso il CoPhiLab. L'esperienza maturata può condurre a programmare sempre meglio (ottimizzando così, tempi e risultati) gli stage che saranno svolti. Riteniamo opportuno, ove possibile, continuare a orientare il lavoro di stage intorno agli interessi specifici dei singoli studenti, in modo da ampliare le loro conoscenze e competenze nello studio della lingua e della letteratura greca antica.

Abbiamo intenzione di creare wordnet di altri autori greci per osservare diacronicamente gli sviluppi lessico-semantici in prospettiva multidisciplinare: il primo passo già pianificato è la risemantizzazione in tragedia di termini omerici. Date le conseguenze attese sull'apprendimento linguistico (dinamico vs statico, autonomo vs dipendente da vocabolario, ragionato vs mnemonico), riteniamo che lo strumento possa avere impieghi non solo di ricerca, ma anche didatticamente rilevanti.

Bibliografia

Bizzoni, Y., et al. 2014. «The making of Ancient Greek WordNet». In *Proceedings of the 9th Annual Conference of LREC*.

Cunliffe, R. J. 2012. *A lexicon of the Homeric dialect*. Norman, OK: University of Oklahoma Press.

Di Mauro Battilana, G. 1985. *"Moira" e "Aisa" in Omero: una ricerca semantica e socioculturale*. Roma: Edizioni dell'Ateneo.

Ebeling, H., cur. 1880–1885. *Lexicon Homericum*. Lipsiae: Teubner.

Fellbaum, C. 1998. *Wordnet, an Electronic Lexical Database*. Cambridge, MA: MIT Press.

Hernandez, M. M. 1997. *Semàntica del Griego Antiguo*. A cura di Clàsicas. Madrid: Ediciones Clásicas.

Mawet, F. 1981. «Evolution d'une structure sémantique: le vocabulaire de la doleur: Apollonios de Rhodes et Homère». *L'Antiquitè Classique* 50 (1-2): 499–516.

Mugler, C. 1964. *Dictionnaire Historique de la Terminologie Optique des Grecs*. Paris: C. Klincksieck.

Paraskevaides, H. A. 1984. *The use of synonymous in the Homeric formulaic diction*. Amsterdam: A. M. Hakkert.

Pavese, C., e F. Boschetti. 2005. *A Complete Formular Analysis of the Homeric Poems*. Amsterdam: Hakkert.

Touch what you see: 3D design of eye tracking data

Barbara Balbi, Università degli Studi Suor Orsola Benincasa, Napoli,
barbara.balbi@centroscienzanuova.it
Emanuele Garzia, Università degli Studi Suor Orsola Benincasa, Napoli,
garziaems@gmail.com
Federica Protti, Università degli Studi Suor Orsola Benincasa, Napoli,
federica.protti@centroscienzanuova.it
Roberto Montanari, Università degli Studi Suor Orsola Benincasa, Napoli,
roberto.montanari@unisob.na.it

1 Introduction

Data representation takes advantage, above all, of the visual channel, by also allowing interaction and manipulation. However, data can be rarely physically manipulated, thus renouncing to the potential of the perceptual channel of touch. The main purpose of this project is to represent in a 3D format intangible data, concerning the perception of works of art, i.e. data belonging to the visual pathway of subjects during art perception. Therefore, a physical artifact shows the results of visual behaviors, i.e. the ocular-motorial tasks performed by people who stand in front of a Caravaggio's masterpiece, with the result of an unexpected synesthesia. The aim of this project is to go beyond the state of the art in digital objects and data *physicalization* field by providing new and innovative devices which allow a new type of sensorial exploration. In this way, we have also suggested a possible development of this research by applying the same methodology to written texts, in particular by designing an experiment that wants to lead to the data physicalization as a reading support for a text closely related to the Caravaggio masterpiece here studied.

2 The study proposal

The concept of "Active touch" refers to the blind recognition, i.e. the manipulation of an object without any visual contact. These manual exploration movements allow the stimulation of the tactile sense, and are able to activate other body functions (muscles, articulations, tendons) as proprioception, which contributes to the recognition of the dimension and other structural properties of the object. Can proprioception be an effective way to better understand virtual amount, typically assigned to visual recognition (e.g. graphics)? Can synesthetic experiences be a merging visual scanning? Can the tactile perceivable amount of the glance intensity be something that could improve artistic experience and understanding?

3 Data physicalization related works

In the framework of data visualization of complex phenomena, the majority of the existing studies confirms the importance of an integrated approach to info-visualization design, which also includes the involvement of various sensorial channels The integration process between the visual and the tactile channels is considered fundamental in order to allow, as well as to facilitate, the comprehension of complex issues (Hornecker 2011; Stusak et al. 2014).

The EMERGE prototype built within the project GHOST (2015) is an excellent example of what is defined as physicalization. In particular, the project is a significant case of data-based user interactions built via a dynamic bar chart. It contains self-actuating rods capable of RGB colors' output and touch detection based on data pushing/pulling; in addition, a surrounding surface enlarges the traditional touch detection (Taher et al. 2015).

Moreover, several studies are referred to the validation and production of physical, visual and interactive information systems, which are mostly dedicated to scientific data representation.

On the contrary, applications for physical data visualization of the artistic fruition experiences for enhancing Cultural Heritage, are quite uncommon.

The only existing examples in this sense can be experienced in the so-called Virtual Cultural Heritages (Hulusic and Rizvic 2013) and are addressed to the reconstruction and rendering of 3D models of both artistic pieces and places used for studying and designing virtual and augmented reality environments. Even in the event that the digitization and virtualization objects are written texts, the physicalization focuses on inscriptions, making a translation from object to object (Mittica, 2014; Amato, 2014). The object of our research, i.e. data concerning the artistic perception, are not often used as tools for designing new forms of augmented reality.

4 Implementation

More specifically, this study is focused on the possibility of building tangible data starting from the visual experience of the visitors, by using an eye tracking device, able to provide information about the visual pathway, as well as the intensity of the gaze. Data were collected from Caravaggio's' masterpieces placed in Pio Monte della Misericordia, i.e. Le sette Opere di Misericordia, located in Naples (Balbi and Protti 2015). During the experiment a sample of 30 users were asked to wear an eye tracker device and to observe the painting. Lately, areas of interest based on the most intense parts observed by visitors have been identified (Fig. 1). Results of the eye-tracking study have been published (Balbi, 2016). By taking as a point of reference the visual pathway, which has resulted to be the most recurrent to the majority of the observers together with the corresponding Area of Interest, a 3D shape deriving from the aggregation of the above mentioned data has been created through the use of Rhinoceros Software and its plug-in Grasshopper, which made data physicalization possible. Results have been properly overlaid in the corresponding area of the paint via a normalization process.

The pattern obtained includes information, which have been physically displayed as coordinates x and y of the light points of the canvas. Data which have been extracted from

Figure 1: Sette Opere di Misericordia with AOI

Figure 2: A render of AOI designed with Rhinoceros and Grasshopper

the visitors' visual pathway are: AOI on the work of art (areas which are observed for the longer time and with higher intensity), areas observed with higher intensity, respectively represented by curves, and three-dimensional ridges. Furthermore, data concerning peaks of luminosity, according to the fact AOI correspond to the brighter areas have been also taken into account.

Results have been printed in 3D with the purpose to create a tactile information device able to communicate data in a physical shape(Fig. 3). This design has resulted in a polystyrene panel on which there are some kind of hills, whose wideness corresponds to the collection of data mentioned above. The modeled data obtained by using the eye tracker device can be considered significant because they were the same for the majority of the visitors. By using this specific deployment of a visual-tactile synesthesia, it is possible for the users to improve their artistic awareness

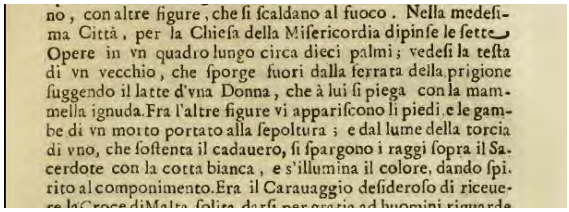
5 Conclusion and future work

Our future research plans are to work on the possibility of associating to the hills defined by the points of higher brightness to other distinctive elements, such as temperature. For instance, by adding temperature, an element to which people are particularly sensitive, it could be possible to raise our awareness of the data related to number of visit counts on a specific area of the canvas. Through the tactile exploration of the above mentioned hills,

Figure 3: 3D printed object

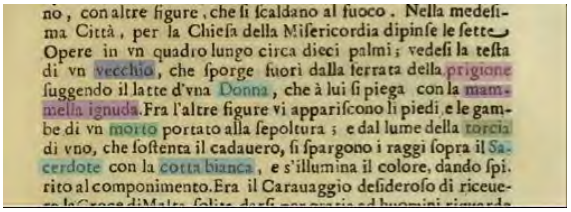
indeed, it could be possible to have effective information concerning the recurring path during the observation of the painting, the duration of the visits, the number of participants, as well as the more observed areas. Moreover, the effectiveness of this project will be evaluated as a support for tactile itineraries addressed to blind users. There are several studies which highlight the role played by the images as narrative device. In this case, data physicalization allows the possibility to deepen the understanding of these devices through active touches. Accordingly, it seems to be reasonable that the same method can also be used on written work. This led us to conceive an experiment on the *physicalization* of data by using written text as independent variable, also as a support in even more understanding the pictures and vice versa. We have selected for this study a short excerpt from the "*Vite de' pittori, scultori e architecti moderni*" written by G. Bellori on 1672, one of the oldest descriptions of the Caravaggio's canvas which addresses the experiments of the present work. The text will be submitted to a sample of subjects that will be asked to perform a reading task. As the text -– digitalized in the highest resolution as it will be happened – is

not easy to be read for non-experts, a sample of expert readers of such type of materials will be recruited, in order to minimize the risk of any bias in the task's execution.



nò, con altre figure, che ſi ſcaldano al fuoco. Nella medeſima Città, per la Chieſa della Miſericordia dipinſe le ſette Opere in vn quadro lungo circa dieci palmi; vedeſi la teſta di vn vecchio, che ſporge fuori dalla ferrata della prigione ſuggendo il latte d'vna Donna, che à lui ſi piega con la mammella ignuda. Fra l'altre figure vi appariſcono li piedi, e le gambe di vn morto portato alla ſepoltura; e dal lume della torcia di vno, che ſoſtenta il cadauero, ſi ſpargono i raggi ſopra il Sacerdote con la cotta bianca, e s'illumina il colore, dando ſpirito al componimento. Era il Carauaggio deſideroſo di riceue-

Figure 4: Extract from G.Bellori "Vite de' pittori, scultori e architecti moderni"



nò, con altre figure, che ſi ſcaldano al fuoco. Nella medeſima Città, per la Chieſa della Miſericordia dipinſe le ſette Opere in vn quadro lungo circa dieci palmi; vedeſi la teſta di vn vecchio, che ſporge fuori dalla ferrata della prigione ſuggendo il latte d'vna Donna, che à lui ſi piega con la mammella ignuda. Fra l'altre figure vi appariſcono li piedi, e le gambe di vn morto portato alla ſepoltura; e dal lume della torcia di vno, che ſoſtenta il cadauero, ſi ſpargono i raggi ſopra il Sacerdote con la cotta bianca, e s'illumina il colore, dando ſpirito al componimento. Era il Carauaggio deſideroſo di riceue-

Figure 5: Extract from G. Bellori with AOI

The subjects whose proposed the ancient text, will wear an eye-tracker device, such as in the experiment of the vision of the *Sette Opere di Misericordia*. In the same way first described, it will be possible to extract data related to the scanning of the written text. The dependent variable considered, once outlined the areas of interest, will probably correspond to the words on which the gaze lingers more time (Fig. 5), that will allow to measure the attention as it is an indicator of the intensity in which a visual stimulus is processed. This measure will depend both on the analysis of the visit count (i.e. the metric which measures the number of visits within an active Areas Of Interest) and the number of fixation (i.e. the number of micro-movements of the fovea).

The figure (Fig. 6) shows an example deriving by the use of an eye-tracker device on the above described text, where the relative visualization of the heat-map of the gaze of a reader are reported. The lower right image is a rendering of how could be the 3D object obtained from these data (Fig. 7). It may, therefore, contain the relevant data to the scanning mode, the most meaningful words, and these data could be intersected with those coming from the vision of the painting, by also highlighting both where data orientation (clearly outlined by the physical layout and landscape) are coherent and / or in contradiction. Thus, the synthetic translation of the data into the 3D format reading mode, could be once again considered as a device to increase the knowledge of the painting, as well as the ancient text.

Figure 6: Screenshot from eye-tracker device: heatmap On proposed text

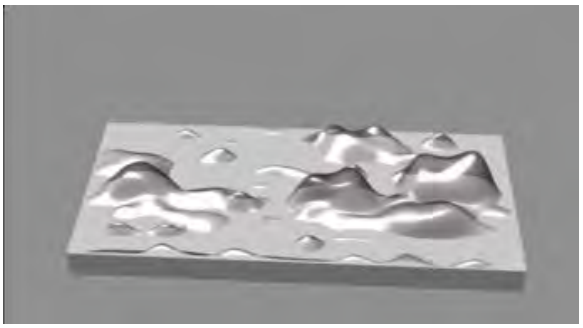


Figure 7: A render of designed object obtained with data text

References

Balbi, B., and F. Protti. 2015. «Caravaggio: Track the dark light. Misurazione dell'esperienza di fruizione dell'opera d'arte». In *Proceedings of Workshop LOSAI, Laboratori Open su Arte Scienza ed Innovazione*.

Brown, C., and A. Hurst. 2012. «VizTouch: Automatically Generated Tactile Visualizations of Coordinate Spaces». In *Proceedings of the Sixth International Conference on Tangible, Embedded and Embodied Interaction*, 131–138. TEI '12. Kingston, Ontario, Canada: ACM. ISBN: 978-1-4503-1174-8. doi:10.1145/2148131.2148160. http://doi.acm.org/10.1145/2148131.2148160.

Hornecker, E. 2011. «The Role of Physicality in Tangible and Embodied Interactions». *Interactions* (New York, NY, USA) 18, no. 2 (): 19–23. ISSN: 1072-5520. doi:10.1145/1925820.1925826. http://doi.acm.org/10.1145/1925820.1925826.

Hulusic, V., and S. Rizvic. 2013. «Story Guided Virtual Environments in Educational Applications». In *Transactions on Edutainment IX*, 132–149. Berlin, Heidelberg: Springer-Verlag. ISBN: 978-3-642-37041-0. http://dl.acm.org/citation.cfm?id=2554449.2554458.

Stusak, S., et al. 2014. «Activity Sculptures: Exploring the Impact of Physical Visualizations on Running Activity». *IEEE Transactions on Visualization and Computer Graphics* 20, no. 12 (): 2201–2210.

Taher, F., et al. 2015. «Exploring Interactions with Physically Dynamic Bar Charts». In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 3237–3246. CHI '15. Seoul, Republic of Korea: ACM. ISBN: 978-1-4503-3145-6. doi:10.1145/2702123.2702604. http://doi.acm.org/10.1145/2702123.2702604.

Illuminated Dante Project (IDP): Una startup per la digitalizzazione e l'analisi della più antica iconografia dantesca (secc. XIV-XV)

Gennaro Ferrante, Università degli Studi di Napoli "Federico II",
gennaro.ferrante@unina.it
Ciro Perna, Università degli Studi di Napoli "Federico II", ciro.perna@unina.it
Paola Denunzio, Università degli Studi di Napoli "Federico II", paola.denunzio@unina.it
Luigi Tessarolo, Web engineer, luigi.tessarolo@fastwebnet.it
Lucia Merolla, Istituto Centrale per il Catalogo Unico, l.merolla@iccu.sbn.it
Laura Ciancio, Istituto Centrale per il Catalogo Unico, l.ciancio@iccu.sbn.it

1 Presentazione generale

Illuminated Dante Project (IDP) punta a creare il più esaustivo archivio di antiche illustrazioni della *Commedia* su libro manoscritto (XIV e XV secolo), interrogabili e confrontabili su base codicologica, stilistica e iconologica. Il progetto, nato in un contesto di studio dell'esegesi del testo dantesco anche attraverso l'apparato illustrativo, si propone di selezionare un *corpus* di manoscritti miniati, che saranno innanzitutto descritti codicologicamente nella piattaforma di *Manus online*, il database dell'Istituto Centrale per il Catalogo Unico delle Biblioteche Italiane (ICCU) espressamente dedicato ai manoscritti. Le schede immesse in MOL saranno di seguito importate su una piattaforma ospitata dai servers di UNINA (http://www.dante.unina.it), che accoglierà tanto il *repository* delle immagini quanto un database iconografico espressamente concepito per analizzare e classificare tutte quelle immagini del corpus ermeneuticamente pertinenti al testo poetico (le «illustrazioni» propriamente dette). L'obiettivo principale della startup IDP, entro la fine del 2017, sarà la pubblicazione di un primo archivio *online* di immagini ad alta definizione e di libero accesso, al quale associare un *database* che sia già pienamente operativo su un sub-*corpus* (c.ca 50 mss.) inizialmente limitato ai manoscritti del Trecento conservati nelle biblioteche statali italiane (118 mss.). Il V convegno dell'AIUCD sarà l'occasione per presentare sia il progetto nelle sue articolazioni e nelle sue potenzialità, che le partnership finora stabilite con Enti di catalogazione nazionali. Il panel proposto si articolerà nei seguenti due slots tematici, rispettivamente di 30 e 20 minuti (suddivisi complessivamente in cinque *papers* brevi di 10 minuti), anticipati da un'introduzione generale di 10 minuti.

2 Introduzione

Gennaro Ferrante, Università degli Studi di Napoli "Federico II", *La sfida di un archivio digitale delle illustrazioni dantesche: qualità della ricerca, interoperabilità*, open access.

L'intervento introduttivo, del responsabile scientifico di IDP, presenterà il progetto nelle sue articolazioni interne e informerà della strutturazione di un'équipe interdisciplinare (filologi, storici dell'arte, codicologi, catalogatori e informatici), delle fasi di realizzazione; nonché della definizione di un protocollo di intervento per l'acquisizione delle immagini, per la concessione del copyright delle stesse per scopi di ricerca e per l'adozione di standard internazionalmente riconosciuti per l'interoperabilità dei metadati testo e immagine (TEI-P5 e IIIF). Si darà altresì conto di possibili sviluppi degli applicativi prodotti e di scenari di collaborazioni future.

3 Prima sezione: *Illuminated Dante Project: metodo e prassi nell'immissione dati e nella concezione del database.*

1. Paper: Ciro Perna (Università degli Studi di Napoli "Federico II"), *Dall'immissione dei dati in* Manus OnLine *alla strutturazione del database* IDP.

L'intervento rappresenterà una introduzione alle relazioni di Paola Denunzio e Luigi Tessarolo, e toccherà talune problematiche relative all'immissione dati nella piattaforma di ManusOnLine, per il versante propriamente codicologico-paleografico, nonché alla strutturazione dei campi del database IDP, sul piano prettamente storico artistico. Nello specifico saranno evidenziate le scelte più significative nell'articolazione delle schede (completezza, uniformità ed esaustività dei dati codicologici) e soprattutto nella co-struzione del DB (descrizione del progetto decorativo, rapporto con le immagini).

2. Paper: Paola Denunzio (CAB-Centro di Ateneo per le Biblioteche, Università degli Studi di Napoli "Federico II"), *L'utilizzo di* Manus OnLine *da parte del team di* IDP.

Nella relazione proposta si provvederà a descrivere per l'appunto le modalità e i criteri seguiti per la compilazione delle schede codicologiche in *Manus online*. L'intervento chiarirà le tappe successive dell'operazione; si illustreranno tutte le problematiche venute in luce durante l'immissione dei dati (per esempio, per quanto riguarda il riallineamento di certa terminologia codicologico-paleografica alla denominazione dei campi prevista da Manus). Si preciseranno inoltre le ragioni che hanno indotto di volta in volta all'assunzione o alla creazione di determinati record di autorità. La possibilità di creare nuove voci destinate ad arricchire il già importante *authority file* di Manus è stata resa possibile dalla collaborazione con i responsabili del grande catalogo elettronico, fiduciosi nel valore culturale del progetto, nella piena affidabilità dei ricercatori coinvolti e della operatività della nuova risorsa proposta. Una serie di sussidi visivi accompagneranno lo sviluppo sintetico della comunicazione, fornendo esempi pratici sia del materiale organizzato che delle modalità d' interrogazione e ricerca dell'OPAC di Manus messo a disposizione della comunità degli studiosi.

3. Paper: Luigi Tessarolo (*Illuminated Dante project*, Web engineer, team IDP), *L'architettura informatica di* IDP.

IDP si avvale, per l'inserimento dei dati codicologici generali, della piattaforma *Manus online* (ICCU), che ha messo a disposizione del progetto la propria interfaccia di *back-end*. Dal database MOL i dati sono esportati in formato XML e trasferiti nel database IDP, in cui sono integrati con quelli di un'approfondita analisi iconografica, iconologica e filologica delle immagini. A tale scopo è stata messa a punto una procedura di *back-end* dedicata, in grado di definire per la singola immagine, avvalendosi il più possibile di campi lemmatizzati, tanto gli elementi descrittivi materiali (illustratore, datazione, origine, tecnica esecutiva, tipologia decorativa), quanto quelli iconografici e iconologici. IDP intende servirsi, per l'esplorazione delle illustrazioni dei codici da parte dell'utente finale, di tecnologie che consentono un efficiente accesso web a immagini di altissima risoluzione (*Tiled Multi-Resolution TIFF* o *jpeg2000*). A tale scopo è stato allestito un repository di grande capacità, gestito da un server web idoneo (*IIPImage Server*, modulo che si integra in *Apache*), che consentirà, attraverso la conformità al protocollo IIIF (*International Image Interoperability Framework*) di: accedere alle immagini attraverso un *viewer* come *Mirador*, molto efficiente e dotato di funzioni di particolare interesse, come quella di confronto complanare tra due o più immagini; di visualizzare immagini collocate in altri *repository* (purché conformi al protocollo IIIF); di mettere a disposizione il proprio contenuto a chiunque operi nel web secondo lo stesso protocollo. Il *front-end* (in via di allestimento) sarà dotato di un apparato di interrogazione operante secondo un'ampia combinazione di criteri, riguardanti anche e soprattutto i dati specifici delle immagini e gli eventuali collegamenti al testo poetico; sarà inclusa comunque una ricerca full-text, operante sulle schede complete come se fossero pagine di testo, dunque indistintamente in tutti i campi. All'utente loggato sarà consentito di memorizzare collegamenti personalizzati tra illustrazioni appartenenti a codici differenti, e di aprirle affiancate mediante *Mirador*.

4 Seconda sezione: *La collaborazione tra IDP e i laboratori dell' ICCU.*

1. Paper: Lucia Merolla (Iccu, Responsabile *Manus online*), *Il database Manus OnLine e l'apertura al progetto* IDP.

Nell'intervento saranno illustrate le nuove possibilità offerte da *Manus OnLine* (MOL) e le modifiche apportate alla procedura. In particolare, si presentano il rinnovato aspetto grafico del sito e la nuova struttura del data base che consentirà l'inserimento di un modulo per la gestione indipendente di progetti di ricerca specialistici e internazionali (in fase di immissione dei dati e in fase di pubblicazione nell'OPAC). con particolare attenzione alle novità e alle problematiche emerse nell'apertura di MOL a un progetto trasversale di immissione dati come IDP. Inoltre, si evidenzia la funzionalità che permette da MOL, tramite un link, la visualizzazione delle immagini dei manoscritti presenti nel portale Internet Culturale (IC) dell'ICCU e descritti nel catalogo MOL. Il secondo intervento si focalizzerà sui termini della partnership con il laboratorio *Manus OnLine* (MOL) dell'ICCU,

2. Paper: Laura Ciancio (ICCU, Responsabile *Internet Culturale*), *La collaborazione di* Internet Culturale *con il progetto* IDP.

Il secondo intervento del terzo slot illustrerà le modalità di metadatazione delle immagini dell'archivio IDP all'interno del portale di Internet Culturale, specificando come

un'archiviazione di tipo tematico e dai presupposti euristici (archivio di soggetti iconografici della tradizione illustrativa antica della *Divina Commedia*) possa essere organizzata, ai livelli di BE e FE.

Bibliografia

Codicologia, Paleografia, Descrizione dei Manoscritti

Barbero, G. 2013. «Manoscritti e standard». *Digitalia* 8:43–65.

Bertelli, S. 2011. *La tradizione della "Commedia" dai manoscritti al testo. I. I codici trecenteschi (entro l'antica vulgata) conservati a Firenze*. Firenze: Olschki.

Boschi Rotiroti, M. 2004. *Codicologia trecentesca della Commedia. Entro e oltre l'antica vulgata*. Roma: Viella.

Commissione permanente per la revisione delle regole italiane di catalogazione, cur. 2009. *Regole italiane di catalogazione. REICAT*. Roma: ICCU.

Jemolo, V., e M. Morelli, cur. 1990. *Guida a una descrizione uniforme dei manoscritti e al loro censimento*. Roma: ICCU.

Marcuccio, R. 2010. «Catalogare e fare ricerca con Manus OnLine». *Biblioteche oggi* 27:33–49.

Petrucci, A. 1984. *La descrizione del manoscritto. Storia, problemi, modelli*. Roma: La Nuova Italia Scientifica.

Solimine, G., e P. G. Weston, cur. 2015. *Biblioteche e biblioteconomia. Principi e questioni*. Roma: Carocci.

Iconologia e Iconografia Dantesca

Battaglia Ricci, L. 2001. «Il commento illustrato alla Commedia». In *Per correr miglior acque*, 1:601–639. Roma.

— . 2008. «Un sistema esegetico complesso: il Dante Chantilly di Guido da Pisa». *Rivista di Studi Danteschi* 8 (1): 40–57.

— . 2009. «Ai margini del testo: considerazioni sulla tradizione del Dante illustrato». *Italianistica* 38 (2): 39–58.

— . 2011. «La tradizione figurata della Commedia». *Critica del testo* 14 (2): 547–579.

Bolzoni, L. 1995. *Le stanze della memoria*. Torino.

Brieger, P., M. Meiss e C. Singleton. 1969. *Illuminated Manuscripts of the Divine Comedy*. 2 vols. London.

Carruthers, M. 2008. *The book of memory. A Study of Memory in Medieval Culture*. 2ª ed. Cambridge.

Ciardi Dupré Dal Poggetto, M. 2003. «Prime osservazioni sulle illustrazioni del Paradiso di Dante». In *Visioni dell'aldilà in Oriente e Occidente*, 167–198. Arte e pensiero.

Ciardi Dupré, M. 1989. ««Narrar Dante» attraverso le immagini: le prime illustrazione della 'Commedia'». In *Pagine di Dante. Le edizioni della 'Divina Commedia' dal torchio al computer. Catalogo della mostra, Foligno 11 marzo-28 maggio 1989; Ravenna, 8 luglio- 16 ottobre 1989, Firenze 1990*, a cura di R. Rusconi, 79–102. Perugia: Electa-Editori Umbri Associati.

— . 1990–1991. «A proposito della "descrizione uniforme dei manoscritti miniati"». *Miniatura* 3–4:123–124.

Hughes Gillerman, D. 1959. «Trecento illustrators of the 'Divina Commedia'». *Annual Report of the Dante Society* 77:1–40.

Mazzucchi, A. 2002. *Chiose Filippine. Ms. CF 2 16 della Bibl. Oratoriana dei Girolamini*. Roma.

Miglio, L. 2003. «I commenti danteschi: i commenti figurati,» in *Intorno al testo*, 377–401. Roma.

Pasut, F. 2006a. «Codici miniati della Commedia a Firenze attorno al 1330: questioni attributive e di cronologia,» *Rivista di Studi danteschi* 6 (2): 379–409.

— . 2006b. «Il "Dante" illustrato di Petrarca: problemi di miniatura tra Firenze e Pisa alla metà del Trecento». *Studi petrarcheschi* 19:115–147 + tt.

— . 2008. «Pacino di Buonaguida e le miniature della Divina Commedia: un percorso tra codici poco noti». In *Da Giotto a Botticelli. Pittura fiorentina tra Gotico e Rinascimento. Atti del convegno internazionale - Firenze, Università degli Studi e Museo di San Marco, 20-21 maggio 2005,* a cura di F. Tripps e J. Tripps. Firenze.

— . 2012. «Florentine Illuminations for Dante's "Divine Comedy": a critical assessment». In *Florence at the dawn of the Renaissance: painting and illumination, 1300-1350*, a cura di C. Sciacca, 155–169. Los Angeles.

Ponchia, C. D. 2014. «A Web Application for the History of Art». In *Collaborative Research Practices and Shared Infrastructures for Humanities Computing,* a cura di M. Agosti e F. Tomasi, 219–228. Padova.

I margini del libro. Indagine teorica e storica sui testi di dedica. Atti del Convegno Internazionale di Basilea, 21-23 novembre 2002. 2004. Roma-Padova: Antenore.

Rijnberk, G. van. 1954. *Le langage par signes chez les moines*. Amsterdam.

Roddewig, M. 1984. *Dante Alighieri. "Die göttliche Komödie"*. Stuttgart.

Villa, C. 1996. «Commentare per immagini». In *Vedere i classici*, 51–68. Roma.

Zanichelli, G. Z. 2006. «L'immagine come glossa». In *Dante e le arti visive*, 109–148. Milano.

Analisi Testuale
Textual Analysis
Talks

Formal ontologies for narrative text analysis

Fabio Ciotti, Università di Roma Tor Vergata fabio.ciotti@uniroma2.it

1 Introduction

In this paper, I present a research program situated at the intersection of narratology and computer sciences, in the main area of Digital Literary Studies. The methodological underpinning of this program is the adoption of formal ontology modeling for the analysis of some relevant concepts in narratology and narrative texts analysis.

Such a computational formalization has various scientific objectives. In the first place, it is an attempt to draw a clear theoretical account of some key concepts of narratology and literary theory. Moretti (2013) in one of its most relevant articles devoted to the methodological foundation of computational text analysis has adapted the epistemological concept of operationalizing proposed by P.W. Bridgman: "Operationalizing means building a bridge from concepts to measurement, and then to the world. In our case: from the concepts of literary theory, through some form of quantification, to literary texts". This purely quantitative conception is too constrictive in my view. I suggest that computational formal modeling is a suitable kind of operational translation of literary theoretical terms and concepts, though not directly producing quantitatively measurable "proxies" of those terms.

In the second place, a formal ontology is a functional tool, as well. We can use it to add semantic annotations to digital textual corpora. This semantic enrichment, based on dedicated reference ontology vocabularies, allows the execution of complex queries and semantic analysis, using automatic reasoning tools (that is, systems able to infer new knowledge based on the ground knowledge that is asserted and on the semantics of the vocabulary that has been used to model that knowledge). The availability of this kind of applications and frameworks can bring relevant advances in the study of literary phenomena (especially in the field of thematic intertextuality) but also in the didactics of literature.

2 State of the Art

Computational narratology is one of the leading sectors of the post-classical narratology (Herman 1999), the purpose of which is "the study of narrative from the point of view of computation and information processing. It focuses on the algorithmic processes involved in creating and interpreting narratives, modeling narrative structure in terms of formal, computable representations" (Mani 2003). It incorporates and enhances the tradition of formalist and structuralist narratology in a computational perspective mixing it with the concepts and models of knowledge representation.

The field of story grammars (Meister 2003) has traditionally been the central theme of the studies in this field. It should be noted, however, that since the late 80 / 90's Gigliozzi (2008) and his team have experimented the modeling of narrative structures and characters, with formalisms such as Lisp. Recently, attention has turned toward the use of ontological technologies (Gruber 1993) for the representation of narrative structures both for the purposes of literary and narratological research (Zöllner-Weber 2007) and for the possible applications of storytelling processes in training, communication and production contexts (Damiano and Lieto 2013). Quite recently the preliminary results of an ontological modeling of narratives by Bartalesi, Meghini and Bartalesi, Meghini, and Metilli (2016) have been presented. This formalization tries to capture some basic notions of narratives namely fabula, narration and plot, and can be considered orthogonal to our ontological model of characters.

3 An ontology for narrative analysis

Starting from the methodological framework we have sketched above, I am developing a formal model of narrative based on ontological languages (and relative semantics), in particular in OWL 2 DL (W3C 2009). The rationale of this choice is of course the computational complexity of the inference algorithms licensed by that formalism and the availability of efficient and scalable semantic stores and inferences engines.

The literary theoretical bases of our project are deeply rooted in some of the relevant concepts developed in classical and post-classical narratology. The object of our model, in fact, are two key notions of narrative:

1. character / actor / actant, following the different level of abstraction proposed in the early works of Greimas (1969). Starting from the work carried out since the early 80s by Gigliozzi on the formal structure of the character in the novel and in the short story, I propose a formal model that allows describing the character as a frame (in Minsky's sense, 1974) composed of nuclear non-negotiable properties and negotiable accessories traits. The character, from this point of view, can be described in terms of a semantic field composed of elements bound together by the force of an isotopic bundle. According to Greimas, two or more elements build an isotopy when they are semantically homogeneous, that is, when they belong to the same semantic level. A cluster that thickens around a root node. In this context we also take into account the recent contributions of cognitive narratology about the concept of "theory of mind" and intentionality in fictional characters (Herman 2003).

2. narrative /fictional world / space, based on the concept of narrative semiotic space of Lotman and of the theories of fiction and narrative as possible worlds (Doležel 1999; Eco 1979; Ryan 1991). There is a deep relation between the notion of character and that of narrative space. According to Lotman (1972), the narrative space to which is bounded defines each character, and the hero is the only character that can move between different narrative spaces. The concept of narrative space can be formalized using the notion of fictional possible world, whose definition is given by Eco (1979, pp. 128-30): "Definiamo come mondo possibile uno stato di cose espresso da un

insieme di proposizioni dove per ogni proposizione o p o ~ p. Come tale un mondo consiste in un insieme di individui forniti di proprietà".

4 Conclusions and next steps

The ontological modeling that we are proposing is amenable to various extensions: we can think, in fact, to the problem of modeling the temporal aspect of the narrative world, or to the possibility of binding the abstract ontology of fictional character to an ontology of narrative motifs and themes.

A further step is the exploration of possible convergence with other narrative ontological frameworks, namely the one developed by Bartalesi, Meghini, and Metilli (2016) and Ontomedia (Jewell, Lawrence, and Tuffield 2005), and when possible the mapping with other relevant ontologies and encoding scheme in the cultural heritage domain for interoperability (e.g., Text Encoding Initiative and CIDOC-CRM). This is particularly relevant in that we want to publish our ontology and data sets as Linked Data, through recommended standard publishing methodologies and to appropriately link them to already existing datasets with similar or partially overlapping information in the Cultural Heritage LOD cloud.

The applications of a formal ontology of narrative, finally, are not limited to literary studies. In these years, the attention towards storytelling has grown up also in social sciences, media and game studies, enterprise communication. Our modeling effort can find relevant intersections and applications inside these domains too. We believe in fact that a model of storytelling based on a rich description of fictional characters and worlds can be more effective than those limited to action theory and story grammars formalization.

References

Bartalesi, V., C. Meghini, and D. Metilli. 2016. «Steps Towards a Formal Ontology of Narratives Based on Narratology». In *CMN 2016*. http://narrative.csail.mit.edu/cmn16/doc/p04.pdf.

Berners-Lee, T., J. A. Hendler, and O. Lassila. 2001. «The Semantic Web: A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities». *Scientific American* 279 (5): 34–43.

Bizer, C., T. Heath, and T. Berners-Lee. 2009. «Linked Data - The Story So Far.» *International Journal on Semantic Web and Information Systems (IJSWIS) – Special Issue on Linked Data* 5 (3): 1–22.

Damiano, R., and A. Lieto. 2013. «Ontological representation of narratives: a case study on stories and actions». In *Proceedings of CMN13*. Hamburg.

Doležel, L. 1999. *Heterocosmica. Fiction e mondi possibili*. Milano: Bompiani.

Eco, U. 1979. *Lector in fabula. La cooperazione interpretativa nei testi narrativi*. Milano: Bompiani.

Gigliozzi, G. 2008. *Saggi di informatica umanistica*. Milano: UNICOPLI.

Greimas, A. 1969. *Del senso*. Milano: Bompiani.

Gruber, T. R. 1993. «A translation approach to portable ontologies». *Knowledge Acquisition* 5 (2): 199–220.

Herman, D., ed. 1999. *Narratologies: New Perspectives on Narrative Analysis*. Ohio: Ohio State University Press.

— , ed. 2003. *Narrative Theory and the Cognitive Sciences*. Stanford: CSLI.

Hyvönen, E. 2012. *Publishing and Using Cultural Heritage Linked Data on the Semantic Web*. San Rafael, CA: Morgan & Claypool.

Lotman, J. 1972. *La struttura del testo poetico*. Milano: Mursia.

Mani, I. 2003. «Computational Narratology». In *The living handbook of narratology*, ed. by P. Hühn et al. Hamburg University.

Meister, J. C. 2003. *Computing Action. A Narratological Approach*. Berlin: De Gruyter.

Minsky, M. 1974. *A Framework for Representing Knowledge*. Cambridge, MA: MIT AI Lab. Memo 306.

Moretti, F. 2013. «Operationalizing: Or, the Function of Measurement in Literary Theory». *New Left Review* 84:103–119.

Ryan, M.-L. 1991. *Possible worlds, artificial intelligence, and narrative theory*. Bloomington: Indiana University Press.

Sanderson, R., P. Ciccarese, and e. Van de Sompel H. 2013. «Open Annotation Data Model». http://www.openannotation.org/spec/core/20130208/index.html.

W3C. 2009. «OWL 2 Web Ontology Language». http://www.w3.org/TR/2009/REC-owl2-overview-20091027.

Zöllner-Weber, A. 2007. «Noctua literaria - A System for a Formal Description of Literary Characters». In *Data Structures for Linguistic Resources and Applications*, ed. by G. Rehm and W. Lemnitzer. Tübingen.

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.

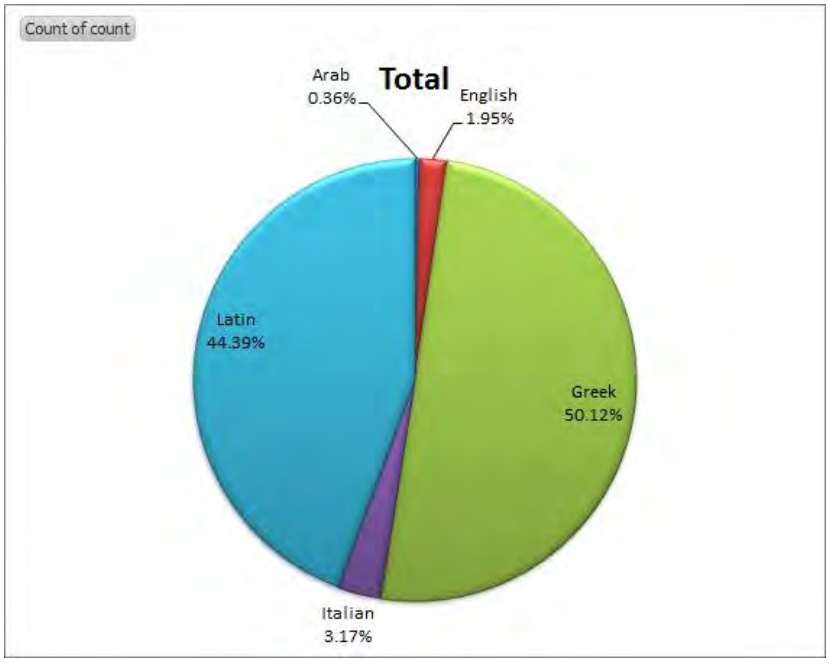


Figura 1

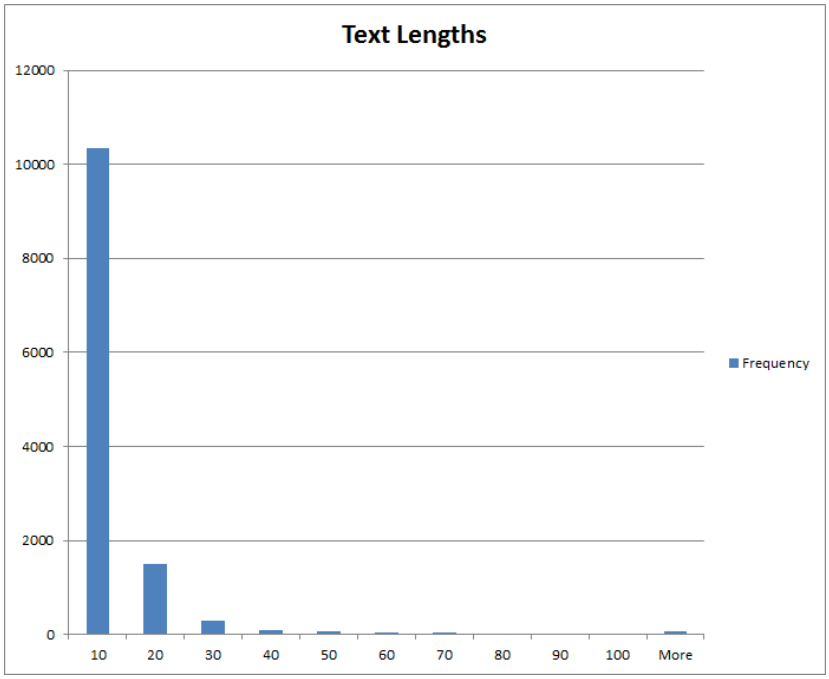


Figura 2

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-

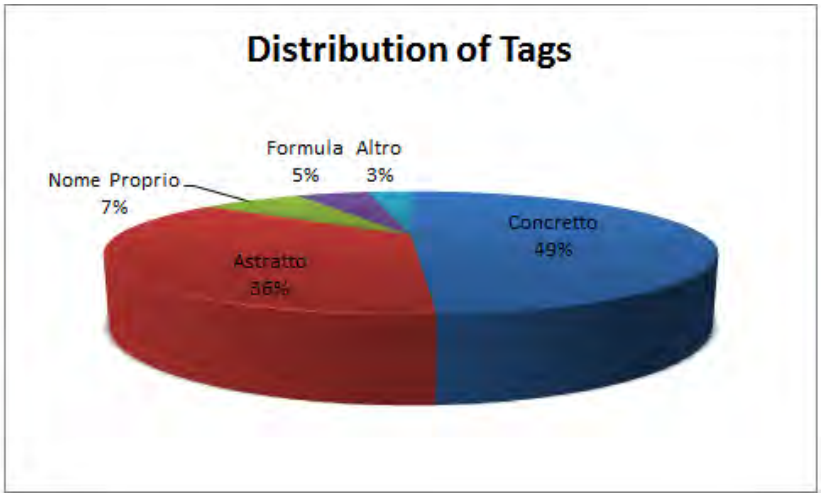


Figura 3

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*.

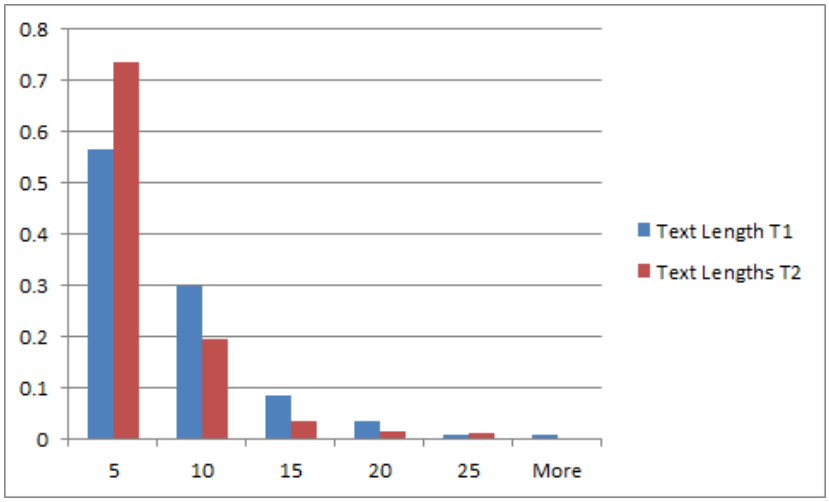


Figura 4: La lunghezza dei testi annotati da T1 e T2 (normalizzato per il numero totale dei testi taggati da T1 e T2).

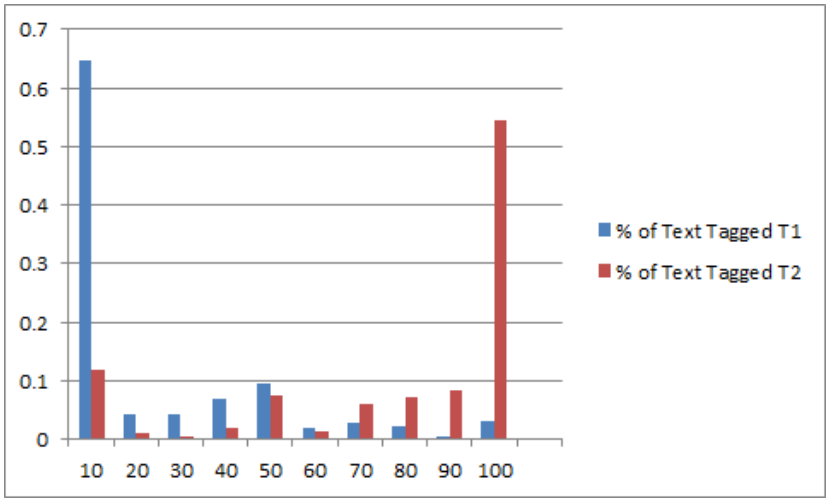


Figura 5: La frequenza della percentuale di ogni testo taggato (normalizzato).

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.

Il corpus di testi arabi in Memorata Poetis

Ouafae Nahli, CNR-ILC, ouafae.nahli@ilc.cnr.it
Federico Boschetti, CNR-ILC, federico.boschetti@ilc.cnr.it
Silvia Arrigoni, Università "Ca' Foscari", Venezia, silvia.arrigoni@unive.it
Luigi Tessarolo, Università "Ca' Foscari", Venezia, luigi.tessarolo@fastwebnet.it

1 Introduzione

Il progetto *Memorata Poetis*[1] è basato sullo studio intertestuale di componimenti poetici brevi di tradizione epigrafica e letteraria, plurilingui e multiculturali. Per superare i limiti della ricerca verbale, è stato creato un motore di ricerca semantico e tematico. Ciò permette di studiare iscrizioni in versi, di provenienza ed epoca varie, in rapporto ai testi letterari 'alti' da cui sono influenzate o che esse stesse influenzano, al fine di far emergere da un lato le mutue riprese e relazioni e dall'altro le affinità tematiche anche in assenza di una influenza diretta.

Il database di *Memorata Poetis* comprende testi in lingua greca, latina, araba, italiana delle origini, lituana antica e inglese, in alcuni casi frutto di riedizioni in formato elettronico (es. i corpora arabo e della poesia italiana delle origini). Attraverso l'allestimento di un indice gerarchico di temi e motivi, è stato poi possibile effettuare un mark-up tematico dei singoli testi[2]. In tal modo, il motore di interrogazione permette di abbracciare tradizioni culturali più che millenarie, con lo sguardo rivolto alla continuità della tecnica espressiva e stilistica e verso i meccanismi con cui agirono dall'interno i processi di riuso semantico.

Della ricerca intertestuale tramite temi e motivi comuni fra componimenti della tradizione classica greco-latina nella collezione di *Memorata Poetis* si è già parlato in altre occasioni, ad esempio al convegno di Venezia *Risorse digitali e strumenti collaborativi per le Scienze dell'Antichità* tenutosi il 2 e il 3 ottobre 2014, i cui atti sono in corso di stampa[3]. Per questo motivo nel presente contributo intendiamo focalizzarci sulla sezione in lingua araba del progetto, e sulla ricerca intertestuale dall'arabo alle altre lingue.

[1] http://www.memoratapoetis.it. A riguardo si vedano Mastandrea e Tessarolo (2014) e Mastandrea (2015), nonché Pistellato (2014). Il progetto è stato svolto nell'ambito del PRIN 2010 / 2011.
[2] Cfr. Ciotti (2014a). Sulla critica tematica si vedano Lefèvre (2006) e Pellini (2008).
[3] Mastandrea (2016).

2 Fasi di lavoro

2.1 Importazione e normalizzazione dei testi

Si è resa necessaria la digitalizzazione di tutto il *corpus* epigrafico arabo, dal momento che non esistono risorse digitali disponibili. Per motivi didattici e scientifici, i testi sono stati vocalizzati e tradotti in italiano. In seconda battuta, è stata compiuta una ricerca approfondita delle fonti arabe per mettere in relazione ogni epigrafe del corpus a poesie della tradizione letteraria. Quando possibile, la poesia a cui fa riferimento l'epigrafe è stata trascritta e tradotta nel campo relativo, la *Scheda di catalogazione* (Fig. 1), contenente indicazioni relative alla cronologia, al supporto del testo, la localizzazione, nonché la bibliografia di riferimento. Lo strumento si è rivelato tanto più utile in relazione al *corpus* arabo per l'inserimento di informazioni indispensabili per l'interpretazione del testo.

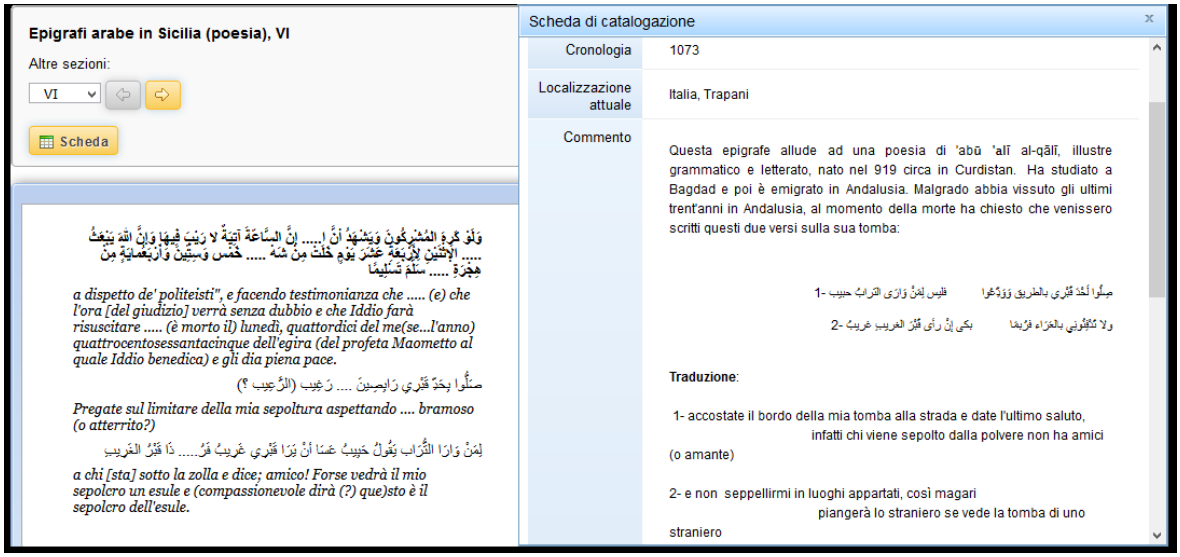


Figura 1: Esempio di epigrafe con la sua scheda di catalogazione

Anche l'utilizzo di note esplicative per termini della lingua araba che non trovano corrispondenza in italiano (toponimi arcaici o antroponimi) è stato introdotto per favorire la fruizione dei testi (Fig. 2).

L'annotazione dei testi di tradizione araba ha richiesto un proficuo arricchimento della tassonomia di Memorata Poetis, incentrata sulla tradizione occidentale[4], per tener conto delle prospettive culturali dei testi oggetto di studio; si veda in questo senso l'aggiunta di alcuni temi attinenti alla religione musulmana, quali *Religio Muslimica, basmala, Mohammed, Qur'ān, shahāda*.

[4]La tassonomia dei temi e dei motivi è stata stabilita da esperti di letteratura greca antica e latina, rispettivamente, Giancarlo Scarpa dell'Università "Ca 'Foscari" di Venezia e Paola Paolucci dell'Università di Perugia.

Figura 2: Esempio di testo con note

2.2 Il *corpus* dei testi in lingua araba in *Memorata Poetis*

Le epigrafi sono state digitalizzate secondo preesistenti edizioni autorevoli[5]:

- le *Epigrafi arabe in Sicilia*, 40 epigrafi prosodiche e 5 epigrafi poetiche;
- le *Epigrafi arabe in Arabia Saudita*, consistenti in 10 poesie;
- una *Scelta di epigrafi funerarie arabe*, raccolta di testi riuniti da al-ǧūzī (morto nel 1200), dei quali sono state digitalizzate 32 poesie;
- *'al-mu'allaqāt*[6], raccolta di poesie pre-islamiche delle quali è stata trascritta e tradotta la poesia di *'imru'u l-qays*[7].

3 Studio di casi

Nell'ambito dello studio dell'intertestualità e della trasmissione della conoscenza, lo stesso concetto può essere espresso attraverso i medesimi temi e motivi, indipendentemente dalla lingua in cui l'autore si esprime. Ad esempio, ricercando il tema del *Tempus fugiens et irreparabile*, è possibile trovare, fra gli altri, questi due testi in greco e latino, qui scelti per la loro notorietà (si veda anche Boschetti, Del Gratta e Lamé 2014b):

Anthologia Graeca 11, 56, Πῖνε καὶ εὐφραίνου. τί γὰρ αὔριον ἢ τί τὸ μέλλον, / οὐδεὶς γινώσκει [....]

(Bevi e sii felice. Nessuno sa come sarà domani o il futuro.)

Catull. *carm.* 5, *Viuamus, mea Lesbia, atque amemus / ... / Nobis cum semel occidit breuis lux, / Nox est perpetua una dormienda.*

[5] Vd. bibliografia araba.

[6] La poesia pre-islamica è costituita dalla raccolta di sette (o dieci) poesie arabe, probabilmente del VI secolo. Per la loro bellezza, queste poesie sono state scritte con inchiostro d'oro su stoffa e appese alla Mecca nella *Ka'ba*, da cui il nome *'al-mu'allaqāt*, "le Appese".

[7] *'imru'u l-qays* è uno dei pochi poeti arabi che ha frequentato l'Occidente; fu assassinato con un sudario avvelenato per aver sedotto la figlia di Giustiniano durante il suo soggiorno a Bisanzio.

al. 2014):

> *Anthologia Graeca* 11, 56, Πῖνε καὶ εὐφραίνου. τί γὰρ αὔριον ἢ τί τὸ μέλλον, / οὐδεὶς γινώσκει [....]
>
> Bevi e sii felice. Nessuno sa come sarà domani o il futuro.

> Catull. 5, *Viuamus, mea Lesbia, atque amemus* / ... / *Nobis cum semel occidit breuis lux*, / *Nox est perpetua una dormienda*.
>
> Viviamo, mia Lesbia, e amiamo ... Quando per noi tramonta la breve luce, noi dobbiamo dormire un'unica notte perpetua.

La ricerca intertestuale permette di mettere a fuoco anche le divergenze tra le civiltà o tra fasi diverse delle civiltà. Lo stesso tema della morte come limite e fine di "questa vita" è presente in buona parte della poesia araba, soprattutto, però, come riflessione sulle azioni terrene da compiere per ottenere la vita eterna:

> وَكيْفَ يَلِذُّ العَيْشَ مَنْ هُوَ عالِمٌ / بِأَنَّ إلَه الخَلْقِ لا بُدَّ سَائِلُه؟
>
> *Come può essere dolce la vita per chi è consapevole / che, sicuramente, il Dio delle creature gli chiederà [di rendere conto],*
>
> فَيَأْخُذُ مِنْهُ ظُلمَه لِعِبادِه / ويُجْزِيهِ بِالخَيْرِ الذِي هُوَ فاعِلُه
>
> *quindi, per le ingiustizie che ha commesso nei confronti dei suoi fedeli, lo punirà / e, per il bene che ha compiuto, lo premierà.*

È infine possibile rilevare concetti universali espressi all'interno di tutte le culture: il tema dello "straniero", ad esempio[8], viene declinato in maniera differente nei testi e identifica una diversa tipologia di persone; ciò nonostante, è possibile riconoscere motivi comuni che legano i testi. Nelle poesie epigrafiche, lo straniero si rivolge spesso ad un altro straniero, perché essi sono mossi dalle stesse esigenze e dalla stessa nostalgia, come si vede negli esempi riportati nella tabella[9].

[8]Per il tema della morte in luogo straniero nei carmina latina epigraphica e nella tradizione letteraria, si veda Cugusi (1996), 200-217.

[9]La prima epigrafe allude ad una nota poesia di 'al-qālī (illustre grammatico e letterato del X secolo) che abbiamo citato nella scheda di catalogazione (Fig. 1).

"straniero", ad esempio[8], viene declinato in maniera differente nei testi e identifica una diversa tipologia di persone; ciò nonostante, è possibile riconoscere motivi comuni che legano i testi. Nelle poesie epigrafiche, lo straniero si rivolge spesso ad un altro straniero, perché essi sono mossi dalle stesse esigenze e dalla stessa nostalgia, come si vede negli esempi riportati nella tabella.

Corpus	Testo in lingua originale	Traduzione
Epigrafi Arabe in Sicilia	(صَلُّوا بِحَدِّ قَبْرِي رَابِصِينَ .... رَغِيب (الرَّعِيب ؟ لِمَنْ وَارَا التُّرَابِ يَقُولُ حَبِيبُ عَسَا أنْ يَرَا قَبْرِي غَرِيبُ فَرُ..... ذَا قَبْرُ الغَرِيبِ	Pregate sul limitare della mia sepoltura aspettando .... bramoso (o atterrito?) a chi [sta] sotto la zolla e dice: amico! **Forse vedrà il mio sepolcro, un esule e (compassionevole dirà (?) que)sto è il sepolcro dell'esule**.
Poesia 'al-qālī	صِلُوا لَحْدَ قَبْرِي بالطريق وَوَدِّعُوا / فليس لِمَنْ وَارَى الترابُ حبيب ولا تَدْفِنُونِي بالعَرَاء فرُبمَا / بكى إنْ رأى قَبْرَ الغريبِ غريبُ	Accostate il bordo della mia tomba alla strada e date l'ultimo saluto, / infatti chi viene sepolto dalla polvere non ha amici (o amanti) e non seppellirmi in luoghi appartati, **così magari / piangerà lo straniero se vede la tomba di uno straniero**
Carmina Latina Epigraphica - *supplementa*	*Hospes, qui tumulum hun*[*c uides*,] | *Si non forte grau*(*e*) *est*, *d*[*isce*]. | *Hic Callo iaceo tellure as*[*pera*]. | *Finibus Italiae lucis ad*[*oleui.*] | *Coniunxs ad superos rem*[*eat* (?)] | *Et frater, quorum est luct*[*us*]. | *Ignotis cara et nulli non gr*[*ata*], | *Quis obitus noster est sin*[*ister*]. | *Nunc hospes hoc titulo lecto* [*dic*] | *discedens* '*Callo, sit tibi terra leuis*'.	**Straniero**, che guardi questa tomba, | se non ti è troppo gravoso, apprendi [queste cose]. | Qui io giaccio, Callo, sotto la dura terra. | Sono cresciuta nei boschi, nei territori dell'Italia | Lo sposo ritornò alle regioni celesti | e [anche] il fratello, le cui morti furono per me causa di dolore. | Fui cara agli sconosciuti e amata da tutti, per i quali la mia morte è infelice. | **Ora, straniero, letta questa epigrafe, [dì], | mentre ti allontani, "Callo, ti sia lieve la terra".**

4 Conclusione

La tradizione poetica araba non sembra avere molti punti di contatto diretto con la tradizione poetica classica. Ad esempio, i testi epigrafici arabi risultano ben distinti dal raggruppamento greco-latino quanto a immagini e metafore specifiche di questo corpus. Le somiglianze maggiori si trovano invece nel trattamento dei medesimi concetti universali che non dipendono da una specifica tradizione.

[8] Per il tema della morte in luogo straniero nei *carmina latina epigraphica* e nella tradizione letteraria, si veda Cugusi 1996, 200-217.

[9] Questa epigrafe allude ad una nota poesia di 'al-qālī (illustre grammatico e letterato del X secolo) che abbiamo citato nella scheda di catalogazione (fig. 1).

Edizioni a stampa dei testi arabi digitalizzati

'abū 'abdi l-lāhi l-ḥusayni bnu 'aḥmad bnu l-ḥusayni az-zūzī, *šarḥ al-mu'allaqāti as-sab'*, 'ad-dār al-'ālamiyyah - 1993.

'abū al-faraǧ 'abdu ar-raḥmāni bnu l-ǧūzī, *mutīr al-ġarām as-sākin ilā 'ašrafi al-'amākin*, casa editrice: dār al-ḥadīṯ - Cairo – 1995, 507 – 516.

'abdu ar-raḥmān bnu nāṣir as-sa'īd, *an-nuqūšu aš-ši'riyyatu fī al-mamolakati al-'arabiyyati as-sa'ūdiyyati*, casa editrice: majalatu darati al-maliki 'abdi al-'azīzi, al-'adadu aṯ-ṯānī - 2013. (http://www.awbd.net/article.php?a=68).

Michele Amari, *Le epigrafi arabiche di Sicilia trascritte, tradotte e illustrate. Parte seconda. Le iscrizioni sepolcrali.* Palermo 1879.

Bibliografia

Boschetti, F., R. Del Gratta e M. Lamé. 2014b. «Computer Assisted Annotation of Themes and Motifs in Ancient Greek Epigrams: First Steps». In *Proceedings of the First Italian Conference on Computational Linguistics CLiCit 2014, 9-10 December 2014*, a cura di R. Basili, A. Lenci e B. Magnini, vol. 1. Pisa.

Ciotti, F. 2014a. «Tematologia e metodi digitali: dal markup alle ontologie». In *I cantieri dell'italianistica. Ricerca, didattica e organizzazione agli inizi del XXI secolo. Atti del XVII congresso dell'ADI Associazione degli Italianisti (Roma Sapienza, 18-21 settembre 2013)*. Roma.

Cugusi, P. 1996. *Aspetti letterari dei Carmina latina epigraphica*. 2ª ed. Bologna: Pàtron.

Lefèvre, M. 2006. «Per un profilo storico della critica tematica». In *Temi e letture*, a cura di C. Spila, 11–29. Roma.

Mastandrea, P. 2015. «Archivi elettronici di poesia latina e opzioni multiple di ricerca intertestuale». *Semicerchio* 53:60–69.

— , cur. 2016. *Risorse digitali e strumenti collaborativi per le Scienze dell'Antichità , Atti del convegno (Venezia 23 ottobre 2014)*. In press. Venezia 2016.

Mastandrea, P., e L. Tessarolo. 2014. «Da Musisque Deoque a Memorata Poetis. Le vie della ricerca intertestuale». In *Collaborative Research Practices and Shared Infrastructures for Humanities Computing (Proceedings of 2nd AIUCD Annual Conference)*, 69–80. Padova.

Pellini, P. 2. 2008. «Critica tematica e tematologia: paradossi e aporie». *Allegoria* 58:25–33.

Pistellato, A., cur. 2014. *Memoria poetica e poesia della memoria: la versificazione epigrafica dall'antichità all'umanesimo*. Studi di archivistica, bibliografia, paleografia 3. Venezia.

Annotazione tematica di testi poetici delle origini: verso l'uso di ontologie formali e geo-referenziazione

Daniele Silvi, Università di Roma 'Tor Vergata', silvi@lettere.uniroma2.it
Fabio Ciotti, Università di Roma 'Tor Vergata', fabio.ciotti@uniroma2.it

1 Introduzione

La critica tematica negli ultimi anni è tornata alla ribalta del dibattito teorico negli studi letterari. Il presente paper propone di inserire l'analisi tematica nel contesto di un approccio computazionale agli studi letterari, in cui l'analisi tematica converge con i metodi di ricerca e con gli strumenti delle Digital Humanities e con quelli dell'epigrafia digitale. In questo contesto, riteniamo di particolare rilievo l'adozione di metodi di annotazione semantica basati su ontologie e di tecniche di geo-referenziazione dei riferimenti spaziali nei testi, al fine di ampliare il quadro offerto dalla strategia di annotazione tematica dei documenti.

2 Contesto e stato dell'arte

Il lavoro che presentiamo si inserisce nell'ambito del contributo apportato dall'unità di ricerca dell'Università di Roma "Tor Vergata" al progetto di ricerca PRIN "Memoria poetica e poesia della memoria. Ricorrenze lessicali e tematiche nella versificazione epigrafica e nel sistema letterario", coordinato dall'Università Cà Foscari di Venezia. Il lavoro da noi condotto nel quadro del progetto complessivo ha riguardato l'annotazione tematica della poesia delle "Origini" in lingua volgare, considerata anche a partire dai suoi labili e mobili confini con la coeva produzione latina. Nella prima fase del progetto sono stati reperiti, digitalizzati e tematizzati tutti i testi poetici di forma breve (intendendo con questa come estensione massima la forma "sonetto") del XIII secolo.

La creazione del database testuale (e tematico) ha altresì permesso la creazione di un parallelo database bibliografico della poesia delle origini in volgare, che costituisce – di fatto – un aggiornamento ed un arricchimento di quello predisposto dall'Opera del Vocabolario Italiano (OVI) per la creazione del TLIO. Questo database bibliografico, che è in corso di ulteriormente aggiornamento nella fase attuale del progetto, verrà trasformato in una forma più evoluta e condivisa dalla comunità scientifica internazionale, ed esposto in formato LOD (Linked Open Data).

Il lavoro sul corpus è attualmente in fase di estensione, integrando il database testuale collettivo, con i testi (poetici, di forma breve) del XIV sec., fino all'anno 1375, che sono

già stati reperiti e digitalizzati. Il lavoro sui corpora testuali di interesse è partito dall'analisi e dall'integrazione della bibliografia fornita dall'Istituto Opera del Vocabolario Italiano (OVI). Successivamente si procederà (dopo averli inseriti nel database già esistente) alla marcatura tematica degli stessi. Tale lavoro è da considerarsi in continua evoluzione per la possibilità concreta di poterlo estendere progressivamente a tutta la produzione poetica italiana.

3 Metodologia

La creazione di un gruppo di lavoro stabile si è resa necessaria per garantire la corretta progressione del lavoro sui vari fronti in cui esso si articola. Il gruppo di lavoro composto da quattro esperti del settore è stato istruito secondo dei criteri metodologici precisi, che qui accenniamo: Criteri storiografici (selezione di autori del '200 secondo bibliografia OVI, criterio 'tematico' per le Tenzoni, scelta delle sole forme poetiche brevi, metricamente corrispondenti o inferiori al sonetto); Criteri di attestazione bibliografica (scelta dell'edizione critica più recente segnalata in OVI, riferimento alle edizioni autorevoli dei testi e degli autori in esame, analisi di casi compositivi particolari come le Tenzoni); Criteri di lavorazione (marcatura in equipe e confronti ciechi dei testi lavorati in parallelo, selezione di giovani laureati nel settore delle digital humanities e divisione dei compiti tra essi, riunioni e confronti mensili sui risultati e le problematiche emerse e eventuale proposizione di queste all'unità centrale del Prin di riferimento); Criteri biografici (classificazione degli autori secondo le seguenti norme: Nome + patronimico (Chiaro Davanzati), Nome + toponimico (Bonvesin da la Riva), Nome + soprannome (Paolo dell'Abaco), conseguente ricostruzione di un indice dei nomi).

Queste indicazioni sono servite a popolare l'attuale database di autori del '200 e sono state il punto di partenza per lo sviluppo delle indicazioni metodologiche che animano la nostra proposta di modellizzazione formale degli aspetti della tematologia mediante l'uso di ontologie e strumenti di georeferenziazione. Queste ulteriori indicazioni verranno esplicitate e commentate durante l'esposizione del progetto e dettagliate in sede di stesura di atti.

Va inoltre notato che tutta la metodologia soggiacente a questa proposta sarà estesa anche ai testi dell'epigrafia latina, già trattati dal gruppo di ricerca CNR di Pisa.

4 Tematologia e ontologie

L'attività di ricerca condotta nell'ambito del progetto ha propiziato anche un'approfondita riflessione metodologica e teorica sui metodi di annotazione adottati e sulle possibilità di sviluppo degli stessi sulla base di sempre più efficaci infrastrutture tecnologiche.

Per quanto riguarda l'aspetto informatico, il linguaggio XML e il suo soggiacente modello di dati non sono in grado di fornire la flessibilità ed efficacia necessarie per la rappresentazione formale della complessa rete di fenomeni testuali e extra-testuali da noi individuati, e la predisposizione di strumenti di analisi e interrogazione di tali informazioni. Le tecnologie e i formalismi che sono attualmente rubricati sotto l'etichetta di Web Semantico cercano di affrontare proprio questo genere di esigenze di elaborazione delle risorse informative digitali, ereditando la tradizione della knowledge representation e dei sistemi

inferenziali sviluppati nel contesto della ricerca in Intelligenza artificiale degli scorsi decenni e rifunzionalizzandoli in vista della loro applicazione a risorse informative in rete. In breve, l'idea alla base del Web Semantico consiste nell'associare alle risorse informative disponibili in varie forme sul Web una descrizione formale del loro significato, cosicché un programma possa elaborare tale informazione in modo significativo (cioè tenendo conto di che cosa essa significhi), dedurne conseguenze, e generare automaticamente nuove informazioni. Questo nuovo orizzonte apre prospettive interessanti anche per la ricerca di ambito letterario, soprattutto nell'ambito della tematologia comparata e nello studio sui fenomeni dell'intertestualità, come testimoniato dalla letteratura internazionale più recente nel campo delle Digital Humanities.

La nostra proposta, dunque, consiste nel procedere a una modellizzazione formale dei concetti di motivo, tema, topos e dei concetti connessi, nonché delle modalità con cui essi circolano nei macrotesti poetici di diversi periodi culturali. L'idea è quella di definire concettualmente l'ambito della tematologia, mediante l'uso di ontologie formali. A nostro avviso le tecnologie ontologiche del Web Semantico forniscono un apparato strumentale idoneo alla creazione di un repertorio tematico che non sia una pura enumerazione o classificazione gerarchica di termini, ma che al contrario si organizzi per multiple stratificazioni tipologiche e al contempo permetta una ricca rete di relazioni orizzontali tra classi e tra istanze di temi e motivi. Questa linea di indagine si congiunge con una formalizzazione in OWL del repertorio tematico, già condotta dalla unità di ricerca dell'ILC di Pisa.

5 Sviluppi: annotazione geografica e geo-referenziazione

Un ulteriore obiettivo del progetto riguarda l'uso di metodi geo-referenziazione. Sono in fase di sperimentazione una serie di strumenti per arricchire la banca-dati testuale della poesia in volgare già descritta mediante annotazione di luoghi che vengono, a loro volta, geo-referenziati. In tale modo i testi potranno essere organizzati su base geografica, consentendone una lettura a partire dallo spazio menzionato che potrà fornire interessanti chiavi di lettura della poesia delle origini (e non solo).

Bibliografia

Boschetti, F. 2015. «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 2015)*. Torino.

Ceserani, R. 2008. «Il punto sulla critica tematica». *Allegoria* 20 (58): 25–33.

Ceserani, R., M. Domenichelli e P. Fasano. 2006. *Dizionario dei temi letterari*. Torino: UTET.

Ciotti, F. 2011. «La rappresentazione digitale del testo: il paradigma del markup e i suoi sviluppi». In *La macchina nel tempo: studi di informatica umanistica in onore di Tito Orlandi*, a cura di L. Perilli e D. Fiormonte. Firenze: Le lettere.

— . 2014b. «Tematologia e metodi digitali: dal markup alle ontologie». In *I cantieri dell'Italianistica. Ricerca, didattica e organizzazione agli inizi del XXI secolo. Atti del XVII congresso dell'ADI – Associazione degli Italianisti (Roma Sapienza, 18-21 settembre 2013)*, a cura di B. Alfonzetti. Roma: Adi editore.

Ciotti, F., M. Lana e F. Tomasi. 2015. «TEI, Ontologies, Linked Open Data: Geolat and Beyond». *Journal of the Text Encoding Initiative* 8. https://jtei.revues.org/1365.

Dionisotti, C. 1967. *Geografia e Storia della Letteratura Italiana*. Torino: Einaudi.

Fiorentino, F., e C. Solivetti. 2013. *Letteratura e geografia. Atlanti, modelli, letture*. Macerata: Quodlibet.

Gruber, T. R. 2009. «Ontology». In *Encyclopedia of Database Systems*, a cura di L. Ling Liu e M. Tamer Ozsu. New York / London: Springer-Verlag.

Guarino, N. 1995. «Formal ontology, conceptual analysis and knowledge representation». *Int. J. Hum.-Comput. Stud.* 43 (5–6): 625–640. doi:10.1006/ijhc.1995.1066.

Heath, T., e C. Bizer. 2011. «Linked Data: Evolving the Web into a Global Data Space». In *Synthesis. Lectures on the Semantic Web: Theory and Technology I*, 1–136.

Khan, F. 2016. «Restructuring a Taxonomy of Literary Themes and Motifs for More Efficient Querying». *MATLIT* 4.

Lefevre, M. 2003. «Tema e motivo nella critica letteraria». *Allegoria* 15:45.

Luzzatto, S., e G. Pedullà. 2010. *Atlante della letteratura italiana*. Torino: Einaudi.

Segre, C. 1985. *Avviamento all'analisi del testo letterario*. Torino: Einaudi.

Sollors, W. 1993. *The Return of Thematic Criticism*. Cambridge(MA) / London: Harvard University Press.

ASED – Annotazione Semantica per Edizioni Digitali

Andrea Bolioli, CELI, andrea.bolioli@celi.it
Riccardo Tasso, CELI, tasso@celi.it
Roberto Rosselli Del Turco, Università di Torino, roberto.rossellidelturco@unito.it

1 Abstract

In questo intervento presentiamo alcuni dei risultati del progetto di ricerca ASED (Annotazione Semantica per Edizioni Digitali), il cui obiettivo principale è stato lo studio della realizzabilità di un "sistema software" per creare e gestire le edizioni digitali online (in particolare le SDE), che consenta l'interoperabilità tra modelli e strumenti di annotazione dei contenuti basati su TEI XML (per quanto riguarda il contenuto testuale) e modelli e strumenti di annotazione semantica basati su RDF, o più in generale i modelli del semantic web, come descritti ad es. in (W3C OADM 2013).

Lo studio di fattibilità è stato cofinanziato dalla Regione Piemonte nel 2015. Il sito web di presentazione sintetica del progetto si trova all'indirizzo https://ased.celi.it.

L'analisi dello stato dell'arte ha evidenziato una grande varietà nei metodi, nelle infrastrutture software e nelle interfacce grafiche di fruizione utilizzate per la realizzazione e la gestione delle SDE, come attestato anche in (Pierazzo 2015c).

Inoltre, nonostante il tema del rapporto tra TEI XML e ontologie sia studiato da parecchi anni, a partire da Tummarello, Morbidoni e Pierazzo (2005) per arrivare a Ciotti e Ciula (2013) o Eide (2014), le edizioni online che presentano queste caratteristiche, cioè integrano in qualche modo questi standard, sono ancora un numero esiguo e presentano differenze significative.

Per capire i motivi di questa situazione e immaginare quali possono essere gli sviluppi futuri, abbiamo cercato di analizzare in dettaglio alcuni esempi recenti rilevanti, oltre a studiare la letteratura scientifica relativa a questi temi. Tra gli esempi rilevanti, in questo intervento presentiamo un confronto tra l'edizione digitale del Codice Pelavicino (Salvatori et al. 2016), Burckhardt Source (Di Donato e Müller 2013), Clavius on the Web (Abrate et al. 2014), le Lettere di Vespasiano da Bisticci (Tomasi 2013b). Nel confronto abbiamo focalizzato l'aspetto per noi centrale, cioè l'annotazione del testo, che si presenta come manuale o automatica, inline o stand-off, individuale o collaborativa, semantica o non semantica. Tra i tipi di annotazione e di linking semantico, quella delle Named Entities (in particolare le persone e i luoghi) rappresenta uno dei punti in comune tra il mondo delle edizioni digitali ed il mondo della linguistica computazionale, come attestato ad esempio in Frontini e Ganascia (2015).

2 Presentazione

In questo intervento presentiamo le caratteristiche comuni e le differenze nella creazione delle annotazioni, le modalità di visualizzazione e distribuzione, le possibilità di accesso, manuale o "programmatico" (ad esempio tramite API). Oltre ad analizzare gli aspetti tecnici e metodologici nella creazione e gestione delle edizioni digitali con annotazioni semantiche, affrontiamo infine il tema della sostenibilità economica di queste attività, e di eventuali "modelli di business" delle SDE (online).

Bibliografia

Abrate, M., et al. 2014. «Sharing Cultural Heritage: the Clavius on the Web Project». In *Proceedings of LREC 2014*, 627–634. http://www.lrec-conf.org/proceedings/lrec2014/pdf/368_Paper.pdf.

Ciotti, F., e A. Ciula, cur. 2013. *The Linked TEI: Text Encoding in the Web – Abstracts of the TEI Conference and Members Meeting 2013: October 2-5*. Rome. http://digilab2.let.uniroma1.it/teiconf2013/abstracts.

Di Donato, F., e S. Müller. 2013. «Biblioteche digitali semantiche. Il progetto Burckhardtsource.org». *Bibliotime* 16 (1). http://www.aib.it/aib/sezioni/emr/bibtime/num-xvi-1/didonato.htm.

Eide, Ø. 2014. «Ontologies, data modeling, and TEI». *Journal of the Text Encoding Initiative*. https://jtei.revues.org/1191.

Frontini, C. B., Francesca, e J.-G. Ganascia. 2015. «Domain-adapted named-entity linker using Linked Data». In *Proceedings of the Workshop on NLP Applications: Completing the Puzzle co-located with the 20th International Conference on Applications of Natural Language to Information Systems (NLDB 2015)*. http://ceur-ws.org/Vol-1386/named_entity.pdf.

Kurland, P. B., e R. Lerner, cur. 1987. *The Founders' Constitution*. Chicago: University of Chicago Press. http://press-pubs.uchicago.edu/founders.

Pierazzo, E. 2015c. *Digital Scholarly Editing: Theories, Models and Methods*. Ashgate Publishing Ltd. https://hal.inria.fr/hal-01182162/document.

Salvatori, E., et al. 2016. «Codice Pelavicino. Edizione digitale». Online. http://pelavicino.labcd.unipi.it.

Tomasi, F. 2013b. *Vespasiano da Bisticci, Lettere*. University of Bologna. ISBN: 9788898010110. doi:10.6092/unibo/vespasianodabisticciletters.

Tummarello, G., C. Morbidoni e E. Pierazzo. 2005. «Toward Textual Encoding Based on RDF». In *Proceedings of ELPUB 2005 Conference on Electronic Publishing, Kath. Univ. Leuven, June*. Leuven: Kath. Univ. Leuven. http://elpub.scix.net/data/works/att/206elpub2005.content.pdf.

W3C Open Annotation Community Group. 2013. «W3C Open Annotation Data Model». Online. http://www.openannotation.org/spec/core.

Strumenti e Architetture
Tools and Architectures
Posters

A Semantic Infrastructure for Scientific Manuscripts

Sahar Aljalbout, CUI-Université de Genève, saharjalbout@gmail.com
Giuseppe Cosenza, CUI-Université de Genève, cs.giuseppe@gmail.com
Gilles Falquet, CUI-Universitéde Genève, Gilles.Falquet@unige.ch
Luka Nerima, CUI-Universitéde Genève, Luka.Nerima@unige.ch

1 Introduction and Requirements

Several projects in the digital humanities field are producing rich digital corpora. Most of these precious data are published using standard techniques. However, the potential of these corpora to generate new knowledge has not yet been fully exploited, mainly because the skills of humanists in the field of knowledge engineering (KE) are generally limited. According to the results of a study of humanists' requirements, we have designed a knowledge representation and management model for digital corpora of manuscripts. We present an advanced infrastructure that is intended to help digital humanist deal with knowledge intensive tasks in the field of scientific manuscript studies. For this purpose, we propose a set of evolving interconnected knowledge resource (ontologies, terminologies, taxonomies, manuscripts, transcriptions), that represent the current state of our knowledge about a corpus of manuscripts ranging from notes taken from the back of notebooks, loose papers, envelops and invitation cards. We applied this model to the scientific manuscripts of *Ferdinand de Saussure* but of course it could be applied to other authors' manuscripts.

Ferdinand de Saussure (1857-1913) is considered as a "formidable linguist" (Joseph 2012) first of all for his works in general linguistics as well as for his contributions in the rather more exclusive field of comparative grammar. However, Saussure published very little. For instance, he never published the theory he developed in the course of general linguistics he taught three times and which is considered as the work of his life. It is on the basis of lecture notes of his students that the book Course in General Linguistics (Cours de linguistique générale CLG) was published in 1916. The legacy of Saussure is fortunately not limited to these monographs but includes a set of about 50,000 handwritten pages deposited in the libraries of Geneva (Bibliothèque de Genève), Paris and Harvard. Recently, all pages in the Geneva library were photographed using a high definition digital camera (about 46.000 photos).

According to a study of the Saussureans' requirements, several needs were detected. Saussurean scholars complain about the exhausting time required to find specific information because they need to read thousands of manuscripts to find the desired one. In fact, the Library of Geneva owns a catalog that gives a shallow description of the manuscripts categories but usually this description is insufficient to be sure that a requested document contains the desired information. That's because the description is too general and does

not take into consideration the detailed thematic categorization of the manuscripts. That means that in the worst case, a Saussurean have to read all categories content in order to find the relevant one.

Furthermore, one of the most important aims of the Saussurean scholars is to uncover spatial and temporal entities related to the manuscripts. In particular, for the majority of the manuscripts, we know neither their date nor their place of writing. This of course hinders the historical or epistemological research works such as the establishment of a clear sequence of ideas in Saussure's work. Thus, it is of primary importance to create a method able to tag each manuscript with an inferred time and space. It may be inferred by following the bibliographic references, names, events, and institutions that Saussure used to quote in his scriptures.

From all these claims and more, the initial idea of our project was to combine the cultural heritage and the knowledge engineering field. Our aim in this research is to obtain a knowledge representation and management model for the digital edition of large corpora of original works (manuscripts). One of the key concept of the proposed research is that a digital edition of corpus of manuscripts is not a single document containing scholarly transcription of manuscripts. It is a set of interconnected resource (manuscripts, transcriptions, terminologies, ontologies) that represents our knowledge about manuscripts.

2 State of the Art

After surveying the state of the art, we noticed that there exist many systems, generally equipped with a web interface to access digital manuscript collections. The "Nietzsche Source[1]" project is dedicated to the scholarly publication of F. Nietzsche's works. The system contains a critical edition of Nietzche's works by Colli and Montinari and digital photographs of the Nietzschean corpus (first editions of his works, manuscripts, his correspondence, biographical documents). Nietzsche Source is often cited as one of the best sites created, thanks to the layout of the critical edition. It may be noted however that there is no link between the critical editions and digital reproductions. The Bentham Project[2] aims to transcribe the manuscripts of the philosopher and jurist Jeremy Bentham. The approach is based on crowdsourcing, where anyone can contribute to the transcript. The transcripts are then validated by experts. Transcripts are encoded with TEI[3] and the entire interface of the site is based on MediaWiki[4] with a modified skin. Ultimately, the developed software tools should be made available to other projects. Some researchers have adopted the ontological approach to improve the access quality to the manuscript content. This method consists in associating annotations or information extracted manually or automatically to the manuscripts images. They used such techniques in the Sharing Ancient Widsoms (SAWS) project in which semantic information are extracted from TEI documents (Jordanous, Stanley, and Tupman 2012). Other projects (Agosti, Ferro, and Orio 2005) worked on providing researchers with innovative ways for accessing digital manuscripts, sharing and transferring knowledge in a digital environment. Hence, relationships between images of manuscripts

[1] http://www.nietzschesource.org
[2] http://www.ucl.ac.uk/Bentham-Project
[3] http://www.tei-c.org/index.xml
[4] https://www.mediawiki.org/wiki/MediaWiki

can be rendered using annotations. A taxonomy of links was provided in this study that indicate the different relationships that can connect digital objects.

The study of digital scientific or technical documents has attracted and is still attracting many theoretical and applied research works. Although they do not address historical documents, these works provide interesting concepts and techniques to automatically classify, analyze, and retrieve the scientific contents of these documents. For instance, (Constantin et al. 2016) and propose ontologies of scientific discourse elements (hypothesis, results, definition, etc.) that are intended to precisely tag scientific texts. A survey of models and techniques for processing scientific discourses can be found in (Shum et al. 2010). Numerous techniques have been developed to extract terms, named entities and relationships from texts. See (Wong et al. 2012) for a survey of these techniques, applied to ontology learning and (Fernandez & Motta, 2012) for a description of semantic indexing techniques applied to information retrieval.

The notion of virtual document (Ranwez and Crampes 1999) and virtual hyperbooks provides another example of the possible connections between a set of document fragments and an ontology (or multiple ontologies). In this case the ontology not only indexes the documents but it serves as a connecting hub to interconnect documents when constructing new derived documents. In the infrastructure we present here, we have adopted a derived document model that is directly inspired by these works.

Since the manuscript knowledge base must comprise several knowledge resources, it is worth mentioning current research work in the domain of knowledge repository management. The increasing importance of ontology engineering and semantic processing has led to the design and development of systems to store and/or retrieve collections of ontologies.

3 A Semantic infrastructure for Scientific Manuscripts

The digital edition of a corpus of manuscripts is not a single document containing a scholarly transcription of the manuscripts. It is in fact a set of evolving interconnected knowledge resources which represents the current validated state of our knowledge on manuscripts and can be one of the following: manuscripts, transcriptions, scholarly annotations, related publications, related terminologies, ontologies, taxonomies. Figure 1 refers to the UML class diagram of our infrastructure model (the attributes have been removed for the sake of clarity)

In order to design our model, we first identified and categorized the resources that we want to represent and the links between these resources. This model is an aggregation of several components. In the following, we list the relevant ones:

1. Manuscripts, their transcriptions and related scientific documents
2. Knowledge resources: ontologies, taxonomies and terminologies
3. Temporally entities
4. Linking structure: Manuscript-to-Manuscript, Manuscript-to-ScientificDocuments, Manuscripts-to-Knowledge.

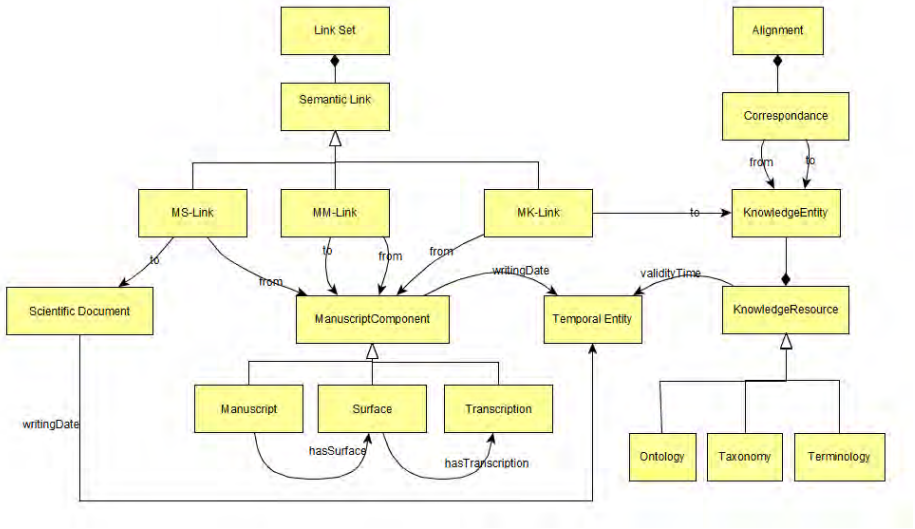


Figure 1: Infrastructure Class Diagram

According to the humanists needs and from the objects and relations indicate above, we will create different provided services. We define a service as an entity that takes as input one or multiple resources and provoke changes on these resources or generate an output. The principal among these are: adding manuscripts and transcriptions, importing knowledge resources, semantic indexing of texts with multiple ontologies, creation and computation of semantic links, information and knowledge retrieval, generation of derived documents.

References

Agosti, M., N. Ferro, and N. Orio. 2005. «Annotating Illuminated Manuscripts: an Effective Tool for Research and Education». In *Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL 2005), June 7-11, 2005, Denver, Colorado, USA*, 121–130. ACM.

Brero, M. 2013. «Système de visualisation, d'annotation et de transcription des manuscrits numérisés de Ferdinand de Saussure». MA thesis, Faculty of Sciences, University of Geneva. http://cui.unige.ch/~nerima/saussure/master_thesis_brero_2013.pdf.

Constantin, A., et al. 2016. «The Document Components Ontology (DoCO)». *Semantic Web* 7 (2): 167–181. http://content.iospress.com/articles/semantic-web/sw177.

Cosenza, G. 2015. «Tra terminologia e lessico: i percorsi di pensiero di F. de Saussure». PhD thesis, University of Calabria.

Gambarara, D., and M. P. Marchese, eds. 2013. *Guida per un'edizione digitale dei manoscritti di Ferdinand de Saussure*. Alessandria: Edizioni dell'Orso.

Ghoula, N., G. Falquet, and H. Nindanga. 2013. «A meta-model and ontology for managing heterogenous alignment resources». In *Proceeding of KDO 2013: First Workshop on Knowledge Discovery in Ontologies, At Atlanta, USA. The 2013 IEEE/WIC/ACM International Conference on Web Intelligence*, 167–170. http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6690720.

Ide, N., and J. Véronis, eds. 1995. *Text Encoding Initiative: Background and Context*. Dordrecht: Kluwer.

Jordanous, A., A. Stanley, and C. Tupman. 2012. «Contemporary transformation of ancient documents for recording and retrieving maximum information: when one form of markup is not enough». In *Proceedings of Balisage: The Markup Conference 2012*, vol. 8. Balisage Series on Markup Technologies. doi:10.4242/BalisageVol8.Jordanous01. http://www.balisage.net/Proceedings/vol8/html/Jordanous01/BalisageVol8-Jordanous01.html.

Joseph, J. E. 2012. *Saussure*. Oxford: Oxford University Press.

Ranwez, S., and M. Crampes. 1999. «Conceptual Documents and Hypertext Documents are two Different Forms of Virtual Document». In *Workshop on Virtual Documents Hypertext Functionality and the Web of the 8th Intl World-Wide Web Conference*. Toronto. http://www.cs.unibo.it/fabio/VD99/ranwez/ranwez.html.

Saussure, F. d. 1974 [1922]. *Cours de linguistique générale, publié par Charles Bally et Albert Séchehaye avec la collaboration de Albert Riedlinger. Édition critique preparée par Tullio de Mauro*. Ed. by T. De Mauro. Paris: Payot.

Shum, S. B., et al. 2010. «Scientific Discourse on the Semantic Web: A Survey of Models and Enabling Technologies». Under review, *Semantic Web Journal: Interoperability, Usability, Applicability*. http://www.semantic-web-journal.net/content/scientific-discourse-semantic-web-survey-models-and-enabling-technologies.

Researches on the *Clitophon*

Pietro Bertocchini, Università di Bologna, pietro_berto@hotmail.it

1 Status Quaestionis

The *Clitophon* is a dialogue that belongs to the *corpus* of Platonic writings. However, it has been suspected of being spurious by many scholars. The modern debate began in 1809 when Schleiermacher took position against its authenticity. Since then, a lot of contributions on the topic have been released and a great number of hypothesis of attribution have been formulated. It is worth mentioning at least Grote's thesis (1865), which speculates that the dialogue had been an introduction (later rejected) to the *Republic*, as well as Kunert's essay (1881), which, for the first time, argued that the attack directed towards Socrates within the *Clitophon* was actually to be understood as directed towards the Socratic Antistenes.

It was starting from this hypothesis that Slings (1999) formulated his proposition about the *Clitophon* in the end of his detailed commentary. According to him, the dialogue could be a literary *pamphlet*, i.e. a criticism of a specific genre of the Socratic literature that downgraded the philosophical research to a mere act of προτρέπειν (exhorting).

On this basis, Slings claims to be persuaded of the authenticity of the text, though not without some reservations. Recently, there have been more scholars who have argued in favour of the Platonic authorship than not (cf. e.g. Bowe 2007). However, they have simply repeated – with some minor changes – the old interpretation of Socrates' attack as a merely symbolic one.

Neither before nor after Slings' commentary (1999) has there been work that analyses in depth the language and style of the dialogue as an interesting field of inquiry for its uncertain attribution. Nonetheless, such an object seems to be the most reliable to base *Clitophon*'s accusation or defence on. I have therefore decided to focus my research on the linguistic and stylistic aspects of the text, aiming at proposing a new and updated analysis, alternative to Slings' one.

Both qualitative and quantitative analysis were conducted in my Master degree's thesis. Herein I will especially concentrate on the quantitative aspect, though my research has been, so far, rather focused on the qualitative one.

2 Qualitative Analysis

In the first place I approached the problem in a traditional way. In fact, I produced a commentary in which I stopped at every passage of the dialogue that was, each one for a different reason, suspected of not being Platonic. For every passage, I analysed the arguments - both in favour and against the authenticity - that were raised by the passage itself. This first stage of the research brought to attention that the reasons for concern are arguably more numerous and more serious than Slings had pointed out in his 1999's commentary: some verb

phrases, single terms, or utterances of various lengths were found to be in contrast with the Platonic language as it is known from the other writings of the *corpus*. Moreover, a limited number of odd passages were identified as possibly caused by some clumsy imitation of other passages belonging to undoubtedly Platonic dialogues, and such a procedure could hardly be ascribed to Plato.

Among the various aspects touched upon in the commentary, it is worth mentioning the problem of the potential date of the dialogue within Plato's production. On this point, the conclusion that had already been reached by others – i.e. that the *Clitophon* needs to be grouped with the dialogues of the last period (*Sophist, Statesman, Philebus, Timaeus, Critias, Laws*) – appeared to be confirmed and, in some cases, even reinforced by new evidence. Yet, the consistency of language and style thus found with these other Platonic writings did not seem sufficient to erase the suspicion of inauthenticity.

3 Quantitative Analysis

Even though the analysis on the single dubious passages is far from being concluded (as therefore is the authenticity's issue), I decided to undertake, in parallel to it, some stylometric researches aimed at gathering new evidence to be compared with that previously found[1]. Though during the qualitative analysis I had taken into consideration some specific passages, at this second stage I tried to assess the overall features of the language and style. In other words, I looked for unperceived stylistic tics, typical constructions, and distinguishing marks, which even the most skilful forger would have trouble to perfectly reproduce. Had the *Clitophon* proved to be significantly different from the norm outlined by the other dialogues, it would have emerged as important evidence against authenticity.

In order to be fully reliable, such a test should compare the dialogue under debate with all the other writings of the Platonic *corpus*[2], or even involve one or more contemporary authors as terms of comparison (e.g. Xenophon, the orators). At this point of the research, however, this was not possible. Nonetheless, as an initial sample for developing some exploratory researches, I was able to gain access to a *corpus* composed of 16 dialogues[3], that is about one-third of the Platonic writings (having taken into account the different lengths of the texts). Among these, 13 are normally considered authentic, while 3 (*Alcibiades II, Hipparcus, Rival Lovers*) are believed to be spurious by most scholars. Such a *corpus* was edited, on behalf of the "Laboratoire d'Analyse Statistique des Langues Anciennes" (LASLA) based in Liège, by Mauro Siviero, who manually[4] assigned information to each word of every di-

[1] Platonic scholars long ago discovered the potentiality of statistical procedures. Yet, they have been mainly employed in the investigation of the long debated problem of the dialogues' internal chronological order (cf. e.g. Campbell 1867, Ledger 1989, Brandwood 1990). As far as Plato is concerned, it is rare to see such procedures applied to questions of authorship attribution. However, an example of this can be found in Ledger (1989), who, among other important contributions, signalled that the *Hippias Major* and the *Rival Lovers* were stylistically closer to Xenophon's writings than to Plato's ones.

[2] On the contrary, a comparison with a *corpus* that comprehended only the dialogues purportedly belonging to the same period (*Sophist, Statesman, Philebus, Timaeus, Critias, Laws*) would have proved to be useless for the question of authenticity. In fact, such a comparison would probably confirm some sort of resemblance between these texts and the *Clitophon*. Yet, as stated earlier, such a result would not be decisive.

[3] *Euthyphro, Apology, Crito, Phaedo, Cratylus, Theaetetus, Sophist, Statesman, Parmenides, Philebus, Symposium, Phaedrus, Alcibiades I, Alcibiades II, Hipparchus, Rival Lovers*.

[4] The manually performed lemmatisation is one of the main qualities of LASLA's *corpus*, together with

alogue, such as a) its lemma (attributed on account of LSJ9), and b) a symbol representing the grammatical category (or POS: *Part Of Speech*) to which the word belonged[5]. Once the same lemmatisation process was implemented for *Clitophon*'s text, it was possible to recompose a database with homogeneous features, comprehensive of 17 dialogues. Some new and automatically deducible information was then added to the already given one: c) word's length; d) POS' variability or invariability; e) lemma's frequency within LASLA's *corpus*.

As this preparatory phase was concluded, I identified ten indicators[6] in order to measure if and how *Clitophon*'s behaviour was inconsistent with the remaining 16 dialogues. Since the analysed features were unmarked linguistic traits, which are involuntary for the most part, it seems that the neutrality of the choice of the phenomena to be analysed has been partly granted. Even so, some part of subjectivity is clearly inescapable. The ten indicators can be summarised as follows:

1. Lemmata variety index.

2. Percentage of the six most frequent lemmata within LASLA's *corpus*.

3. Average percentage of lemmata belonging to some specific POS (substantives, adjectives, verbs, adverbs) within LASLA's *corpus*.

4. Coexistence ratio of variable and invariable words.

5. Average length (n° of letters) of the words.

6. Coexistence ratio of ὅς and ὅστις.

7. Index of dissimilarity between POS frequency distribution of each dialogue and of LASLA's *corpus*.

8. Index of dissimilarity between lemmata frequency distribution of each dialogue and of LASLA's *corpus*.

9. Percentage of lemmata non occurring in other dialogues.

10. Percentage of three lemmata sequences non occurring in other dialogues.

the annotation for each word of its respective part of speech (POS). For future research it would certainly be useful to gain access to the databases provided with all the dialogues. A database of this type is, for example, the one edited by Bombacigno as a part of Radice's lexicon (2004). Even though in this case the lemmatisation was performed through semi-automatic procedures, the curators' *équipe* assures to have manually intervened on each case the software was not able to solve.

[5]This information comes from LASLA's official website. A detailed illustration of the criteria adopted by LASLA's group is found in Denooz (1988, II-IV).

[6]Some others indicators were rejected as redundant since they measured the same phenomenon in two different ways.

4 Length factor

The results' analysis and interpretation required taking into account the distortions that could derive from the extreme shortness of the *Clitophon* (1.564 words). In fact, among the writings included in LASLA's *corpus* – together with those normally held to be spurious (*Alcibiades II, Hipparcus and Rival Lovers* which amount, respectively, to 4.211, 2.270 and 2.418 words) – the only one whose length is comparable to *Clitophon*'s length is the *Alcibiades II* (1.650). The *Crito* (4.164 words), the *Euthyphro* (5.187) and the *Apology* (8.786) too can be grouped as relatively short dialogues if compared to the other nine, whose dimensions are radically superior (they range from *Sophist*'s 16.199 words to *Phaedo*'s 22.440). Two different procedures have been followed in order to avoid the potential results' distortions caused by lengths that are so variable. On the one hand, throughout the analyses, I took into consideration a correlation coefficient that measured the relationship between each indicator and the length of each dialogue in terms of number of words. On the other hand, the same statistical analyses which were carried out for the single dialogues were also performed for *tranches* or groups composed by 1.564 words, i.e. groups of the same length of the *Clitophon*.

5 Conclusions

As for the results, it must be reported that the first assumption – i.e. the *Clitophon* presenting features that are inconsistent with the features of the other dialogues – was not confirmed, except for a few cases. In other words, as for the majority of the analysed phenomena, *Clitophon*'s language appears to be consonant with the other dialogues' language[7]. However, the data's analysis brought to light another interesting and unexpected result: in fact, as for different indicators, a special resemblance was noted between the *Clitophon* and the *Rival Lovers, Hipparchus and Alcibiades II* – these three dialogues in the Platonic *corpus* usually being believed to be inauthentic.

Therefore, an attempt was made to verify such similarity through the calculation of a synthetic indicator that reunited the ten indicators adopted before. To do this, it was necessary to transform the gathered data in two different ways. At first, the single indicators' data were transformed in order to measure not the feature itself, but rather the distance from the *Clitophon* in regards to that feature. Such a procedure resulted in a new set of indicators assessing the greater or smaller resemblance of each dialogue to the *Clitophon*. Secondly, a data standardisation was performed aimed at changing the values into new ones expressed in a single unit of measurement. This way they could be synthesized in an index obtained through the calculation of their simple arithmetic average.

It was ultimately chosen to differentiate the result of this new synthetic indicator according to the greater or smaller influence of the length's factor. Thus, I either included or excluded the single indicators from the calculation based on the value they recorded in the correlation coefficient between their values and the number of words (reaching a positive maximum of +1, a negative one of -1 and being null when 0). Hence, the first column of the table below shows the ranking of the dialogues (from the one that was most simi-

[7] For reasons of space, any specific result of the analysis was omitted.

synthesized in an index obtained through the calculation of their simple arithmetic average.

It was ultimately chosen to differentiate the result of this new synthetic indicator according to the greater or smaller influence of the length's factor. Thus, I either included or excluded the single indicators from the calculation based on the value they recorded in the correlation coefficient between their values and the number of words (reaching a positive maximum of +1, a negative one of -1 and being null when 0). Hence, the first column of the table below shows the ranking of the dialogues (from the one that was most similar to the *Clitophon* to the one that was least) calculated as for the nine indicators with a correlation coefficient varying from -0,75 to +0,75. The second column shows the ranking calculated as for the eight indicators varying from -0,50 to +0,50. The third shows the ranking calculated as for the four indicators varying from -0,25 to +0,25. At a glance, each of the three rankings confirms the *Rival Lovers* and the *Hipparchus* as particularly similar to the *Clitophon*, while the *Alcibiades II* is revealed to be not as close to it as expected.

Ranking of the dialogues in order of distance from the Clitophon

	Indicators with correlation coefficents in the range of:		
	-0,75 / +0,75	-0,50 / +0,50	-0,25 / +0,25
Euthyphro	10	12	13
Apologie	9	10	8
Crito	14	14	14
Phaedo	7	4	6
Cratylus	2	2	3
Theaitetus	13	11	11
Sophista	6	7	10
Politicus	11	8	7
Parmenides	15	15	16
Philebus	8	9	12
Symposium	5	5	4
Phaedrus	12	13	9
Alcibiades I	16	16	15
Alcibiades II	4	6	5
Hipparchus	3	3	2
Amatores	1	1	1

Figure 1: Ranking of the dialogues in order of distance from the Clitophon

[8] For reasons of space, any specific result of the analysis was omitted.

6 Ranking of the dialogues in order of distance from the Clitophon

The special resemblance between the *Clitophon* and the probably spurious *Rival Lovers* and *Hipparchus* stands out as an important outcome for the question of authenticity. Such a result encourages to broaden the comparison to all the Platonic works not included in the LASLA's *corpus* and, at the same time, it urges to examine more in depth the other potential similarities between this dialogue, those two, and the other Platonic spuria and dubia. Therefore, the aim of the next stages of the research will be to verify whether the special affinity observed here is indeed a symptom, for the *Clitophon* too, of an un-Platonic origin or, even, of a context of production consistent with that of the *Hipparchus* and the *Rival Lovers*.

References

Bowe, G. S. 2007. «In defense of Clitophon». *Classical Philology* 102 (3): 245–64.

Brandwood, L. 1990. *The Chronology of Plato's Dialogues*. Cambridge: Cambridge U. Press.

Campbell, L. 1867. *The Sophistes and Politicus of Plato*. Oxford: Clarendon Press.

Denooz, J. 1988. *Aristote. Poetica: index verborum, liste de frequence*. Liège: CIPL.

Grote, G. 1865. *Plato and the Other Companions of Sokrates*. London: J. Murray.

Kunert, R. 1881. *Quae inter Clitophontem dialogum et Platonis Rempublicam intercedat Necessitudo*. Greifswald: Gryphiswaldiae.

Ledger, G. R. 1989. *Recounting Plato. A Computer Analysis of Plato's Style*. Oxford: Clarendon Press.

Radice, R., ed. 2004. *Lexicon. I. Plato*. Milano: Biblia.

Slings, S. R. 1999. *Clitophon*. Cambridge: Cambridge U. Press.

All good with the hood? Towards a methodical reconsideration of the co-occurrence analysis

André Bünte, Universität Leipzig, abunte@uni-leipzig.de
Hannes Kahl, Universität Leipzig, hannes.kahl@uni-leipzig.de
Charlotte Schubert, Universität Leipzig, schubert@uni-leipzig.de

Abstract

In our poster we show a new approach for analysing co-occurrences of a pair of terms in a large text corpus. Our prototype uses a different algorithm than the conventional approaches for a co-occurence analysis, which are usually based on the bag-of-words-model. We think that our concept of counting co-occurrences is able to overcome the difficulties of the conventional model thus enhancing the quality of the results. Our main goal is to implement syntactical information as base for counting and for taking the range of reference from the unstructured text data.

1 The Problems of co-occurrence analysis

Conventional co-occurrence analysis as it is commonly used in Digital Humanities lacks of some critical features, which effects its outcome. The bag-of-words model behind the approach fails short to the linguistic elaboration of a sentence, a passage or a whole text. As a result the definition of range for reference is arbitrary and the results are not comparable. Secondly, the presentation of the results - either graphically as a network or textually as a list - does not allow the recognition of the linguistically relevant (semantic and/or syntactic) relationship between a term and its co-occurrents. Thirdly, the application of certain measurements are necessary in order to decrease the amount of results for a gain of usability. Depending on the measure co-occurrences are considered to be significant or not. As most of these measures include probabilistic evaluation the results are less controllable as well as understandable for the common user (Bünte 2011).

2 Approaching unstructured text data with syntactical knowledge

We've put these matters into consideration and started from scratch to develop a different approach. We think that a discrete definition of neighbourhood is not the right way

and should rather be deduced from observance by analysis of syntactical and/or semantical properties of the occurrences in focus (the context). So the results are categorized according to their degree of "neighbourhoodness", which is based on a concept for the determination of distance on grammatical grounds (derived from generative grammar, as detailed description we use Scalise 1986, Moore 1934, Brandenstein 1954). Especially for close "neighbours" also the syntactical relationship of the co-occurrents will be apt for recognition by the user. This method yields some advantages for the evaluation of the results. It allows a better control of the outcome gaining a much higher usability especially in cases of co-occurrents as well as co-occurrences with high frequencies in large corpora. By using grammatical concepts we can partly rely on the syntactic and semantic structures inherent in the texts, which renders the range of reference less arbitrary.

3 How does it work?

For an example we choose a passage in Herodotus book 5.98: "Ἀρισταγόρης δὲ προπλώσας καὶ ἀπικόμενος ἐς τὴν Μίλητον, ἐξευρὼν βούλευμα ἀπ᾽ οὗ Ἴωσι μὲν οὐδεμία ἔμελλε ὠφελίη ἔσεσθαι, οὐδ᾽ ὧν οὐδὲ τούτου εἵνεκα ἐποίεε ἀλλ᾽ ὅκως βασιλέα Δαρεῖον λυπήσειε, ἔπεμψε ἐς τὴν Φρυγίην ἄνδρα ἐπὶ τοὺς Παίονας τοὺς ἀπὸ Στρυμόνος ποταμοῦ αἰχμαλώτους γενομένους ὑπὸ Μεγαβάζου, οἰκέοντας δὲ τῆς Φρυγίης χῶρόν τε καὶ κώμην ἐπ᾽ ἑωυτῶν:" ("Aristagoras sailed before the rest, and when he came to Miletus, he devised a plan from which no advantage was to accrue to the Ionians (nor indeed was that the purpose of his plan, but rather to vex king Darius). He sent a man into Phrygia, to the Paeonians who had been led captive from the Strymon by Megabazus, and now dwelt in a Phrygian territory and village by themselves.") Our software tries to determine the noun phrases such as "τοὺς Παίονας τοὺς ἀπὸ Στρυμόνος ποταμοῦ αἰχμαλώτους γενομένους ὑπὸ Μεγαβάζου, οἰκέοντας δὲ τῆς Φρυγίης χῶρόν τε καὶ κώμην ἐπ᾽ ἑωυτῶν" by some typical patterns such as the use of articles (e.g. articular NP), nonoccurrence of verbs and agreement.[1]. Since the classical period Greek texts show an extensive use of articles. Especially in cases of more complex NPs the initial use of an article can be generally observed. That's why we use articles as a primary mark for the start of a NP. The ending is often marked by the noun in agreement with the article if there are not other nouns following, which agree with the same article and are somehow conjunct. Nouns in the Genitive case tend to be adhesive to other NPs, because they can modify other nouns. Thus they may be recognised as part of another NP, e.g. when appearing between article and the corresponding noun such as in "τοὺς ἀπὸ Στρυμόνος ποταμοῦ αἰχμαλώτους". The Nominative marks generally the subject to a verb or a predicate noun or adjective. Even if there are multiple Nominatives in a sentence, they are in most cases semantically connected. So combining multiple Nominative-NPs may be important for showing very closely related co-occurrents. As in "Ἀρισταγόρης δὲ προπλώσας καὶ ἀπικόμενος ἐς τὴν Μίλητον ἐξευρὼν βούλευμα ἀπ᾽ οὗ Ἴωσι" we can combine the proper name Ἀρισταγόρης with the two participles ἀπικόμενος and ἐξευρὼν and for example recognise also the phrase "βούλευμα ἀπ᾽ οὗ Ἴωσι" and "Ἀρισταγόρης" as very close neighbours. While verbal phrases are recognised separately from the NPs, participles may be seen as part of NPs. As Herodotus uses a lot of NPs (Bakker 2009), this text serves as a

[1]See screenshot of the software http://ecomparatio.net/~khk/venedig/0.png (login: venedig & pw: famosissima)

great scenario for our development. Our idea is that "very close neighbours" belong to the same NP (e.g. as modifiers). A "close neighbourhood" is formed by two or more NPs that are conjunct. This measure has the advantage of showing discrete inherent properties of the text rather than being predefined. As we depend on as much as information as is possible to be derived from the unstructured text data, we need to recognise built in structural information. Therefore we focus on conjunctions, particles, articles, prepositions, etc. In other approaches exactly these parts of speech are dismissed as "stop words". We believe that they contain extreme valuable data for the recognition of the text structure (see Hermann 1923, Denniston 1959). We use them for the definition of the range of reference. As they appear both within complex NPs and between different NPs, these "small words" organise the relationship of the NPs as well as their internal organisation respectively. In our tool the co-occurrence of "Παίονας" and "Στρυμόνος" or their base forms "Παίονες" and "Στρυμών" would be considered as being "very close neighbours", whilst e.g. "Ἀρισταγόρης" and "Παίονας" would be considered as "close neighbours". In the usual way "Ἀρισταγόρης" and "Παίονες" would occur as significant co-occurrence, while "Ἀρισταγόρης" and "Παίονας" would not, as they occur together only once in Herodotus.

4 The poster

On our poster we are representing some screenshots of an online implementation that we've built for experimenting with the new approach. The implementation of the new method for co-occurrence analysis is realised as a browser based program (JavaScript). It is suitable for the computation of small sized corpora up to approximately 300000 words on consumer hardware. Any unstructured text data (containing polytonic Greek characters) can be pasted into a box and will be analysed subsequently. The results will be displayed immediately.[2]

References

Bakker, S. J. 2009. *The Noun Phrase in Ancient Greek. A Functional Analysis of the Order and Articulation of NP Constituents in Herodotus*. Leiden / Boston: Brill.

Brandenstein, W. 1954. *Griechische Sprachwissenschaft. 1, Einleitung, Lautsystem, Etymologie*. Berlin: de Gruyter.

Bünte, A. 2011. «Documentation for the use of the eAQUA function 'explorative search'». *Working Papers Contested Orders* 3:19–32.

Denniston, J. D. 1959. *The Greek Particles*. 2nd ed. Oxford: Clarendon Press.

Hermann, E. 1923. *Silbenbildung im Griechischen und in anderen indogermanischen Sprachen*. Göttingen: Vandenhoeck & Ruprecht.

Moore, R. W. 1934. *Comparative Greek and Latin Syntax*. London: G. Bell & Sons.

Scalise, S. 1986. *Generative Morphology*. Berlin / New York: de Gruyter.

[2]For Screenshots of the results see http://ecomparatio.net/~khk/venedig/1.png & http://ecomparatio.net/~khk/venedig/2.png (login: venedig & pw: famosissima)

Bridging the Gap between Digital Humanities and Philology: A Case Study of Herrera's Poetry

Laura Hernández Lorenzo, University of Seville, lhernandez1@us.es

1 Digital Humanities and Corpus approaches

This paper aims to combine the methodologies of Digital Humanities, Corpus and Computational Linguistics and Philology in order to solve questions which emerge in literary texts.

After decades of struggle, Digital Humanities methodologies and research are finally gaining presence in universities over the world. The arisement of Big Data and the development of Computational systems have encouraged the massive analysis of literary texts, following the new perspectives of 'macroanalysis' (Jockers 2013) or 'distant reading' (Moretti 2007b). These studies -- carried out using statistical procedures and huge amount of texts by disciplines such as Corpus Stylistics or Stylometry – differ from traditional close reading and Textual Criticism used in Philology (Pérez Priego 2011). Believing it is necessary to bridge the gap between Digital Humanities and traditional Philology in the study of literary text, at this work we combine these methodologies using Corpus and Computational Linguistic tools.

2 A Case Study: Fernando de Herrera's Poetry

The objective is to enlighten the textual problems and variants affecting the poetic works of one of the most important poets in Spanish Golden Age, the writer Fernando de Herrera (1534-1597). The writer who was called 'The Divine' published a book of poetry during his life in 1582, titled *Some Works* and known as *H*. This book was carefully prepared by the author, who even revised printed proofs. However, some years after his death, the painter Francisco Pacheco, great admirer of Herrera as a writer, published a new book with Herrera's poetry in 1619 titled *Verses by Fernando de Herrera* and known as *P*. It included new poems by Herrera and different versions of some of the poems published in *Some Works*. Great differences and variants between *Some Works* and *Verses* had been largely discussed by experts and philologists without reaching an agreement about the authenticity of the variants (Cuevas and Herrera 1985): are they final versions (Macrí 1972) or deturpation of texts (Blecua and Herrera 1975)? Certainly, some advances in research were achieved, but it was still necessary to study the entire corpus using the new tools and possibilities offered now-a-days by Corpus Linguistics, Computational Linguistics and Digital Humanities.

3 Methodology and conclusions

The first step was to digitalise the entire corpus of Herrera's poetry, so the text was available in machine-readable form. We selected Blecua's annotated edition (Blecua and Herrera 1975) for this process which not only contains the entire corpus but is the most authorised edition of Herrera's poetry. The text resulted was revised and the spelling mistakes of the OCR were corrected. After that, the text was converted in UTF-8 plain text and a second revision was undertaken to eliminate notes from the original Blecua's edition.

Once the text was prepared, the poems which appeared corrected in P were isolated and the two versions of each poem were collated automatically with Juxta (`http://www.juxtasoftware.org`). After that, instead of using a pre-existent corpus software, a DIY tool was developed on Java, titled *Litcon*. This software offers the habitual features for corpus analysis, such as concordances, wordlists or keywords. It was developed thinking on poetic texts and, therefore, concordances are shown by lines, an option that habitual corpus software do not possess, as they are thought for texts written in prose. Additionally, it has a tool for the automatic generation of a list of concordances using all the words in the document, and another one that compares two texts and extracts common words between them, as well as words appearing only in one of them.

Wordlists of *H* and *P*, in addition to data of common and different words between the two books were used for the statistical and quantitative part of the analysis. The most frequent function words were extracted from both of the wordlists and compared. At the same time, the qualitative analysis was undertaken looking at concordances – one of the basic features of Corpus analysis according to McEnery and Hardie (2012) – of the words included in the variants. Finally, quantitative and qualitative results are contrasted and interpreted taking into account the studies about the matter that experts on Herrera's poetry have produced and most especially, those about Herrera's poetic language by Kossoff (1966) and Macrí (1972), contributing at the same time to these studies, as we already did through a Corpus approach Lorenzo (2016). We will demonstrate how the results point out to an evolution in style from *H* to *P* poems, in both concordances and statistical procedures, in an agreement with the theories of some of the most authorised philologists experts in the field Macrí (1972). To conclude with, this paper bridges the gap between Digital Humanities, Computational tools and traditional Philology in order to solve problems in the study of texts which couldn't be solved without integrating all these disciplines.

References

Blecua, J. M., and F. d. Herrera. 1975. *Obra poética I-II*. Madrid: Real Academia Española.

Chiappini, G. 1985. *Fernando de Herrera y la escuela sevillana*. Madrid: Taurus.

Cuevas, C., and F. d. Herrera. 1985. *Poesía castellana original completa*. Madrid: Cátedra.

Elson, D. K. 2015. «Literature Lifts Up». *Computational Linguistics Linguistic Issues in Language Technology (LiLT) 12* 1:1–4.

Jockers, M. L. 2013. *Macroanalysis: digital methods and literary history*. Urbana: University of Illinois Press.

Kossoff, A. D. 1966. *Vocabulario de la obra poética de Herrera*. Madrid: Real Academia Española.

Lorenzo, L. H. 2016. «The Poetic Word of Fernando de Herrera. An Approach through Corpus and Computational Linguistics». In *CILC2016*, ed. by A. M. Ortiz, vol. 30. EPiC Series in Language and Linguistics. Forthcoming.

Macrí, O. 1972. *Fernando de Herrera*. Madrid: Gredos.

McEnery, T., and A. Hardie. 2012. *Corpus Linguistics: Method, Theory and Practise Cambridge*. Cambridge: Cambridge University Press.

McEnery, T., and A. Wilson. 2005. *Corpus Linguistics*. Edinburgh: Edinburgh University Press.

Moretti, F. 2007b. *La literatura vista desde lejos*. Barcelona: Marbot Ediciones.

Pérez Priego, M. Á. 2011. *La edición de textos*. Madrid: Síntesis.

Schreibman, S. 2004. *A Companion to Digital Humanities*. Oxford: Blackwell.

Siemens, R., and S. Schreibman, eds. 2008. *A Companion to Digital Literary Studies*. Oxford: Blackwell.

Formalisierung digitaler Abbildungen früher Aldinen zur Berechnung der Klassifikation des Satzspiegels, sowie der optischen Buchstabenerkennung

Hannes Kahl, University of Leipzig / University of applied science Erfurt

1 Einleitung

Aldus Manutius war Quelle vieler Neuerungen, die er in seinem Streben nach Weitergabe des hellenistisch Wissens, praktisch einführte. Die herausgehobene Stellung seiner Editionen für die Weitergabe der griechischen Sprache und der Überleiferung der Manuskripte in die frühe Zeit der Renaissance machen seine Druckerzeugnisse für die Alte Geschichte besonders bedeutsam. Die Größe, die es verlangte Editionen vor der Erfindung des kritischen Edierens[1] auf den Weg zu bringen, und damit den Ausgangspunkt des Editionswesens mitzubegründen und die Größe, die es verlangt die typographischen Probleme zu lösen, die mit dem Druck altgriechischer Texte (vor allem in den mischsprachlichen Ausgaben) verbunden sind, ist Grundlage der Bedeutung, die die Aldinen haben[2]. Dieser Bedeutung und der weiteren Konservierung der Artefakte trägt die Praxis der Bibliotheken, ihre wertvollen Aldinenausgaben zu digitalisieren, Rechnung. Ein weit aus kleineres Unterfangen will ich mit meinem Poster zeigen.

Es geht um die Herstellung von Software zur Verarbeitung von digitalen Abbildungen früh standartisierter Druckwerke. Zur Grundlage meiner Arbeit dienen Digitalisate der berühmten Aldinen. Ob das Bestreben der automatischen Dokumentanalyse, Layoutklassifikation und *Optical Character Recognition* für die Verarbeitung der Incunablen und Nachwiegendruckausgaben sinnvoll ist, kann berechtigt bestritten werden. Worum es sich eigentlich handelt ist eine Kunst: die Kunst komplexe visuelle Analysen automatisch durch zu führen und im Vorfeld die richtigen Annahmen zu einer formalen Beschreibung des Abbildes zusammen zu führen. Das Fehlen richtiger Formalisierungen kann daran abgelesen werden, dass die über lange Jahre entwickelte OCR-Methode wenig die eigentlichen

[1] Zum Charakter des Kritischen an den Editionen von Aldus Manutius siehe S. 121 - 123 in Martin Sicherl in „Griechische Handschriften und Aldinen, Eine Ausstellung anläßlich der XV. Tagung der Mommsen-Gesellschaft in der Herzog-August Bibliothek Wolfenbüttel "(Wolfenbüttel, 1978) S. 121-123

[2] Zur Diskussion der Lösung für griechische Typen in dieser Zeit / For a discussion of the solution to greek fonts see Robert Proctor „The printing of Greek in the 15th Century "(Oxford, 1868) und Nicolas Barker „Aldus Manutius and the Development of Greek Script & Type in the Fifteenth Century "(Fordham Univ Press, 1992)

Strukturen der abgebildeten Artefacte berücksichtigt und es an einer funktionierenden Layoutklassifikation völlig fehlt.[3] Ich beschäftige mich in meiner Dissertation mit dem Thema der Formalisierung[4] der Abbildungen der Aldinendrucke, wie sie von den Bibliotheken zur Zeit angeboten werden. Meine bisher erkämpften Ergebnisse will ich nun gern einem breiten Publikum vorstellen, um von der Notwendigkeit meiner Vorgehensweise hören zu lassen und auf die Lücken in der Theorie, die die Computertechnik in die wissenschaftliche Tradition gerissen hat, hinzuweisen. Ich will die Resultate und die technische Umsetzung demonstrieren, um Interessierten die Möglichkeit zu bieten auf meine Umsetzung mit ihren praktischen Notwendigkeiten einzuwirken. Die live Vorführung der Software, die meine Posterpräsentation begleiten wird, soll zugleich eine Beweisführung sein, die die theoretischen Ausarbeitungen stützt.[5]

2 Ansatz: Software und Methoden im Überblick

Die Entwicklung der Informatik, fand in einer Zeit statt, als das grundlegende Medium der Wissenschaft, das typographische Medium, als zentrales Bildungsorgan der Normen der Wissenschaft, bereits duch Neues ersetzt wurde[6]. Daß die Wissenschaftlichkeit dabei nur durch ein erhöhtes Begriffsbewusstsein; die Wissenschaft, als geistige Einheit nur durch eine Übernahme als Begriffstradition weitergeführt werden kann, liegt auf der Hand. Die mediale Transformation ist nicht mehr als Tradition der Artefakte implizit gegeben, sondern muss erneut auf einer abstrakteren Ebene geschehen. Diese Transformation blieb an vielen Stellen aus, so zeichnen sich die Veröffentlichungen der Informatik durch einen Begriffswildwuchs aus, der mit den Begriffen der mathematischen und rechnenden empirischen Wissenschaften bricht. Anstelle der Wissenschaft steht die Informatik, als Mythos der Analyse und Weiterentwicklung als Feld zur Verfügung. Dieses grundlegende Überlie-

[3] Die nötigen Erweiterungen für OCR wurden von Springmann, U. http://www.cis.uni-muenchen.de/~springmann/ (2014), Boschetti et al. (2009) und Federbusch M. (2013) benannt. Zur Wichtigkeit der Layoutklassifikation wurde ich von Springmann unterrichtet (Graz 2015). Die Theorie der Typographie, die für die Aldinen, sowie Fälle der französischen Druckereien und Zacharias Callierges, um einige zu nennen, erheblich ist, wurde von Burnhill beschrieben / For the relevant theory of typography see Peter Burnhill „Type spaces in-house norms in the typography of Aldus Manutius "(London, 2003).

[4] Zur Ableitung der Berechtigung der Fomalisierung aus dem Begriff des Naturgesetzes Michael Hampe „Eine kleine Geschichte des Naturgesetzbegriffs "(Frankfurt am Main, 2007)

[5] Ein Screenshot der Software findet sich unter http://ecomparatio.net/~khk/venedig/screenshot.png und eine Arbeitsprobe die als Vorarbeit zur Satzklassifikation verstanden werden muß ist ebenfalls online http://ecomparatio.net/~khk/venedig/baselpre.png (login: venedig & pw: famosissima)

[6] Michael Giesecke „Der Buchdruck in der frühen Neuzeit, Eine historische Fallstudie über die Durchsetzung neuer Infromations- und Kommunikationstechnologien "(Suhrkamp, 4. Auflage, Frankfurt am Main, (1991) 2006), S. 672 ff. Was den Aspekt der Abstraktion von der Erfahrung und den Ersatz der Erfahrung im menschlichen Erkenntnisprozess angeht vgl. besonders S. 678/679. Die Diskretisierung der Information scheint in hohem Maße die Grundlage der Erkenntnismethode beeinflusst zu haben. Was geschrieben stand war bereits abstrakt im Vergleich zur Erfahrung, durch die mediale Neuorganisation wurde diese Abstraktion jedoch so bedeutend, dass anstelle der Erfahrung die Anführung des Abstrakten genügte. Medialer Wandel scheint gekennzeichnet durch einen erhöhten Abstraktions- und Ordnungszwang: So manifestiert sich im Handeln des Medienmenschen die Funktion des Mediums ebenfalls. Wo Mannigfaltigkeit/Unendlichkeit der Erfahrung durch das strengere Prinzipielle ersetzt wird, da wird dies dann analog durch ein strenges System nur medial transportfähig. Modell und Leben sind zwei über mehrere Medien hinweg gegensätzliche Positionen.

ferungsproblem der Informatik macht ihre Errungenschaften fraglich, ihre Textproduktion hinderlich. Für die Arbeit an den Methoden, die zu meiner Software führen, lehne ich daher die Form kurzer Veröffentlichungen ab, bzw. halte ich es für meist unnötig diese zu lesen. Die Arbeit an der Theorie geht an die Grenze der Computerisierung zurück und beginnt mit den Veröffentlichungen der mathematischen Logik um 1900. Es kann gezeigt werden, das moderne Logik, die Theorie der Maschinen, die Theorie der Berechenbarkeit, die Theorie der Funktion, sowie die moderne Statistik begrifflich eng mit der Wissenschaft dieser Zeit verbunden ist - dies aber in den meisten Veröffentlichungen der Informatik nicht deutlich wird.

Unabhängig von der Limitierung der Quellen der Theorie, gibt es eine weitere Limitierung die sich für die praktische Softwareentwicklung um einiges erheblicher auswirkt. Bei gegebenen knappen zeitlichen, technischen, sowie personellen Ressourcen wird eine Entwicklungsumgebung benötigt, die eine einfache und schnelle Entwicklung, wie Implementierung der Funktion erlaubt. Aus diesen Gründen wird die Software vollständig in JavaScript geschrieben. Dies verschiebt die Lösung aus dem Bereich des HPC[7], wie üblich für OCR Aufgaben, hin zur client-seitigen Verwendung und Distribution über das Internet. Technisch ist dies durch die jüngeren Entwicklungen der Browser begünstigt. Neben einer weit reichenden Entwicklungsumgebung, sind Nebenläufigkeit und der Zugriff auf die Grafikhardware möglich geworden. Der bisherige Entwicklungsstand meiner Software, lässt erkennen, dass Operationen im Bildpunktraum bis zur Bildgröße von 12 Megapixeln, machbar sind. Die sich anschließende Operation im Objektraum, reduziert den Rechenaufwand nochmals erheblich.

Die für dieses Poster vorbereiteten Methoden umfassen folgendes: (a) Objektrandermittlung auf Basis von Gradientenfeldern im RGBA Farbvektorraum, (b) GPU-Maskierung der Faksimiles zur Ermittelung des abgebildeten Blattrandes auf Basis dynamischer Programmierung, (c) RGBA-Histogramm basierte Methode zur Auswahl von Hintergrundpixeln, (d) RGBA-Grauwertraumwechsel mit dynamischem Schwellwert, (e) Ermittlung der zusammenhängenden Komponenten, (f) horizontale und vertikale Definition von Zusammenhang zwischen den Komponenten und die Ermittlung der Drucklinien auf Basis eines gestuften Mittelwertverfahrens, (g) Drucklinien-Optimierung durch dynamische Programmierung und (h) Boxmodell der Druckbereiche (Layoutklassifikation) durch Drucklinien-gruppierung. Von Verfahren zur frühen Entzerrung der Faksimiles kann abgesehen werden, die Zielfaksimiles sind ausreichend streng strukturiert bzw. wird dies dann Teil der Klassifikation von Formen.

3 Ergebnisse: Die Ergebnisse als Bildbeispiel

Da es nicht versucht werden muss an dieser Stelle, in dieser Kürze, auch nur eine Methode algorithmisch, den Annahmen und den Geltungsbereichen nach aufzuzeigen, sollen die Ergebnisse anhand von bildlichen Representationen angedeutet werden.

[7] High performace computing

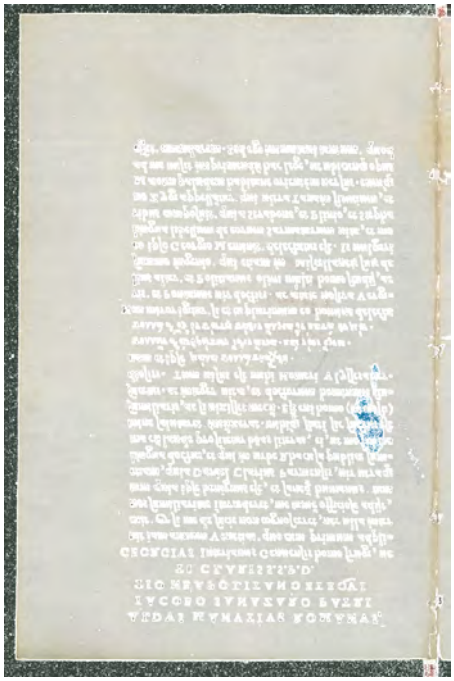


Abbildung 1: Beispielbild zu Gradientenfeldermittelung und Maskierung der Objektkanten, weiße Bereiche entsprechen den Objektgrenzen

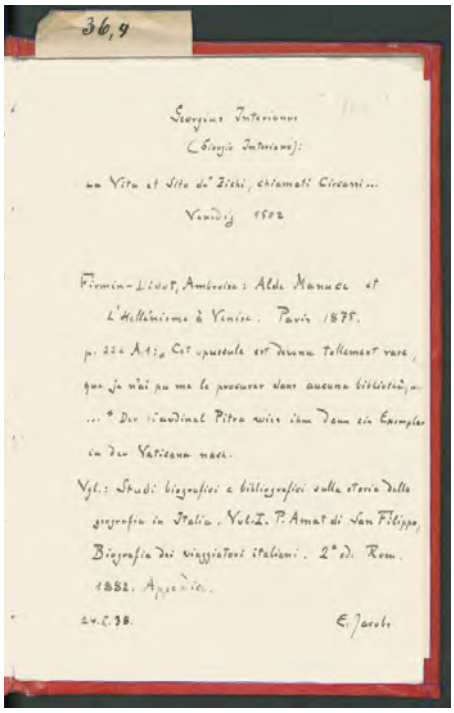


Abbildung 2: Beispielbild zur Blattrandermittelung (b), die blaue Line stellt das Ergebnis des Algorithmus dar

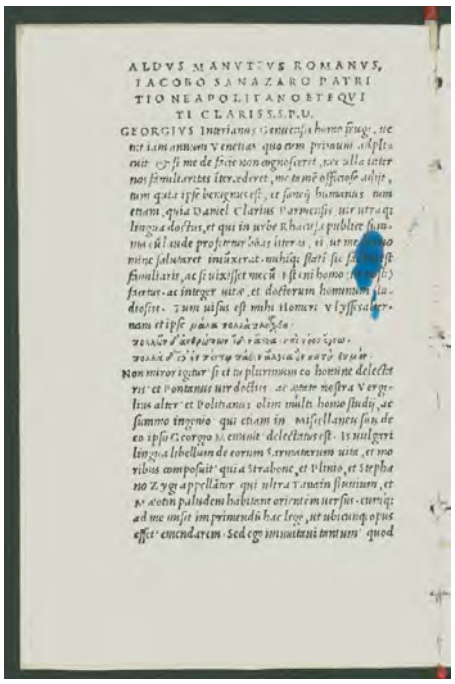


Abbildung 3: Beispielbild zur Auswahl der Papierfarbepixel, die einheitliche graue Farbe, die anstelle der Abbildung des Papiers eingezeichnet wird stellt das ergebnis dar

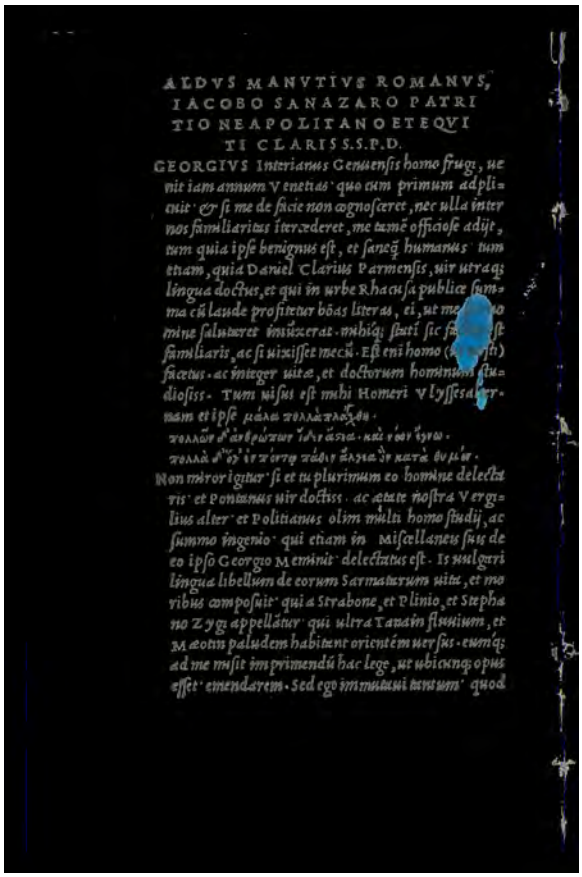


Abbildung 4: Beispielbild der Schwellwertanwendung im Intensitätsraum, die schwarzen Bereiche sind jene, die nach dem Schwellwert als zu hell zu gelten haben

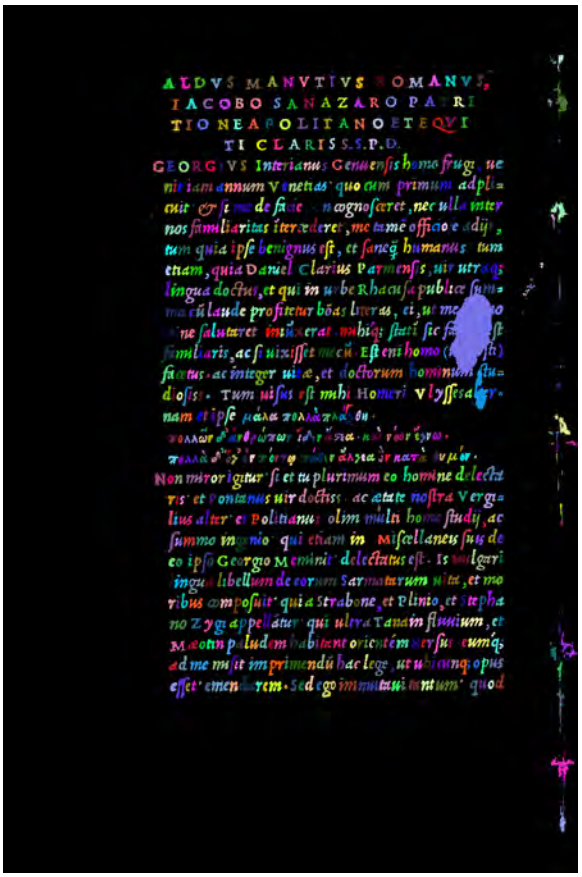


Abbildung 5: Beispielbild der Auswahl der verbunden Komponenten, die einheitlich eingefärbten Pixelregionen sind Komponenten

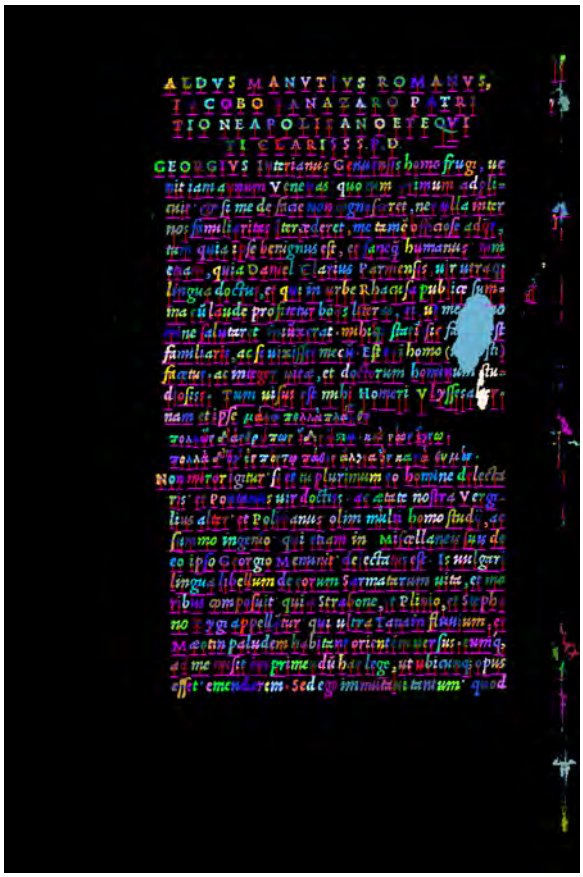


Abbildung 6: Beispielbild Definition von horizontalen und verticalen Zusammenhengen, unter den horizontal zusammenstehenden Komponenten wird die typographische p-Line (Linie die die Unterlänge einschließt) eingezeichnet, hier noch ohen Optimierung und nur auf den vertikalen oder horizonatlen Zusammenhang bezogen

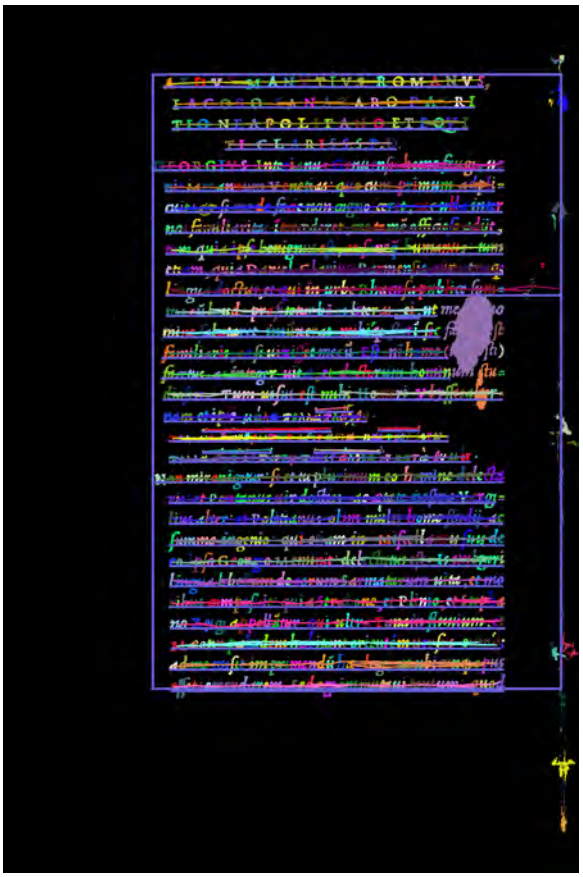


Abbildung 7: Frühes Beispielbild zur Bestimmung der zusammengehörigen Druckzeilen, selbsterklärend

Sguardi sulla tradizione a stampa della fiaba dei fratelli Grimm *Märchen von einem, der auszog, das Fürchten zu lernen* con metodi computazionali

Nicola Nunes, Università di Pisa, nicola.nunes90@gmail.com

1 Introduzione

Questo lavoro propone un'analisi linguistica e testuale della fiaba *Märchen von einem, der auszog, das Fürchten zu lernen* contenuta nella celebre raccolta dei fratelli Grimm *Kinder und Hausmärchen*. La fiaba è stata studiata nelle sue otto edizioni, rielaborate dai fratelli Grimm fra il 1812 e il 1857, anno dell'ultima edizione. L'analisi è stata effettuata grazie a metodi computazionali per l'allineamento e l'analisi linguistica. Successivamente mi sono concentrato su due recenti adattamenti della fiaba in un libro per bambini illustrato e un fumetto, con lo scopo di indagare come viene riadattato il testo per destinatari diversi.

2 La raccolta *Kinder und Hausmärchen*

L'attività dei Grimm tocca molteplici aspetti degli studi linguistici e letterari, fra cui la trascrizione di leggende legate alla Germania, con l'intento primario di recuperare una tradizione orale che stava scomparendo. I due fratelli sfruttarono la stampa per mettere su carta il patrimonio culturale Tedesco e salvarlo dall'oblio.

Jacob e Wilhelm procedono in maniera diversa: il primo ha un approccio scientifico, mentre il secondo si volge verso ambiti fantastici delle fiabe. Queste differenze emergono dalle edizioni, che tra una variante e l'altra risentono di modifiche profonde.

La prima edizione delle fiabe (1812), ha uno scopo puramente filologico, con scarsa attenzione agli aspetti narrativi (Zipes 2014, XX-XXI). Infatti, i Grimm inizialmente trascrivono il racconto orale, fornendoci implicitamente uno spaccato della società di quegli anni. Per i fratelli questa raccolta è lo stimolo per creare un' identità nazionale, poiché i testi sono l'eredità di una cultura orale, e metterli su carta significa arricchire il patrimonio letterario tedesco (Mittner 1978).

Lo spaccato della società fornito dalle nuove edizioni, ci mostra che le fiabe risentono di mutamenti culturali, e della nuova potenza acquisita dal ceto borghese, che diventa primo fruitore della raccolta. Questo porta ad una eliminazione di alcuni contenuti che non si allineavano con la nuova classe sociale. Per quanto riguarda la fiaba, effettivamente risulta da un primo sguardo di confronto fra la versione del 1812 *Gut Kegel und Kartenspiel* e le

versioni successive, una notevole quantità di differenze, non solo ortografiche, ma anche semantiche e tematiche. Nelle prime edizioni della fiaba scelta, ad esempio, il sagrestano viene defenestrato e muore, mentre a partire dalla versione del 1840, si rompe semplicemente una gamba. Partendo dal tema che accomuna i due episodi, cioè il danneggiamento di un personaggio secondario a causa della sua stupidità, esso sfocia in due diversi motivi: danneggiamento fisico da un lato, morte dall'altro. La morte rimane presente fino alla versione del 1840, dopodiché viene sostituita con il danneggiamento fisico. Probabilmente i Grimm hanno modificato il testo riadattandolo sia alle convenzioni culturali e religiose del loro tempo, e sia al suo nuovo intento, cioè quello di manuale educativo (Zipes 2014, XXIX-XXX).

3 Analisi

Per eseguire la comparazione fra le diverse edizioni del testo mi sono avvalso di metodi, risorse e strumenti messi a disposizione dalla corpus linguistics e dalla filologia collaborativa e cooperativa. Ciò mi ha permesso, di facilitare le procedure di collazione delle edizioni, di automatizzare le procedure, ma anche di fornire uno sviluppo all'analisi linguistica che si avvalga delle nuove tecnologie e nuovi criteri di analisi.

Inizialmente mi sono concentrato su sette delle otto versioni, dal 1818 al 1857, i cui testi possono essere ricondotti allo stesso numero di paragrafi, e mostrano contenuti più o meno simili. Ho proceduto alla loro collazione automatica tramite algoritmi di allineamento mutuati dalla bioinformatica per allineare sequenze di DNA (basati sull'algoritmo di Needleman-Wunsch). L'algoritmo permette di individuare differenze fra i testi delle diverse edizioni, allineati parola per parola. Diventa facile quindi comparare cambiamenti dovuti alle riforme ortografiche che la lingua tedesca ha subito, ma anche variazioni tematiche fra un'edizione e l'altra. Certi contenuti, che si sarebbero allontanati dalla morale borghese vengono progressivamente eliminati. Non solo, ma avvicinandoci all'ultima versione, il nucleo narrativo risulta essere più sviluppato, con arricchimenti che provvedono ad una maggiore coerenza testuale.In particolare risulta evidente come fra edizioni temporalmente vicine fra loro le differenze ravvisibali siano di ordine ortografico, lessicale, o di punteggiatura, mentre versioni temporalmente lontane mostrano anche arricchimenti narrativi e vere e proprie variazioni tematiche.

Poichè lo scarto fra la fiaba del 1812 è significativo rispetto alle altre versioni, ho deciso di procedere diversamente nella sua analisi: applicando il modello funzionale di Vladimir Propp fornito in *Morfologia della fiaba*, ho potuto ricondurre le macrostrutture della fiaba ad una serie di sigle. Da ciò risulta che, comparando le due strutture sintetizzate dalle sigle di Propp, la fiaba del 1812 mostra un nucleo centrale interamente contenuto in quella del 1818, con cui l'ho comparata, e la cui unica differenza a livello strutturale è data solo dalla cornice narrativa che troviamo dal 1818 in poi. Questo lascia pensare che la storia del 1812 abbia subito una integrazione con una seconda storia, da cui emerge il tema della paura e la necessità di un finale moralizzante.

Per i due rifacimenti ho proceduto in maniera ancora diversa: Per quanto riguarda il libro per bambini, mi sono concentrato sui criteri di riduzione rispetto alla fiaba originale (si passa da un testo di 3444 tokens ad un testo di 1045 tokens), e ho provveduto ad un'analisi statistica sul lessico della versione precedentemente citata con la versione di riferimento del

1857, così da rilevare automaticamente le parole con frequenze significativamente lontane dai valori attesi, e valutare se nel processo di rielaborazione si sia preferito rimanere vicini all'edizione di riferimento oppure distaccarsi da essa. Per quanto riguarda il fumetto invece ho lavorato sui discorsi diretti presenti in esso in comparazione con i discorsi diretti della versione del 1857. Ho diviso i discorsi diretti in tre tipologie: *T1* per i discorsi attestati anche nella versione originale; *T2* per i discorsi diretti che nella versione del 1857 risultano essere espressi implicitamente o come parte di narrazione; *T3* per i discorsi presenti soltanto nel fumetto. Anche per questa parte ho verificato la pertinenza del nuovo testo rispetto alla versione redatta dai Grimm del 1857.

4 Conclusioni

Per quanto riguarda le differenze fra le sette versioni redatte dai Grimm, mi sono servito di un'applicazione web (http://cophilab.ilc.cnr.it:8080/hypergrimm) che mostra in parallelo coppie di comparazioni in modo che la valutazione delle variazioni venga effettuata contemporaneamente su quattro testi, scelti fra le diverse edizioni. Questo mi permette di farmi un'idea del periodo in cui i fratelli Grimm hanno deciso di modificare fortemente il testo, per riadattarlo alla nuova funzione pedagogica che avrebbe dovuto avere.

Per quanto riguarda il libro da bambini, risulta evidente la vicinanza alla versione del 1857, attraverso l'espediente di glossare quei termini che lessicalmente risulterebbero incomprensibili ad un bambino.

Diversi sono i risultati del fumetto, che pur mantenendo la struttura narrativa della fiaba originale, preferisce sostituire con parole recenti quelle parole ormai cadute in disuso.

L'utilizzo di metodologie e strumenti computazionali per questo lavoro possono sembrare superflui, dato che l'analisi è applicata ad un unico testo, ma ciò è dovuto ai vincoli temporali del mio lavoro, che non si poteva estendere quindi all'intera raccolta *Kinder und Hausmärchen*. Tuttavia questo si configura come un punto di partenza che può essere allargato a tutto il lavoro dei fratelli Grimm: prendendo questa tipologia di analisi come modello, essa potrà essere applicata a tutte le fiabe dei Grimm così da comprendere se i due autori hanno lavorato alla stessa maniera per tutta la raccolta, o se invece la fiaba che ho analizzato risulta essere un'eccezione.

Bibliografia

Boschetti, F. 2008. «Alignment of variant readings for linkage of multiple annotations». In *Proceedings of the ECAL 2007 Electronic Corpora of Ancient Languages, Prague 16–17 November 2007*, a cura di P. Zemánek, 11–24. http://usj.ff.cuni.cz/system/files/Boschetti-Ch-2007.pdf.

Grimm, J., e W. Grimm. 1857. «Märchen von einem der auszog, das Fürchten zu lernen», a cura di D. Buchhandlung, 15–25. Göttingen. https://de.wikisource.org/wiki/M%C3%83%C2%A4hrchen_von_einem,_der_auszog_das_F%C3%83%C2%BCrchten_zu_lernen_(1857).

Jöken, K. 2006. «Grimms Märchen», 99–128. Köln: Ehapa comic collection.

Mittner, L. 1978. *Storia della letteratura Tedesca: Dal pietismo al romanticismo*. 1:923–925. Piccola Biblioteca Einaudi. Torino: Einaudi.

Nicolai, T., e K. Pannen. 2010. *Die Märchenmäuse erzählen Von einem, der auszog, das Fürchten zu lernen*. Göttingen: Lappan.

Nunes, N. 2016. *Sguardi sulla tradizione a stampa della fiaba dei fratelli Grimm* Märchen von einem, der auszog, das Fürchten zu lernen *con metodi computazionali*. Tesi di laurea trienn., Università di Pisa. Pisa.

Propp, V. 1992. *Morfologia della Fiaba: le radici storiche dei racconti di magia*. Roma: Newton Compton editori.

Zipes, J. 2014. *The Original Folk and Fairy Tales of the Brothers Grimm: The Complete First Edition*. Princeton: Princeton University Press.

iAligner: A tool for syntax-based intra-language text alignment

Tariq Yousef, University of Leipzig, tariq.yousef@uni-leipzig.de
Chiara Palladino, University of Bari and Leipzig, chiarapalladino1@gmail.com

1 Introduction

The aim of the poster is to introduce an in-development tool for intra-language and syntax-based text alignment.

Intra-language alignment is the alignment of texts in the same language. The topic was first raised within the Hear Homer project at DEVLAB (Haentjens Dekker et al. 2014). Recently, intra-language alignment methods have been applied in the field of Textual Criticism, with the aim to detect textual variants across various witnesses, in order to support the philological process of collation (Makedon 1998) and the reconstruction of textual transmission (West 1973). Current applications of alignment for semi-automated collation are particularly focused on the detection of multi-variants on "alive" texts, i.e. where the authorial process is documented by multiple manuscript versions. However, ancient texts provide insights into a different situation, where authorial intervention can only be reconstructed from circumstantial evidence: in this case, the relations amongst various witnesses are not always clear, and instances of intertextual relations and reuse are decisive as well.

The tool is available as a web service at the address http://www.i-alignment.com. The actionable Python code is also provided in the GitHub repository (https://github.com/OpenGreekAndLatin/ILA_python).

2 Methodology

The main aim of the tool is to facilitate various degrees of textual comparison: in critical editorial practice, it allows the detection of manuscript variants across several witnesses, including non-literal variants in instances of textual re-use; it also provides comparison across multiple editions.

The alignment is performed through a modified version of the Needleman-Wunsch algorithm (Needleman and Wunsch 1970), also used in Bioinformatics to perform optimal alignment of DNA sequences. The algorithm is optimized by reducing the search space: given two sentences **S1** and **S2** with length **n** and **m** respectively, the algorithm in its basic form compares each word of **S1** with each word of **S2**, producing a search space = **n * m**. As we do not need to compare each word of **S1** with each word of **S2**, our algorithm compares a word **W** in **S1** with words **[W-5, W+5]** in **S2**. Therefore, the search space is reduced from **n * m** to **10 * m** (Fig. 1). Various language-dependent refinement options are additionally

chosen by the user: diacritics and punctuation can be detected as single tokens or ignored, and Levensthein distance metric can be applied to adjust the tolerance threshold, in order to restrict or amplify the tolerance in the detection of variants.

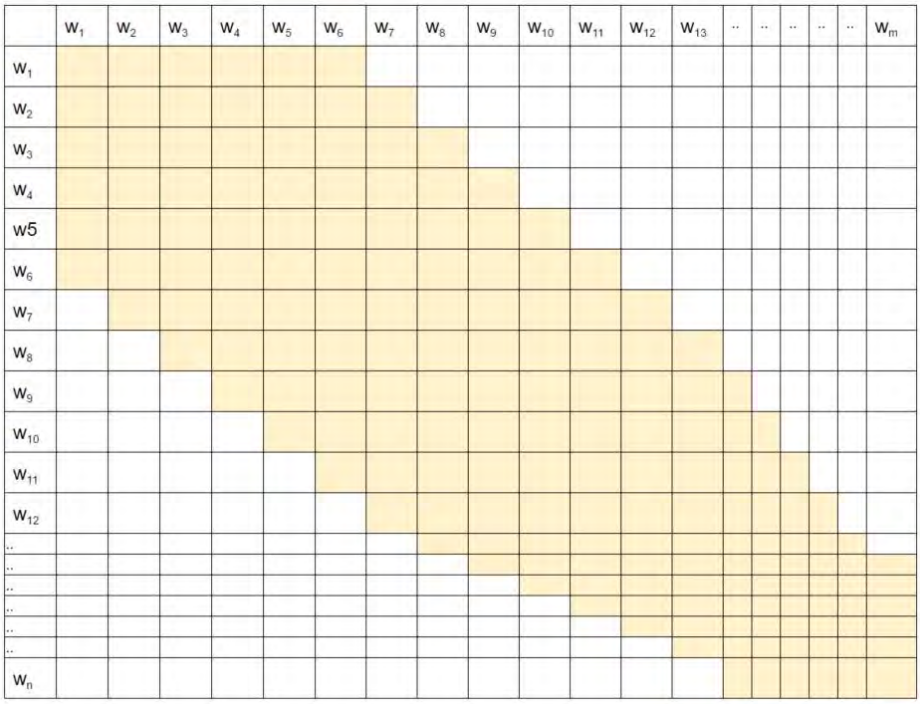


Figure 1: Reduction of the search space in the optimized Needleman-Wunsch algorithm. The white cells represent the search space of the algorithm in its normal form, the yellow cells the optimized search space.

3 Workflow

The alignment can be performed either by uploading the text in tabular CSV format, or by pasting chosen junks in plain format. Further development will also allow direct upload of texts in XML and JSON format.

The uploaded file is parsed into a list of parallel sentences, which are then passed to tokenization and converted to a vector of single tokens: we used a simple tokenizer, which takes white spaces and punctuation marks as delimiters to split the sentence in a vector of single words. Once the parallel sentences are tokenized, they are processed by the alignment algorithm and further elaborated by the user according to the various refinement options. At present, the texts are supposed to be initially structured at paragraph or sentence level. Future implementations will also allow a preprocessing stage for initially not aligned texts, by means of algorithms on length-based methods (Thompson 1994).

Current results

The several refinement criteria are provided to the user in order to allow a specific performance of the algorithm, according to the purpose: for example, for the detection of manuscript variants, we did not apply any criteria of tolerance, in order to be able to individuate every minute difference between various exemplars: this approach was experimented on three manuscript witnesses of Plato's *Crito* [Image 2].

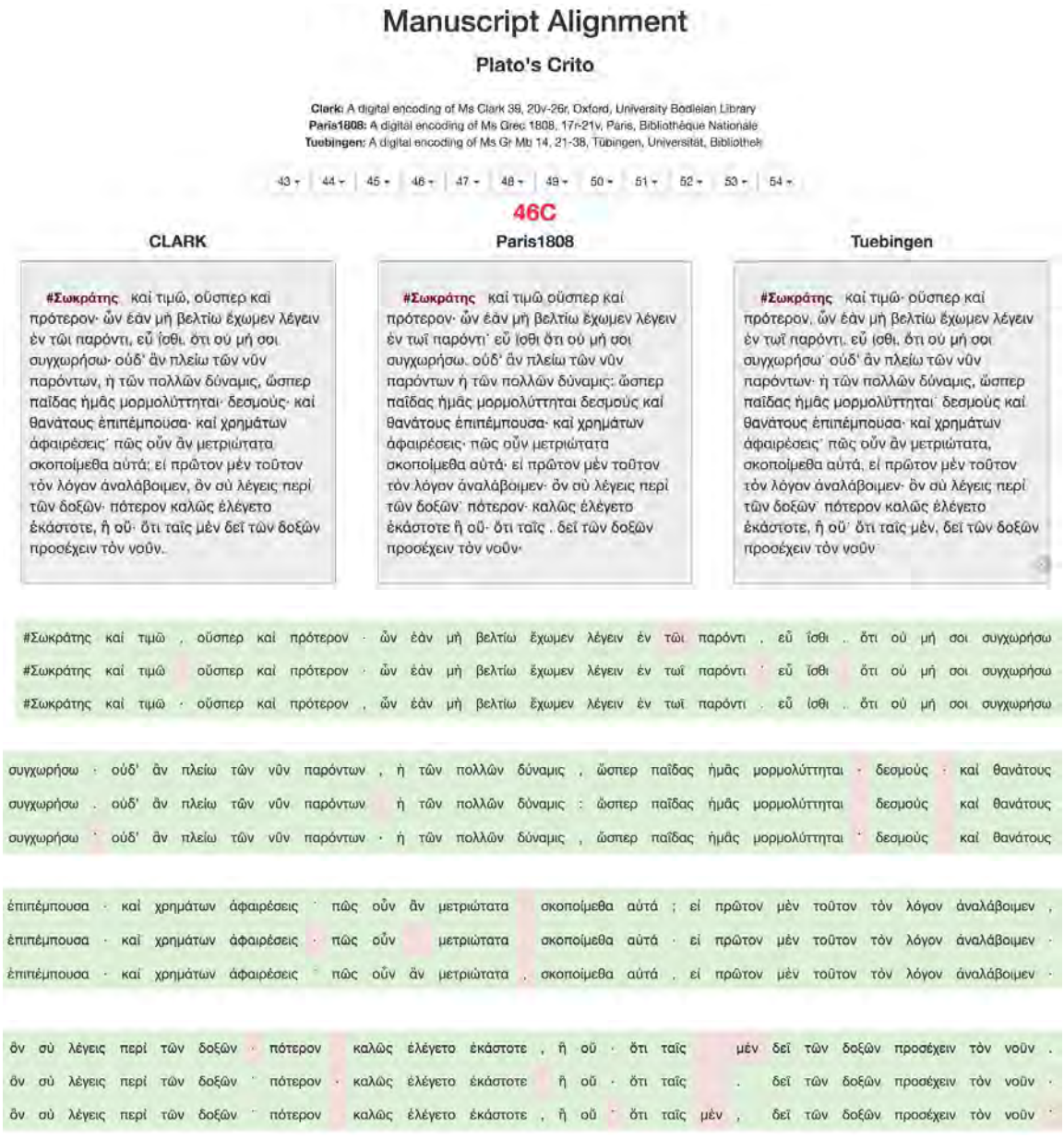


Image 2: alignment of manuscripts of Plato's *Crito*.

Figure 2: Alignment of manuscripts of Plato's *Crito*.

On the other hand, user refinement criteria proved to be essential for more complex instances. We applied the workflow to the in-progress born-digital critical edition of Agathemerus' *Sketch of Geography*, a Greek geographical work whose transmission offers a good

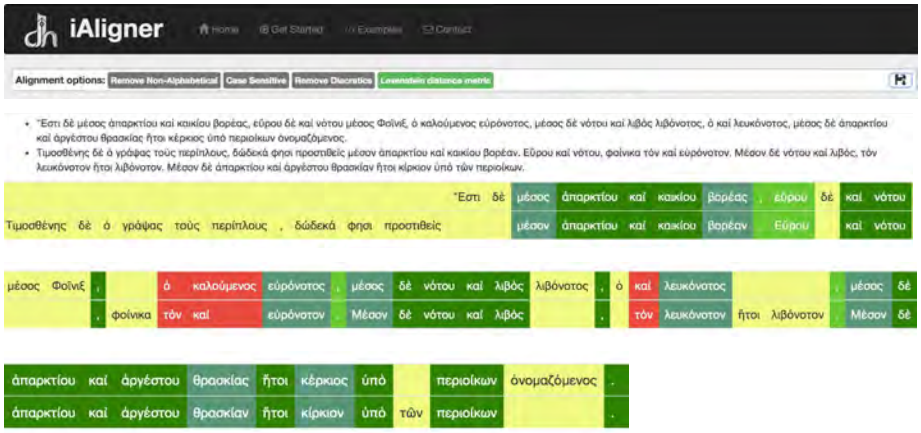


Image 3: An extract from aligned Excerpt a.

Figure 4: An extract from aligned Excerpt a.

Future work

Current results encourage both applications on scholarly editorial practice and on larger efforts for the detection of a high amount of variants. A further stage is going to establish a workflow for automatic alignment of OCR outputs for postcorrection. Current limits in the alignment output, due to the use of a syntax-based algorithm, will also be addressed

5 Future work

Current results encourage both applications on scholarly editorial practice and on larger efforts for the detection of a high amount of variants. A further stage is going to establish a workflow for automatic alignment of OCR outputs for postcorrection. Current limits in the alignment output, due to the use of a syntax-based algorithm, will also be addressed combining syntactic and semantic matching by means of lexical information (e.g. lemma, synonyms and PoS tagging).

Various options for the graphic display of the alignment are currently evaluated, including visualization of the syntax with different coloring options according to the user's refinement criteria, highlight of individual variants by means of graphs, and adaptations of customisable existing graphs for comparison based on one reference text (Jänicke 2014).

References

Brown, P. F. 1991. «Aligning sentences in parallel corpora». In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, 169–176. Berkeley.

Haentjens Dekker, R., et al. 2014. «Computer-supported collation of modern manuscripts: CollateX and the Beckett Digital Manuscript Project». *Digital Scholarship in the Humanities*. doi:10.1093/llc/fqu007.

Jänicke, S. 2014. «Improving the Layout for Text Variant Graphs». In *VisLR workshop at LREC Conference*. Reykjavík.

Levenshtein, V. I. 1966. «Binary codes capable of correcting deletions, insertions, and reversals». *Soviet Physics Doklady* 10:707–710.

Makedon, F. 1998. «HEAR HOMER: A multimedia-data access remote prototype for ancient texts». In *Proceedings of ED-MEDIA'98*. Freiburg.

Needleman, S. B., and C. D. Wunsch. 1970. «A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins». *Journal of Molecular Biology* 48:443–453.

Thompson, J. 1994. «CLUSTAL W: Improving the Sensitivity of Progressive Multiple Sequence Alignment through Sequence Weighting, Position-Specific Gap Penalties and Weight Matrix Choice». *Nucleic Acids Research* 22:4673–4680.

West, M. 1973. *Textual criticism and editorial technique*. Stuttgart: Teubner.

Linked Books: a bibliometric history of the history of Venice

Matteo Romanello, École Polytechnique Fédérale de Lausanne (EPFL) – Digital Humanities Laboratory (DHL)
Giovanni Colavizza, EPFL-DHL
Martina Babetto, EPFL-DHL
Silvia Ferronato, EPFL-DHL
Frédéric Kaplan, EPFL-DHL

1 Introduction

The use of data to analyze and evaluate research has a long tradition (Smith 2012), which has been greatly impacted by citation indexing services since the publication of the Science Citation Index by Eugene Garfield in 1964. Among the first to recognize the importance of this newly available tool in order to analyze bibliometric processes was a historian, Derek de Solla Price, who introduced the idea of the cumulative advantage in citation networks (Price 1976). Despite its drawbacks, such as the indiscriminate use of indicators of performance, the availability of citation data has provided both an increased understanding of science as a communication process, and an invaluable wealth of information retrieval means to the modern researcher. Nothing comparable yet exists for the Arts and Humanities (Sula and Miller 2014).

The Linked Books project is a joint effort of the Swiss Federal Institute of Technology in Lausanne and a growing consortium of partner libraries[1]. Our goal is to digitize the available scholarly literature on the history of Venice, index and analyze the citations it contains and make this data available through a search engine. This will enable scholars of Venice to navigate citation data within the catalog (references to secondary sources) and between the catalog and the information system of the State Archive of Venice (references to primary sources). Our project is articulated in a series of steps, from the selection of the literature to the implementation of the search engine, which entail specific challenges. This contribution briefly discusses the structure of the project in terms of its pipeline, and further expands on the development of the first version of an environment to navigate and annotate our digitized sources: the Linked Books Catalog.

[1] In alphabetical order: Archivio di Stato di Venezia, Biblioteca Marciana (Venice), Fondazione BEIC (Milan), ICCU (Rome), Istituto Veneto di Scienze, Lettere ed Arti (Venice), Sistema Bibliotecario di Ateneo e Biblioteca di Area Umanistica Ca' Foscari (Venice).

2 Structure of the project

The pipeline of the Linked Books project is articulated as a possibly never-ending circuit, as potentially neverending is the production of new scholarship on Venice (Fig. 1). We consider the library catalog both our starting point and our end point, in the sense that the citation data we extract can be of two typologies:

1. **Primary sources**: references to documentary evidence, often from the State Archive of Venice.
2. **Secondary sources**: references to other published scholarly literature.

The nature of citation data is relational: a citation is a link between resources. We start from the catalog, in the sense that we select our literature from it, and we end in the catalog meaning that we ultimately enrich it with relations among catalog resources and with archival resources. The cycle is also self-incremental, meaning that a snowball effect will eventually take place, by which we will be able to select further literature to digitize and process from the previously extracted citation data.

The selection of the first seed of literature was done using all available means: library catalog (especially by subject and consultation shelves), domain bibliographies, expert advice. The result was a set of circa 2000 books and 4 journals, for another approximately 1000 issues. A common approach to citation extraction is supervised classification, meaning a subset of the data are to be annotated by hand by experts, and will be used in order to train a set of parsers and recognize: 1. footnotes, 2. references, 3. components of every reference (e.g. author, title, etc.). A cycle of annotation, parsing and extraction goes on until results of sufficient quality are reached.

The last step is the look-up: a matching system which is able to couple extracted references with the resources they refer to in the catalog or in the information system of the archive. The result of this process are links among resources.

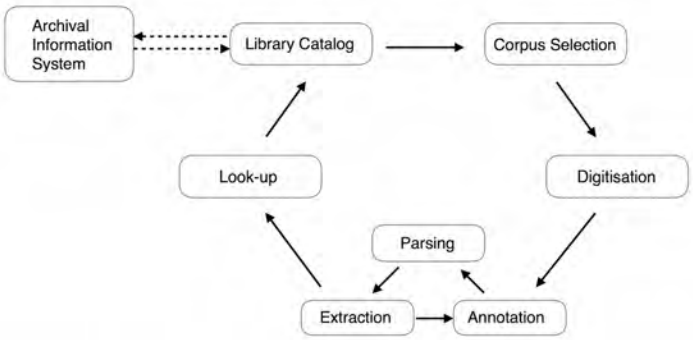


Figure 1: Linked Books pipeline.

3 Linked Books Catalog

The main role of the Linked Books Catalog is to allow for the rapid annotation, enrichment and verification of all the descriptive and structural metadata, as well as the contents of the

literature. It is positioned between digitization and annotation in the pipeline. Since we deal with a considerable amount of journals, a finer grained enrichment of catalog metadata is needed.

The current version of the Catalog supports the following tasks:

1. The correction of digitization metadata (Fig. 2).
2. The reconciliation of the numbering of the digitized images with the pagination of the printed issue.
3. The identification of the pages that contain the Table of Contents (ToC) of a journal issue.
4. The semi-automatic reconstruction of the contents of an issue: the results of parsing the ToC are corrected manually so as to obtain a precise list of articles contained in each issue (Fig. 3). This allows us to exclude from the processing steps those sections that are less relevant and would just introduce/cause unnecessary noise (e.g. book reviews, obituaries, etc.).
5. The annotation of footnotes (Fig. 4).

We are currently exploring the possibility of deploying a version of the Catalog that can be used in situ by users of the libraries participating to the project.

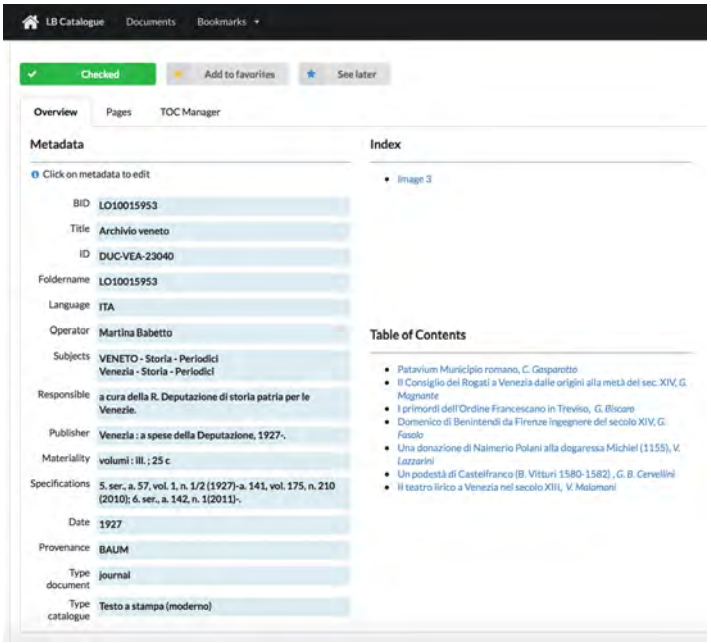


Figure 2: One issue (n. 1, 1927) of Archivio Veneto.

In this paper we have described the overall structure and some preliminary results of Linked Books, a project that aims at digitizing, analyzing and making searchable the available scholarly literature on the history of Venice. The main focus of the project is on the extraction of references – both to primary sources and to secondary literature – as a way of mapping and making more effective searchable the available literature in this domain.

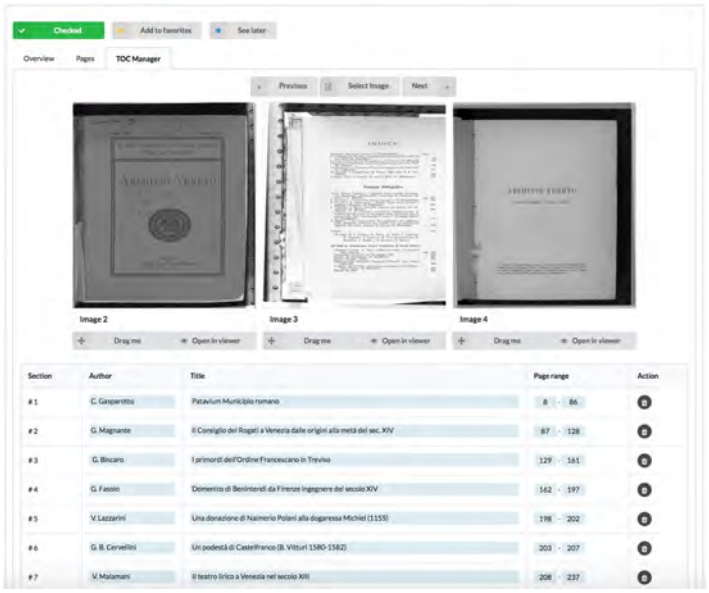


Figure 3: Structure of the issue.

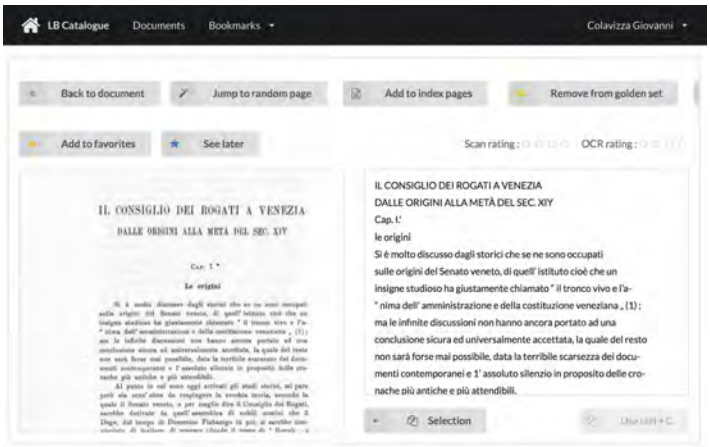


Figure 4: Check of contents and annotation of footnotes.

References

Price, D. d. S. 1976. «A general theory of bibliometric and other cumulative advantage processes». *Journal of the American Society for Information Science* 27 (5–6): 292–306. http://garfield.library.upenn.edu/price/pricetheory1976.pdf.

Smith, D. R. 2012. «Impact factors, scientometrics and the history of citation-based research». *Scientometrics* 92:419–427.

Sula, C. A., and M. Miller. 2014. «Citations, contexts, and humanistic discourse: toward automatic extraction and classification». *Literary and Linguistic Computing* 29 (3): 452–464.

Strumenti e Architetture
Tools and Architectures
Talks

Vantaggi dell'Astrazione attraverso l'Approccio Orientato agli Oggetti per il Digital Scholarly Editing

Angelo M. Del Grosso, CNR-ILC, angelo.delgrosso@ilc.cnr.it
Federico Boschetti, CNR-ILC, federico.boschetti@ilc.cnr.it
Emiliano Giovannetti, CNR-ILC, emiliano.giovannetti@ilc.cnr.it
Simone Marchi, CNR-ILC, simone.marchi@ilc.cnr.it

1 Introduzione

La comunità delle Digital Humanities (DH) sta diventando sempre più inclusiva, non solo nei confronti di gruppi di ricerca in ambito computazionale, ma anche nei confronti delle comunità che praticano le discipline umanistiche con metodi non digitali, con le quali nel passato c'era un difetto di comunicazione. Grazie a questo dialogo ritrovato, è necessario che l'umanista digitale accolga le esigenze di questa allargata multidisciplinarietà. Le DH sono chiamate quindi ad essere sempre meno autoreferenziali e sempre più attente a definire metodi che producano risultati rilevanti per le discipline tradizionali. La collaborazione fra umanisti e informatici, spesso con l'intermediazione degli umanisti digitali, è ormai un fatto assodato e porta a progetti collaborativi che promuovono l'interoperabilità delle risorse ma non garantiscono generalmente la riusabilità delle componenti software e la personalizzazione delle implementazioni.

Infatti, gli strumenti realizzati all'interno di progetti collaborativi difficilmente riescono a recuperare moduli sviluppati in altri progetti complementari, limitando sensibilmente la possibilità di cooperare senza pesanti adattamenti e ristrutturazioni del software. Per affrontare questo stato di cose, è necessario dunque lo sviluppo di modelli condivisi, astratti e formali finalizzati alla costruzione di strumenti flessibili, estendibili e riusabili rivolti allo studio filologico del testo (fra gli altri cfr. Boschetti e Del Grosso 2015). Proprio in questa direzione stanno andando le grandi iniziative infrastrutturali quali DARIAH[1], DiXIT[2], CLARIN[3], TAPOR[4], DiRT[5], impegnate inizialmente nella catalogazione delle risorse, degli strumenti e alla descrizione dei desiderata (requisiti utente). E' quindi opportuno procedere verso la definizione di un processo community-driven volto alla formalizzazione di tipi di dato astratti (Gabbrielli e Martini 2010; in inglese: *Abstract Data Type*, ADT) in grado di standardizzare i requisiti raccolti.

[1] http://dariah.eu
[2] http://dixit.uni-koeln.de
[3] https://www.clarin.eu
[4] http://www.tapor.ca
[5] http://dirtdirectory.org

2 Contesto

La ricerca di modelli generici per lo studio del testo digitale ha prodotto negli ultimi anni molte riflessioni e dibattiti (si vedano per esempio Robinson 2013b; Sahle 2013; Vanhoutte 2010; Shillingsburg 2006b; Thaller 2006). Per esempio, già nel lavoro "Humanities Computing" McCarty 2005, si evidenziava la necessità di definire sistematicamente, attraverso un processo rigoroso, modelli astratti che fossero condivisi dalla comunità.

A distanza di oltre 10 anni, la discussione sulla necessità di realizzare modelli concettuali per la definizione delle entità di dominio resta ancora una questione aperta (cfr. Pierazzo 2015b, in particolare il capitolo 2 - Modelling Digital Texts; cfr. Schreibman, Siemens e Unsworth 2016) Del resto, l'AIUCD ha tra i suoi obiettivi principali la condivisione delle best practices maturate nelle diverse discipline che hanno come centro di interesse o come interesse secondario il trattamento del testo (Agosti e Tomasi 2014; Tomasi 2015).

3 Approccio metodologico

Le esperienze maturate in seno a progetti nei quali siamo stati coinvolti negli ultimi anni hanno messo in evidenza che astrazioni in grado di rappresentare risorse di natura diversa, come ad esempio, nel caso dell'allineamento, singole parole, parole e immagini, unità di testo a diverse granularità, consentono di sviluppare sistemi più flessibili e riusabili. Tra i nostri progetti citiamo: l'ERC "Greek into Arabic[6]" (testi greci e arabi in parallelo); il PRIN "F. de Saussure[7]" (testo e immagini affiancate); il progetto di Traduzione del Talmud Babilonese[8] (segmenti di testo in lingua originale con la relativa traduzione italiana); il progetto "Clavius on the Web[9]" (testo e immagini allineati); il CoPhiProofreader[10] (allineamento dei risultati del processo di OCR); il software Euporia[11] (testi con traduzione a fronte da annotare).

L'approccio qui proposto prevede l'adozione di un processo rigoroso e orientato agli oggetti (in inglese: *Object Oriented Analysis and Design & Object Oriented Programming*, OOAD&OOP) mutuato da pratiche di ingegneria del software ormai consolidate (Dathan e Ramnath 2015). Questo prevede, inoltre, il coinvolgimento della comunità interessata allo studio filologico del testo digitale fin dalle prime fasi della sua attuazione. Ci proponiamo, infatti, di seguire un processo community-driven (Vernon 2013) e user-centered (Gibbs e Owens 2012) che tenga in particolare considerazione le esigenze degli utenti finali.

Obiettivo della metodologia è quello di definire tipi di dato astratti in grado di cogliere le caratteristiche essenziali del dominio di interesse. L'approccio si articola nei seguenti passi:

1. individuazione dei requisiti utente tramite la definizione di *user stories* (Cohn 2004) (ad esempio: "*come* editore critico, *voglio* ricercare tutte le varianti di una lezione di un testo *al fine di* confrontare i vari testimoni");

[6] http://www.greekintoarabic.eu, https://github.com/literarycomputinglab/G2A_Wapp
[7] http://webilc.ilc.cnr.it/viewpage.php/sez=ricerca/id=917/vers=ita
[8] https://www.talmud.it
[9] http://claviusontheweb.it
[10] https://github.com/CoPhi/cophiproofreader
[11] https://github.com/CoPhi/EUporiaJsf

2. identificazione delle entità di dominio (Del Grosso et al. 2016) (ad esempio: l'entità *Source* denota il concetto di fonte primaria, mentre l'entità *Content* denota il contenuto informativo della fonte stessa);

3. definizione degli ADT (Boschetti et al. 2014) attraverso l'identificazione a) dei modelli astratti (ad esempio *Document* per la rappresentazione digitale di un documento) e b) delle relative operazioni (ad esempio aggiunta di una annotazione linguistica ad una porzione di un documento).

Una volta definiti gli ADT sarà possibile realizzare, a partire da essi, dei componenti software, da intendersi come unità uniformi di servizi (per esempio analisi linguistica, annotazione del testo, ecc.) invocabili tramite interfacce di programmazione (API). La realizzazione di interfacce grafiche (GUI) completerà il processo di sviluppo.

Come si è detto all'inizio di questa sezione, ci siamo occupati di vari tipi di allineamento (testo a diversi livelli di granularità, testo e immagine, ecc.) e per questo abbiamo cercato di generalizzare l'approccio proponendo un tipo di dato astratto schematizzato in Tab. 1.

Grazie a questa formalizzazione siamo in grado di standardizzare l'uso del servizio e garantire l'allineamento di liste di oggetti di tipo diverso. La flessibilità di tale soluzione risiede nella personalizzazione della strategia di allineamento. Ad esempio se si allineano stringhe la strategia farà uso di meccanismi basati sull'*edit distance*, mentre se si allineano oggetti più complessi la strategia valuta la distanza nello spazio vettoriale delle sue *features*.

A questo proposito stiamo lavorando anche alla definizione di modelli per la rappresentazione del testo e delle relative risorse contestuali (lessici, terminologie e ontologie) e alla realizzazione di componenti software per il loro trattamento. Nel caso specifico del problema dell'allineamento, potremo, infatti, avvalerci di nuove *features* di natura lessico-semantica. Sarà possibile valutare, ad esempio, l'allineamento tra due parole sulla base di una similarità semantica rappresentata attraverso una relazione di sinonimia specificata tra i rispettivi sensi lessicali (nel nostro caso codificati attraverso il modello *lemon*[12]).

4 Conclusioni

Troppo spesso gli strumenti sviluppati nell'ambito delle DH si limitano a risolvere problemi specifici (tipicamente attraverso interfacce grafiche) senza particolare attenzione alla riusabilità dei modelli e dei moduli software; una possibile spiegazione si ritrova nella effettiva carenza di astrazioni delle entità di dominio e delle librerie software che possano manipolarle.

Come descritto in questo contributo, si intende proporre alla comunità delle DH un approccio ingegneristico, orientato agli oggetti, che possa contribuire a uno sviluppo più sistematico e condiviso di entità software da trattare come vere e proprie risorse (al pari di quelle testuali / documentali) da arricchire, estendere e condividere.

Un lavoro di questo tipo, per definizione, richiede il coinvolgimento dell'intera comunità, alla quale si chiede di specificare, in modo chiaro e puntuale, i requisiti d'uso fondamentali e condivisi dai quali partire per realizzare componenti software che siano utilizzabili (previ adattamenti minimi) da un numero maggiore possibile di utenti.

[12] http://lemon-model.net

public class **AlignmentTable**		
	AlignementTable()	Costruttore
void	**addList**(*List*<S> list)	Aggiunta di una lista di elementi da allineare con gli elementi di altre liste
void	**alignAll**(*Strategy*<S,T> strategy)	Processo di allineamento con strategia personalizzata
Record<T>	**getRecord**(int idx)	Ritorna l'allineamento alla posizione specificata da idx
Element<T>	**getElement**(int recordIdx, int idx)	Ritorna l'elemento della tabella specificato dalla posizione *idx* del Record specificato da *recordIdx*
List<*Element*<T>>	**getColumn**(int idx)	Ritorna la lista allineata alla colonna specificata da idx
boolean	**isAligned**()	Verifica se le liste sono state allineate
...	...	Iteratori...

Tabella 1: AlignmentTable Abstract Data Type. L'ADT specifica le operazioni (il comportamento) che possono essere effettuate sugli oggetti che lo istanziano. La rappresentazione dello stato è nascosto all'utente dell'ADT.

Tabella 1. *AlignmentTable Abstract Data Type*
L'ADT specifica le operazioni (il comportamento) che possono essere effettuate sugli oggetti che lo istanziano. La rappresentazione dello stato è nascosto all'utente dell'ADT.

Conclusioni

Troppo spesso gli strumenti sviluppati nell'ambito delle DH si limitano a risolvere problemi specifici (tipicamente attraverso interfacce grafiche) senza particolare attenzione alla riusabilità dei modelli e dei moduli software; una possibile spiegazione si ritrova nella effettiva carenza di astrazioni delle entità di dominio e delle librerie software che possano manipolarle.

Come descritto in questo contributo, si intende proporre alla comunità delle DH un approccio ingegneristico, orientato agli oggetti, che possa contribuire a uno sviluppo più sistematico e condiviso di entità software da trattare come vere e proprie risorse (al pari di quelle testuali/documentali) da arricchire, estendere e condividere.

Un lavoro di questo tipo, per definizione, richiede il coinvolgimento dell'intera comunità, alla quale si chiede di specificare, in modo chiaro e puntuale, i requisiti d'uso fondamentali e condivisi dai quali partire per realizzare componenti software che siano utilizzabili (previ adattamenti minimi) da un numero maggiore possibile di utenti.

Il convegno dell'AIUCD, coerentemente alla vocazione fortemente inclusiva e collaborativa esplicitata nella *call for papers*, potrà fornirci l'opportunità di lanciare pubblicamente questa iniziativa.

Operativamente, si intende proporre la costituzione di uno Special Interest Group aperto a tutti coloro che siano disposti a confrontarsi nella raccolta dei requisiti degli utenti, delle entità di dominio, degli ADT e delle API.

La metodologia di lavoro prevede l'utilizzo di un processo iterativo (es. Agile use case driven e Domain-driven design) nel quale i requisiti potranno evolvere in accordo ai contributi provenienti dalla comunità fino a stabilizzarsi in una forma che sia la più largamente condivisa: sarà possibile far fronte alla loro natura intrinsecamente instabile proprio attraverso l'adozione del paradigma di analisi, progettazione e sviluppo orientato agli oggetti.

Questo processo si potrà concretizzare, ad esempio, attraverso la costituzione di un portale pubblico comprensivo di forum, wiki e chat, che fungerà da hub per la raccolta dei contributi provenienti dai gruppi di lavoro impegnati nelle DH interessati a partecipare a questa iniziativa.

Il risultato atteso sarà la pubblicazione di documenti di specifica i quali guideranno lo sviluppo di strumenti software flessibili ed estendibili, in grado di soddisfare i bisogni propri della nostra comunità.

Il convegno dell'AIUCD, coerentemente alla vocazione fortemente inclusiva e collaborativa esplicitata nella *call for papers*, potrà fornirci l'opportunità di lanciare pubblicamente questa iniziativa.

Bibliografia

Agosti, M., e F. Tomasi, cur. 2014. *Collaborative Research Practices and Shared Infrastructures for Humanities Computing*. Proceedings of revised papers of the Annual Conference of the Associazione per l'Informatica Umanistica e la Cultura Digitale (AIUCD). 11-12 December 2013 Padova: Cooperativa Libraria Editrice Università Di Padova (CLEUP). ISBN: 9788867872602.

Boschetti, F., e A. M. Del Grosso. 2015. «TeiCoPhiLib: A Library of Components for the Domain of Collaborative Philology». A cura di A. Ciula e F. Ciotti. [on line journal], *Journal of the Text Encoding Initiative*, n. 8 (). ISSN: 2162-5603. doi:10.4000/jtei.1285. http://jtei.revues.org/1285.

Boschetti, F., et al. 2014. «A top-down approach to the design of components for the philological domain». In *Book of abstract of Digital Humanities Conference*, a cura di M. Terras, 109–111. Lousanne. http://dharchive.org/paper/DH2014/Paper-673.xml.

Cohn, M. 2004. *User Stories Applied: For Agile Software Development*. Agile Software Development. Crawfordsville, Indiana: Addison-Wesley Professional. ISBN: 9780321205681.

Dathan, B., e S. Ramnath. 2015. *Object-Oriented Analysis, Design and Implementation*. Second. Undergraduate Topics in Computer Science, 1863-7310. 10.1007/978-3-319-24280-4. Springer International Publishing. ISBN: 9783319242804.

Del Grosso, A. M., et al. 2016. «Defining the Core Entities of an Environment for Textual Processing in Literary Computing». In *Digital Humanities 2016: Conference Abstracts*, a cura di M. Eder e J. Rybicki, 771–775. Kraków: Jagiellonian University & Pedagogical University. ISBN: 9788394276034. http://dh2016.adho.org/abstracts/425.

Gabbrielli, M., e S. Martini. 2010. *Programming Languages: Principles and Paradigms*. First. Undergraduate Topics in Computer Science, 1863-7310. DOI: 10.1007/978-1-84882-914-5. Springer London. ISBN: 9781848829138.

Gibbs, F., e T. Owens. 2012. «Building Better Digital Humanities Tools: Toward broader audiences and user-centered designs». A cura di M. Carassai e E. Takehana. *Digital Humanities Quarterly* 6 (2). ISSN: 1938-4122. http://www.digitalhumanities.org/dhq/vol/6/2/000136/000136.html.

McCarty, W. 2005. *Humanities Computing*. London: Palgrave Macmillan. ISBN: 9781403935045.

Pierazzo, E. 2015b. *Digital Scholarly Editing : Theories, Models and Methods*. Digital Research in the Arts and Humanities. Farnham Surrey: Ashgate. ISBN: 9781472412119.

Robinson, P. 2013b. «Towards a Theory of Digital Editions». A cura di W. Van Mierlo e A. Fachard. ISBN13: 9789042036321, *Variants: The Journal of the European Society for Textual Scholarship*, Variants, n. 10:105–131. ISSN: 1573-3084.

Sahle, P. 2013. *Digitale Editionsformen: Textbegriffe und Recodierung (Teil 3). Zum Umgang mit der Überlieferung unter den Bedingungen des Medienwandels*. Vol. 9. Schriften des Instituts für Dokumentologie und Editorik. Books on Demand. ISBN: 9783848239665.

Schreibman, S., R. Siemens e J. Unsworth, cur. 2016. *A new companion to digital humanities*. Vol. 93. Blackwell companions to literature and culture. Chichester, West Sussex: John Wiley & Sons. ISBN: 9781118680599.

Shillingsburg, P. L. 2006b. *From Gutenberg to Google: Electronic Representations of Literary Texts*. New York, USA: Cambridge University Press. ISBN: 9780521683470.

Thaller, M. 2006. «Waiting for the Next Wave: Humanities Computing in 2006». In *Literatures, Languages and Cultural Heritage in a digital world*. King's College London. http://legacy.cch.kcl.ac.uk/clip2006/content/abstracts/paper04.html.

Tomasi, F., cur. 2015. *Humanities and their Methods in the Digital Ecosystem*. Proceedings of revised papers of the Annual Conference of the Associazione per l'Informatica Umanistica e la Cultura Digitale (AIUCD). 18-19 Settembre 2014, Bologna: Association for Computing Machinery (ACM), New York, NY, USA. ISBN: 9781450332958. http://dl.acm.org/citation.cfm?id=2802612.

Vanhoutte, E. 2010. «Defining Electronic Editions: A Historical and Functional Perspective». In *Text and Genre in Reconstruction: Effects of Digitalization on Ideas, Behaviours, Products and Institutions*, a cura di W. McCarty, 119–144. Digital Humanities. Open Book Publishers. ISBN: 9781906924263.

Vernon, V. 2013. *Implementing domain-driven design*. First. Westford, Massachusetts (US): Addison-Wesley Professional. ISBN: 9780321834577.

Un modello ed un linguaggio ad oggetti per documenti testuali

Renzo Orsini, Università Ca' Foscari Venezia, orsini@unive.it

1 Abstract

Viene presentata un'alternativa all'uso dei marcatori per la rappresentazione di testi basata su un modello ad oggetti, simile a quelli usata nella moderna ingegneria del software. L'obiettivo è la rappresentazione di testi con molteplici gerarchie e un numero potenzialmente illimitato di livelli di annotazione, insieme a un linguaggio ad alto livello per ricerche e manipolazioni complesse di testi.

2 Introduzione

2.1 Motivazioni

L'affermarsi di XML negli anni recenti ha portato diversi benefici nell'area delle discipline umanistiche, fra cui: a) la possibilità di codificare testi interoperabili attraverso un "minimo comun denominatore", costituito da un'alfabeto standard, Unicode, e da una struttura del documento che può essere sia validata, che interpretata da programmi generici, come editor, visualizzatori, sistemi di interrogazione; b) lo sviluppo di linee guida per la codifica di ampie classi di documenti, come quelle proposte dal comitato TEI (originariamente nate per SGML, ma successivamente adattate a XML), che offrono un modello e un vocabolario di riferimento di "marcatori" XML.

D'altra parte è noto da tempo che XML soffre di significative limitazioni, sia riguardo alla rappresentazione dei testi, che al loro utilizzo.

Per la rappresentazione dei testi, XML prevede direttamente la possibilità di strutturare un documento attraverso un'unica composizione gerarchica. Nel caso molto frequente di gerarchie multiple, invece, come quando si vuole rappresentare la struttura fisica di un documento insieme a quella metrica e linguistica, è necessario ricorrere ad artifici, quali l'uso delle codifiche multiple degli stessi dati (*stand-off markup*), marcatori di elementi vuoti che delimitano i confini di strutture non annidate (*milestones*), divisione del testo in segmenti da usare per ricostruzione di elementi virtuali, ecc. (si veda ad esempio DeRose 2004). Questa situazione è aggravata quando si ha la necessità di associare al testo del documento più serie di annotazioni di vario tipo, ad esempio filologiche, grammaticali, semantiche, o generici commenti. In questo caso la complessità del documento aumenta in maniera tale da renderlo praticamente inutilizzabile attraverso l'*editing* manuale, vanificando di fatto il principale beneficio della marcatura. In termini informatici, possiamo dire che XML non

è "scalabile", cioè cessa di essere utilizzabile in maniera pratica al superamento di una certa soglia di complessità.

Naturalmente la complessità della struttura si riflette negativamente nell'uso di questo tipo di documenti, sia per semplici operazioni di ricerca che tengano conto delle diverse gerarchie, sia per operazioni di analisi testuale, come *content analysis*, *text mining*, ecc. Si pensi, ad esempio, alla complessità di interrogazioni espresse nel linguaggio XQuery quando il documento è rappresentato con una notazione di tipo *stand-off* e si rende necessario fare contemporaneamente una ricerca sul testo e su due diversi livelli di annotazioni. Sono state proposte delle estensioni a XPath e XQuery per ridurre questi problemi (ad esempio Iacob e Dekhtyar 2005), ma queste estensioni sono ancora limitate per quello che riguarda lo sviluppo di applicazioni complesse di analisi testuale.

3 Panoramica del modello

La proposta è quella di utilizzare XML come standard di fatto di interoperabilità, per lo scambio di documenti, e di *abbondonare il concetto di marcatura come strumento principale di rappresentazione di testi*, usando invece un modello e di un linguaggio basati su *oggetti testuali astratti*, ispirato ai modelli ad oggetti utilizzati da tempo in informatica per padroneggiare la complessità dei sistemi software. Per ragioni di spazio riportiamo qui le principali caratteristiche di questa proposta, rimandando il lettore a Maurizio 2009; Maurizio e Orsini 2010 per maggiori dettagli.

La proposta prevede:

1. Un modello in cui il testo viene visto con un duplice aspetto: come una serie di caratteri, e insieme come un qualunque numero di gerarchie di strutture logiche, "appoggiate" su di essi, dette "oggetti testuali", entità software con un tipo, dotate di stato e di comportamento. Lo stato di questi oggetti rappresenta la parte di testo coinvolta, e una serie di proprietà, che possono essere oggetti testuali componenti, oppure attributi che possono assumere valori di qualunque tipo. Il comportamento è costituito da una collezione di procedure locali, che definiscono proprietà calcolate o eseguono operazioni sugli oggetti.

2. Un linguaggio di programmazione ad alto livello con un sistema di tipi statico che incorpora gli oggetti testuali, e un insieme di operatori su di essi. Una collezione persistente di documenti testuali ("digital library") può essere importata in un programma, e i suoi oggetti possono essere manipolati all'interno del programma.

3. Un modello per la memorizzazione efficiente e permanente di oggetti testuali.

4. Un'architettura di riferimento per un sistema che includa strumenti per lo scambio con altri sistemi attraverso XML, interfacce grafiche, gestione della persistenza e dell'accesso concorrente ai dati.

4 Confronto con lavori precedenti

In letteratura sono presenti varie proposte con un approccio confrontabile a quello qui presentato. Fra questi citiamo i due recenti lavori Schmidt 2012 e Boschetti e Grosso 2014.

Il primo prevede l'abbandono della marcatura inserita nel testo e l'uso intensivo di marcatura di tipo *stand-off*, combinata con strumenti automatici per la fusione dei documenti. Il secondo presenta uno strumento software per la trasformazione di testo annotato di tipo *stand-off* in una struttura basata su oggetti, per permettere elaborazioni complesse e visualizzazioni sofisticate di documenti testuali.

Approcci simili sono presentati anche in Coombs, Renear e DeRose 1987b; DeRose et al. 1997, con un modello in cui il testo è visto come una o più gerarchie di oggetti, come base per la fondazione di sistemi complessi come quelli di Carletta et al. 2003; Petersen 2002; Deerwester, Waclena e LaMar 1992.

Rispetto a queste, la proposta qui presentata prevede una soluzione complessiva, formata non solo da un modello per la rappresentazione dei dati, inclusi dati non testuali, ma anche da un linguaggio definito esplicitamente per i problemi di analisi testuale, e un'architettura che prevede la memorizzazione dei dati in una base di dati persistente, per permettere annotazioni concorrenti e condivise da parte di molti utenti, ricercabili e visualizzabili insieme al testo.

5 Conclusioni

Un primo prototipo di linguaggio, insieme ad alcune componenti del sistema, è stato realizzato e sperimentato. È stato studiato anche un approccio alternativo basato su un'interfaccia per linguaggi tradizionali (interfaccia API). Inoltre, sono state sviluppate alcune interfacce grafiche basate su web. Il lavoro da fare è ancora molto: le prossime fasi riguarderanno una ri-ingegnerizzazione del modello e del linguaggio, aperto ai contributi dei ricercatori di area umanistica, strumenti per l'acquisizione (semi)-automatica dei dati in formato XML, sviluppo di interfacce grafiche per utenti non esperti, ottimizzazione del modello di memorizzazione dei dati.

Bibliografia

Boschetti, F., e A. M. D. Grosso. 2014. «TeiCoPhiLib: A Library of Components for the Domain of Collaborative Philology». *Journal of the Text Encoding Initiative*, n. 8.

Carletta, J., et al. 2003. «The NITE XML Toolkit: flexible annotation for multi-modal language data». Special issue on Measuring Behavior, *Behavior Research Methods, Instruments, and Computers* 35 (3).

Coombs, J. H., A. H. Renear e S. J. DeRose. 1987a. «Markup systems and the future of scholarly text processing». *Commun. ACM* (New York, NY, USA) 30 (11): 933–947. ISSN: 0001-0782. doi:http://doi.acm.org/10.1145/32206.32209.

— . 1987b. «Markup Systems and the Future of Scholary Text Processing.» *Commun. ACM* 30 (11): 933–947. http://dblp.uni-trier.de/db/journals/cacm/cacm30.html#CoombsRD87.

Daniels, P. 1993. «The Unicode Consortium: The Unicode standard». *Language: journal of the Linguistic Society of America* 69 (1): 225–225.

Deerwester, S. C., K. Waclena e M. LaMar. 1992. «A Textual Object Management System.» In *SIGIR*, a cura di N. J. Belkin, P. Ingwersen e A. M. Pejtersen, 126–139. ACM. ISBN: 0-89791-523-2.

DeRose, S. J. 2004. «Markup Overlap: A Review and a Horse.» In *Extreme Markup Languages*. http://dblp.uni-trier.de/db/conf/extreme/extreme2004.html#DeRose04.

DeRose, S. J., et al. 1997. «What is text, really?» *ACM SIGDOC Asterisk Journal of Computer Documentation* 21 (3): 1–24.

Iacob, I. E., e A. Dekhtyar. 2005. «Processing XML documents with overlapping hierarchies». In *Proceedings of the 5th ACM/IEEE-CS joint conference on Digital libraries*, 409. ACM.

Iacob, I. E., A. Dekhtyar e W. Zhao. 2004. *XPath Extension for Querying Concurrent XML Markup*. Rapp. tecn. Citeseer.

Ide, N. 1994. «Encoding standards for large text resources: The Text Encoding Initiative.» In *COLING*, 574–578.

Maurizio, M. 2009. «Manuzio: an Object Language for Annotated Text Collections». Tesi di dott., Dipartimento di Informatica, Università Ca' Foscari Venezia.

Maurizio, M., e R. Orsini. 2010. «Manuzio: A Model for Digital Annotated Text and Its Query/Programming Language». In *Proceedings of the 14th European Conference on Research and Advanced Technology for Digital Libraries*, 478–481. ECDL'10. Glasgow, UK: Springer-Verlag. ISBN: 3-642-15463-8, 978-3-642-15463-8. http://dl.acm.org/citation.cfm?id=1887759.1887834.

Petersen, U. 1999. «The Extended MdF model». Available online, Aarhus University, Denmark. http://emdros.org/New-EMdF.pdf.

— . 2002. «The Standard MdF Model». *Unpublished article. Obtainable from URL: http://emdros.org*.

— . 2004a. «Emdros: a text database engine for analyzed or annotated text». In *COLING '04: Proceedings of the 20th international conference on Computational Linguistics*, 1190. Geneva, Switzerland: Association for Computational Linguistics. doi:http://dx.doi.org/10.3115/1220355.1220527.

— . 2004b. «MQL Programmer's Guide». Available online.

— . 2004c. «MQL User's Guide». No more available online.

Schmidt, D. 2012. «The role of markup in the digital humanities». *Historical Social Research/Historische Sozialforschung*:125–146.

Sperberg-McQueen, C., L. Burnard e S. Bauman. 1994. *Guidelines for Electronic Text Encoding and Interchange*. TEI Consortium.

Text Encoding Initiative Consortium. 2008. *TEI P5: Guidelines for electronic text encoding and interchange*.

Un innovativo graphic matching system per la ricerca in database di manoscritti antichi

Nicola Barbuti, Università degli Studi di Bari Aldo Moro, nicola.barbuti@uniba.it
Stefano Ferilli, Università degli Studi di Bari Aldo Moro, stefano.ferilli@uniba.it
Tommaso Caldarola, D.A.BI.MUS. S.r.l., t.caldarola@dabimus.com

1 Introduzione

Storicamente, lo sviluppo della conoscenza in ambito scientifico si è evoluto secondo due paradigmi: quello teorico e quello sperimentale. Negli ultimi due decenni, si sono affermati due ulteriori paradigmi: la *simulazione computazionale* (Ken Wilson, premio Nobel per la fisica 1982), il terzo paradigma dal quale sono originate le scienze computazionali; gli studi scientifici *data-driven based* (*data intensive scientific discovery*, Gordon Bell, 2012), il quarto paradigma che ha dato origine alle science informatics. Quest'ultimo, in particolare, è oggi ampiamente utilizzato nell'analisi di dati scientifici grazie alla crescente disponibilità di enormi quantità di dati, che permettono un approccio *in silico* finalizzato alla generazione di conoscenza.

Per applicare lo stesso cambiamento di prospettiva alle *humanities*, si deve partire dall'osservare che, negli ultimi decenni, uno sforzo notevole è stato dedicato alla generazione di grandi database cui è possibile accedere on-line. Alcuni esempi sono:

Thesaurus Linguae Graecae: raccoglie la letteratura greca da Omero (VIII sec. a.C.) alla caduta di Bisanzio (1453 d.C.) [http://stephanus.tlg.uci.edu];

Integrated Archaeological Database (IADB): risponde alle esigenze di gestione dei dati per tutta la durata della vita dei progetti di scavo archeologico [http://www.iadb.org.uk];

World Digital Library (WDL): raccoglie le versioni digitalizzate dei libri rari, mappe, manoscritti, fotografie [https://www.wdl.org/en];

Musisque Deoque: archivio digitale di poesia latina [http://www.mqdq.it/public];

Trismegisto: banca dati relativa a documentazione su papiro ed epigrafica [http://www.trismegistos.org].

Tutti questi database propongono meccanismi di interrogazione dai diversi livelli di complessità, che forniscono agli studiosi supporto per ricerche per lo più specifiche (ad esempio, il recupero di tutte le poesie scritte utilizzando una certa metrica). Vale a dire, replicano l'approccio metodologico tradizionale delle *humanities* in base al quale è necessario che, preliminarmente all'interrogazione, lo studioso abbia già formulato precise ipotesi di ricerca, sulle quali si aspetta poi di ottenere conferme grazie all'utilizzo di tecnologie digitali.

L'approccio metodologico suggerito dal quarto paradigma è del tutto diverso: gli algoritmi si sviluppano e applicano per trovare nuove ipotesi di lavoro tramite la scoperta di *pattern* dedotti direttamente da database anche di grandi dimensioni.

Per esempio: gli algoritmi possono essere applicati al fine di identificare i gruppi (*cluster*) di testi poetici o letterari 'simili' tra loro per argomento, o usi linguistici, o formulari, in digital libraries biblioteconomiche o archeologiche, o in corpora letteraria digitali, dai quali inferire nuove ipotesi su cui avviare ricerche utilizzando approcci tradizionali.

Nel presente intervento si descrive il modulo di *graphic matching* M-Evo (Multi-Evolution) del sistema di riconoscimento digitale ICRPad (Brevetto n. 0001407881), sviluppato tra il 2010 e il 2013 dalla spin off dell'Università degli Studi di Bari Aldo Moro D.A.BI.MUS. S.r.l. con la collaborazione di ricercatori del Dipartimento di Studi Umanistici della medesima università.

Il modulo permette di interrogare grandi database di manoscritti storici applicando l'approccio metodologico definito dal quarto paradigma, grazie all'utilizzo di un algoritmo di graphic matching basato sul concetto di *shape-contour recognition* che non necessita di processi di segmentazione del contenuto dell'immagine. Il processo è stato testato con obiettivi differenti sia su medi e grandi database di varia documentazione, sia su singoli oggetti digitali riproducenti documentazione manoscritta e a stampa antica o moderna.

2 La prospettiva del *data science* negli studi sui *data humanities*: una proiezione ancora *in progress*

Alla risoluzione del problema del riconoscimento digitale di immagini riproducenti beni documentali manoscritti o a stampa antichi sono stati dedicati negli ultimi 15 anni numerosi progetti di ricerca finalizzati a creare sistemi OCR o *pattern recognition*.

Oltre a non avere prodotto risultati di rilievo in termini di efficace indicizzazione, i diversi prototipi hanno in comune il limite di presupporre quale obiettivo da raggiungere non lo sviluppo di sistemi che favoriscano nuovi approcci metodologici nello studio dei diversi segmenti del dominio di interesse del ricercatore, ma la sola costruzione di strumenti tecnologici che lo agevolino negli studi condotti sempre secondo metodi tradizionali. Inoltre, tutti i sistemi di riconoscimento a oggi sviluppati con tali finalità sono stati testati su database di piccole dimensioni oppure su singole risorse digitali.

Approcci che, con lo sviluppo di database sempre più ampi e complessi, risultano ormai inadeguati a soddisfare pienamente i bisogni di studiosi e ricercatori.

Nello specifico, le ricerche di settore si sono basate principalmente su due diversi processi di sviluppo: a. segmentale; b. olistico.

a. L'approccio segmentale è il più utilizzato in un ampio numero di prototipi, per lo più strutturati secondo lo Hidden Markov Model (HMM).

b. L'approccio olistico è preferito in alcune più recenti sperimentazioni che hanno prodotto risultati di rilievo per percentuale di contenuto riconosciuto, ma i test sono stati eseguiti su campioni di immagini del tutto esigui e perciò non assumibili come significativi.

Entrambi presentano i seguenti limiti comuni, che a nostro parere ne ridimensionano notevolmente l'implementabilità in strumenti in grado di costruire nuovi percorsi di studio e ricerca:

- i diversi prototipi sono stati testati su campioni di immagini quantitativamente irrilevanti e riproducenti contenuti omografi;

- la fase di sviluppo necessita di attività manuali preliminari lunghe e complesse, che inevitabilmente incidono su risultati apparentemente rilevanti: l'eccessivo lavoro manuale, infatti, limita notevolmente la possibilità di utilizzare questi sistemi su database di grandi dimensioni, limitandoli a singole risorse digitali o a database contenenti poche risorse e del tutto omogenee tra loro, in quanto diversamente il lavoro umano preliminare sarebbe del tutto insostenibile;

- i campioni utilizzati nelle diverse ricerche sono stati selezionati accuratamente tra gli oggetti digitali maggiormente adatti al tipo di sperimentazione (immagini singole con grafia manoscritta o a stampa assolutamente omografa e uniforme, quanto più possibile scevre da rumorosità e ulteriormente ripulite con accurati interventi di post processing): ovviamente, non vi sono indicatori che attestino altrettanta qualità di risultati qualora i prototipi siano utilizzati su database di medie o grandi dimensioni contenenti immagini eterogenee.

3 Il modulo M-Evo del sistema ICRPad

L'algoritmo utilizzato nel modulo M-Evo di ICRPad è stato sviluppato avendo quale obiettivo la costruzione di uno strumento tecnologico che consentisse agli studiosi di paleografia di avvalersi nelle proprie ricerche dei database digitali esistenti, interrogandoli sia secondo metodi di approcci tradizionali (primo e secondo paradigma), sia utilizzando l'approccio definito dal quarto paradigma, del tutto nuovo nel dominio di riferimento, di modo da poter inferire nuove o inattese ipotesi di ricerca dall'analisi dei dati risultati dall'interrogazione dei database. L'algoritmo si basa sul concetto di *shape contour recognition*, che consente di evitare laboriose attività manuali o complessi training preliminari per la segmentazione del layout e il riconoscimento delle regioni grafiche. L'utente seleziona direttamente sul layout di un'immagine da lui preliminarmente scelta una regione grafica, che l'algoritmo codifica come lo *shape model* da utilizzare quale chiave di ricerca per recuperare regioni omografe o graficamente simili in una o più immagini di destinazione.

Per eseguire il matching con le immagini di destinazione, l'algoritmo utilizza non i valori in scala di grigio dell'immagine, ma i pixel della forma che costituisce il modello scelto dall'utente e il parametro del numero di livelli della piramide che ne strutturano la rappresentazione iconica.

In tal modo, il processo di interrogazione del modulo M-Evo consente la massima efficacia nella ricerca e, contestualmente, le più ampie potenzialità di effettuarla sia secondo metodi tradizionali che secondo il quarto paradigma, in quanto: permette di collegarsi *real time* come client a *n* database esistenti on line le cui immagini sono fruibili liberamente, grazie alla funzione di selezione e scelta di "repository" prevista nel sistema;

consente di visualizzare ed esplorare le immagini contenute nei diversi database per valutare eventuali elementi di interesse, anche secondo scelta casuale, da selezionare per creare *shape models* da utilizzare quali chiavi di ricerca; consente di variare, modulare e personalizzare in qualsiasi momento i parametri di setting per la ricerca, la quantità e la qualità delle

risposte, in relazione alle attese di maggiore o minore quantità di dati da rilevare (soglie di deformazione, etc.);

consente di creare gli shape models in tempo reale secondo le esigenze dell'utente: visualizzate una o più immagini, egli può selezionare le regioni di interesse direttamente sulle immagini e modellarle secondo le sue necessità (fermarsi a un singolo grafo, comprendere più grafi, un'intera parola, etc.); un tool di rilevazione delle rumorosità dell'immagine gli consente di verificare i livelli di "sporcizia" che potranno in qualche modo compromettere l'affidabilità della ricerca;

consente di personalizzare le ricerche salvando le regioni selezionate e utilizzate come modelli per la ricerca in apposita repository di sistema.

Sono stati eseguiti numerosi test per verificare le funzionalità del sistema e la sua validità. I test hanno riguardato sia oggetti digitali riproducenti diverse tipologie documentali manoscritte e a stampa antiche e moderne, sia database di diversa dimensione e complessità.

Come detto all'inizio, in questa sede si presentano i risultati delle sperimentazioni eseguite in ambiti della ricerca paleografica, privilegiando l'approccio metodologico definito dal quarto paradigma.

Abbiamo ipotizzato un metodo che permette agli studiosi di paleografia di eseguire ricerche in diversi tipi di database interrogandoli secondo un approccio *assumption-free*: ciò significa che il sistema non è destinato a trattare con un singolo, specifico, pre-definito database, ma può essere collegato in tempo reale con molteplici database disponibili on-line, che l'utente seleziona al momento e a sua discrezione in quanto potenzialmente rilevanti per i suoi obiettivi di ricerca.

Dopo aver selezionato i database, l'utente sceglie l'immagine (o le immagini) dalla quale estrarrà i grafi da utilizzare come modelli per l'interrogazione. Quindi, estratti casualmente i modelli grafici grazie al tool di creazione automatica *create model*, lancia le sue interrogazioni nei database collegati.

Nella cornice classica, l'interrogazione dovrebbe essere effettuata presupponendo una determinata ipotesi di ricerca, con l'obiettivo di verificare se i risultati della query la confermino. Nell'impostazione metodologica utilizzata nella sperimentazione, la ricerca è casuale, cioè senza alcuna aspettativa sui risultati, e le suggestioni per la costruzione di nuove ipotesi di ricerca sono dedotte dall'analisi dei dati risultanti dall'interrogazione.

Questa metodologia può anche presupporre modalità di interrogazione dei database partendo da precise ipotesi di ricerca, ma, ottenuti risultati positivi, ulteriori ipotesi di studio inizialmente impreviste possono essere dedotte anche dall'analisi attenta dei risultati cosiddetti 'falsi positivi': questi ultimi, infatti, possono rivelarsi di grande interesse, perché alcuni di loro, anche se formalmente diversi rispetto al modello utilizzato, possono tuttavia rivelare somiglianze non facilmente rilevabili (o del tutto non rilevabili) a occhio nudo, ma nondimeno reali, che aprono la via a ipotesi diverse e interessanti da indagare. Sono, questi ultimi, due degli approcci del *data science* che aprono la via al rinnovamento metodologico nello studio di *data humanities* tramite la consultazione di banche dati, e aprendo nuove frontiere di ricerca.

Si illustra di seguito la sperimentazione eseguita. Quale contesto di riferimento, si è considerata la metodologia di studio di un ricercatore impegnato in studi paleografici e storici sui manoscritti, che si propone di esplorare nuove ipotesi ricerca relative al dominio

di suo interesse. A tal fine, decide di interrogare alcuni database on-line al fine di raccogliere indicazioni utili per la sua ricerca dedotte dai risultati delle sue interrogazioni.

Utilizzando il modulo M-Evo di ICRPad, lo studioso si è connesso a un database registrato on-line, per esplorare un gruppo di sette manoscritti latini (che denomineremo con lettere alfabetiche da A a G), notoriamente considerati del tutto diversi tra loro in quanto provenienti da scriptoria di diversa dislocazione geografica e prodotti tra i secoli XI e XIII. Obiettivo dell'interrogazione è verificare se vi sia qualche possibilità che, usando l'algoritmo di matching, possano essere dedotte dall'analisi dei dati risultanti dall'interrogazione istanze tali da consentire di formulare nuove ipotesi di ricerca non assumibili tramite gli studi tradizionali. In tempo reale, lo studioso si connette al repository in cui sono memorizzati i codici, sceglie casualmente un'immagine digitale di una pagina da uno di questi, che chiameremo ms A, e tramite il tool *create model* estrae automaticamente da questa immagine una regione in cui è rappresentato il glifo &, che potrebbe essere comune a tutti i manoscritti del gruppo di suo interesse. Questo costituisce il modello da utilizzare per lanciare la query al fine di verificarne l'eventuale presenza con le stesse caratteristiche grafiche in altri manoscritti. Salva il modello appena creato nella sua repository personalizzata inclusa nel sistema, quindi lancia la ricerca. L'algoritmo scansiona l'intero gruppo e restituisce tutte le occorrenze che soddisfano i parametri da lui utilizzati e sono quindi considerate uguali o simili al modello.

La sperimentazione ha prodotto i seguenti risultati:

- corrispondenze rilevate con il ms A, glifo &:
  - mss B ed E:

    veri positivi (omografi di &): 90% per ognuno

    falsi positivi: 10% per ognuno; tra questi, alcuni presentano tratti perfettamente sovrapponibili ai corrispondenti del modello cercato (per es., il glifo a, o il gruppo 'et' con legatura tra i grafi che lo rende assimilabile all'& classico, ma non uguale, in cui la curva inferiore risulta essere del tutto sovrapponibile a quella corrispondente del modello)
  - ms D:

    veri positivi (omografi di &): 70%

    falsi positivi: 30%; tra questi, alcuni presentano tratti perfettamente sovrapponibili ai corrispondenti del modello cercato (per es., il glifo a)
  - mss C, F, G:

    veri positivi (omografi di &): 0% per ognuno

    falsi positivi: 100% per ognuno (10% in ms A, il 15% in ms B); tra i quali:

I risultati "veri positivi", sia pure incerti, potevano essere in qualche modo "attesi" dall'utente al momento del lancio della ricerca, e tuttavia potrebbero non essere sufficienti, da soli, a sostenere l'ipotesi che entrambi i manoscritti siano opera di un solo amanuense. In tale direzione, i risultati falsamente positivi, del tutto inattesi, assumono un rilievo decisivo, in quanto risultano contenere tratti perfettamente omografi rispetto al modello, il che è possibile solo ipotizzando una modalità scrittoria del tutto identica per entrambi

i manoscritti. Ne rileva che, approfondendo l'analisi dei risultati dell'interrogazione, dati normalmente definibili come "rumorosità" che pregiudicano l'efficacia di un algoritmo diventano invece fonti attendibili da cui dedurre nuove ipotesi di ricerca altrimenti non formulabili, in quanto comunemente rifiutate dagli studiosi che operano secondo approcci metodologici tradizionali.

I dati analizzati, infatti, sollecitano ulteriori approfondite indagini sia sul digitale che sul manufatto fisico, e diventano il punto di partenza per formulare nuove ipotesi sulla paternità dei due manoscritti, per esempio:

- che alcuni dei manoscritti siano opera di uno stesso amanuense operativo in un singolo scriptorium in un determinato arco di tempo, o in scriptoria differenti in archi temporali non distanti tra loro;
- che alcuni dei manoscritti siano opera di due diversi amanuensi attivi in tempi diversi (anche nel corso di secoli diversi) nel medesimo scriptorium, che perciò hanno utilizzato il medesimo canone riuscendo a perfezionare la loro abilità scrittoria al punto da rendere le grafie quasi omografe;
- che alcuni dei manoscritti siano stati prodotti nel medesimo periodo storico ma in due scriptoria diversi, nei quali era utilizzato il medesimo canone;
- che alcuni dei manoscritti siano stati prodotti in secoli diversi e in scriptoria differenti, ma collocati nella stessa area geografica, che quindi avrebbero utilizzato e perfezionato un canone scrittorio comune mantenutosi nel corso del tempo;

e così via ipotizzando.

Ovviamente, non spetta a noi fornire risposte precise a queste possibili istanze. Quanto interessa è aver cercato di mostrare in questa sede come il modulo di graphic matching M-Evo del sistema ICRPad renda possibile applicare le metodologie di ricerca del *data science* alla ricerca sui *data humanities*, e come questo nuovo approccio metodologico possa realmente apportare una decisa e decisiva innovazione anche nella stessa formazione nella ricerca umanistica, a oggi ancora saldamente ancorata alle metodologie tradizionali.

Bibliografia

Adamek, T., E. N. O' Connor e A. F. Smeaton. 2007. «Word matching using single closed contours for indexing handwritten historical documents». *International Journal of Document Analysis and Recognition (IJDAR)* 9 (2–4): 153–165.

Barbuti, N., e T. Caldarola. 2012. «An innovative character recognition for ancient book and archival materials: A segmentation and self-learning based approach.» In *Digital Libraries and Archives, IRCDL 2012*, a cura di M. Agosti et al., 354:261–270. Communications in Computer and Information Science. Heidelberg: Springer.

Fischer, A., e H. Bunke. 2011. «Character prototype selection for handwriting recognition in historical documents I». In *Proceedings of 19th European Signal Processing Conference (EUSIPCO)*, 1435–1439.

Hey, T., S. Tansley e K. Tolle. 2009. *The Fourth Paradigm: Data-Intensive Scientific Discovery*. WA: Microsoft Research.

Indermühle, E., M. Eichemberger-Liwicki e H. Bunke. 2008. «Recognition of Handwritten Historical Documents: HMM-Adaptation vs. Writer Specific Training». In *Proceedings of 11th International Conference on Frontiers in Handwriting Recognition*, 186–191. Montreal, Quebec, Canada.

Le Bourgeois, H., F.and Emptoz. 2007. «DEBORA: Digital AccEss to BOoks of the RenaissAnce». *IJDAR* 9 (2–4): 193–221.

M., B., e S. L. 2007. «Automatic Handwriting Identification on Medieval Documents». In *ICIAP 2007: 14th International Conference on Image Analysis and Processing*, 279–284.

Rath, M. T., R. Manmatha e V. Lavrenko. 2004. «Search Engine for Historical Manuscript Images». In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 369–376.

Srihari, S., C. Huang e H. Srinavasan. 2005. «A Search Engine for Handwritten Documents». *Document Recognition and Retrieval XII* 154 (3): 66–75.

Stokes, P. A. 2009. «Computer-aided Palaeography, Present and Future». In *Codicology and Palaeography in the Digital Age, Schriften des Instituts fur Dokumentologie und Editorik,* a cura di M. Rehbein, vol. 2. Norderstedt: Book on Demand GmbH.

Morphology beyond inflection. Building a word-formation-based lexicon for Latin

Eleonora Litta, CIRCSE - Università Cattolica del Sacro Cuore,
eleonoramaria.litta@unicatt.it
Marco Passarotti, CIRCSE - Università Cattolica del Sacro Cuore,
marco.passarotti@unicatt.it
Chris Culy, independent scholar, chrisculy@mac.com

In the construction of computational language resources, derivational morphology has constantly been overlooked compared to inflectional morphology, which plays a central role in fundamental annotation tasks such as PoS tagging. Yet enhancing textual data with derivational morphology tagging has the potential to provide solid results.

The study of derivation allows for a fine-grained organisation of the lexicon by linking words that share the same lexical ancestor or the same word formation process. Likewise, because core semantic properties are shared at different extent by derivate words built by a common word formation process, derivational morphology can act like a kind of interface between morphology and semantics. This especially holds true in the context of linguistic investigations of textual material.

In recent times, the computational linguistics world has gradually been focussing its interests on researching and building new derivational morphology resources and tools. Among them are the lexical network for Czech DeriNet (Ševcıková e Zabokrtský 2014), the derivational lexicon for German DErivBASE (Zeller, Snajder e Padó 2013) and that for Italian derIvaTario (Talamo, Celata e Bertinetto 2016).

On the Classical languages front, although the number of lexical resources and NLP tools, especially for Latin, is now large and varied (corpora, treebanks, computational lexica, and digital libraries)[1], there has not been any attempt to create a derivational morphology tool, where lemmas are segmented and analysed into their derivational components, to establish relationships between them on the basis of word formation.

The first steps towards building such a word formation lexicon for Latin were made by Passarotti e Mambrini (2012), who described a model for the semi-automatic extraction of word formation rules from the list of lemmas of Lexicon Totius Latinitatis by Forcellini (1940, fifth edition) and the subsequent pairing of lexical entries and their derivational ancestor(s).

[1] See for example the *Ancient Greek Dependency Treebank* (https://perseusdl.github.io/treebank_data), PROIEL corp us (http://foni.uio.no:3000/users/sign_in), *Index Thomisticus Treebank* (*http://itreebank.marginalia.it*), *Thesaurus Linguae Latinae* (http://www.thesaurus.badw.de/english), *Latin WordNet* (as part of Multi-WordNet http://multiwordnet.fbk.eu/english/home.php), *Perseus Digital Library* (http://www.perseus.tufts.edu)

In this context, the Word Formation Latin project has received funding from the European Union's Horizon 2020 research and innovation programme (Marie Sklodowska-Curie grant agreement No 658332-WFL) to create the definitive derivational lexicon for Classical Latin. This will ultimately be included in the automatic morphological lemmatiser for Latin LEMLAT (http://www.ilc.cnr.it/lemlat/lemlat/index.html, accessed 20/05/2016), creating a 360° resource for the study of Latin morphology.

This proposal contains a brief description of the methodology used to build the resource, followed by a preview on how the data will be ultimately accessed and presented to the user in the finished resource.

The lexical basis used for the word formation based resource is the one featured in LEMLAT (http://www.ilc.cnr.it/lemlat/lemlat/index.html), which results from the collation of three Latin dictionaries (Georges and Georges, 1913-1918; Glare 1982; Gradenwitz 1904). It counts 40,014 lexical entries and 43,432 lemmas (as more than one lemma can be included into the same lexical entry). Recently, the lexical basis of Lemlat was further enlarged by adding 26,250 lemmas from the Onomasticon provided by Forcellini (1940).

The word formation lexicon is built in two steps:

- Word formation rules are detected.
- They are applied to lexical data.

Word formation rules (WFRs) are grouped in two classes: 1. compounding; 2. derivational. Derivational rules are divided in two further categories: a. affixal (in turn split into prefixal and suffixal), and b. conversion, a derivation process that affects PoS without including any affix. Compounding and conversion WFRs are automatically detected, by considering all the possible combinations of main PoS (verbs, nouns, adjectives). Affixal WFRs are found both according to previous literature on Latin derivational morphology (Jenks 1911; Fruyt 2011; Oniga 1988) and in semi-automatic fashion. Each morphologically derived lemma is assigned a WFR. All those lemmas that share a common (not derived) ancestor belong to the same "morphological family". For instance, lemmas *formatio* ("formation"), *formo* ("to form") and *formosus* ("beautiful", lit. "finely formed") all belong to the morphological family whose ancestor is the lemma *forma* ("form").

Lemmas and WFRs are paired by using a MySQL relational database, and a number of MySQL queries provide the candidate lemmas for each WFR.

Given the high number of homographs in Latin, the procedure described above is not necessarily sufficient for building the morphological families. Thorough manual checking allows the identification of false results, duplication and lacunas resulting from the automatic process. Such duplicate results need to be analysed and rectified.

Some morphotactically obscure word formation processes, like some kinds of compounding, need to be completely hardcoded. The word formation based lexicon is accessible on-line through a visualization query system (temporarily at http://wfl.marginalia.it). The lexicon can be browsed either by WFR, affix, input and output PoS or lemma. Drop down menus provide the available options for each selection, like for instance the list of affixes and lemmas. Results are visualized as tree graphs, whose nodes are lemmas and edges are WFRs. Trees are interactive. Clicking on a node shows the full derivational tree ("word formation cluster") for the lemma reported in that node. For example, Fig. 1 shows the word formation cluster for the lemma *bellum*. Clicking on an edge shows the lemmas

built by the WFR concerned in the edge. Lemmas are provided both as a derivational graph and as an alphabetical list. For instance, double clicking on the edge going from *bello* to *rebello* in Fig. 1 shows the lemmas built by the derivational WFR that builds new verbs from first conjugation verbs (V1) with prefix re–. Fig. 2 presents a close-up of the derivational graph for this rule.

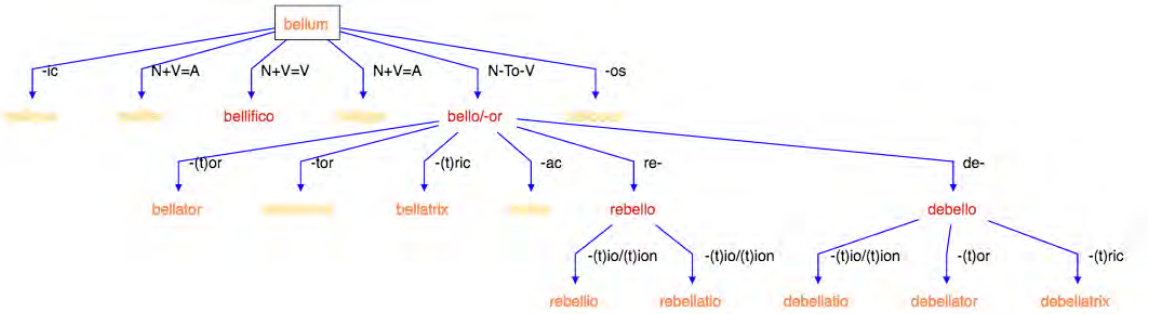


Figura 1: Word formation cluster for *bello*

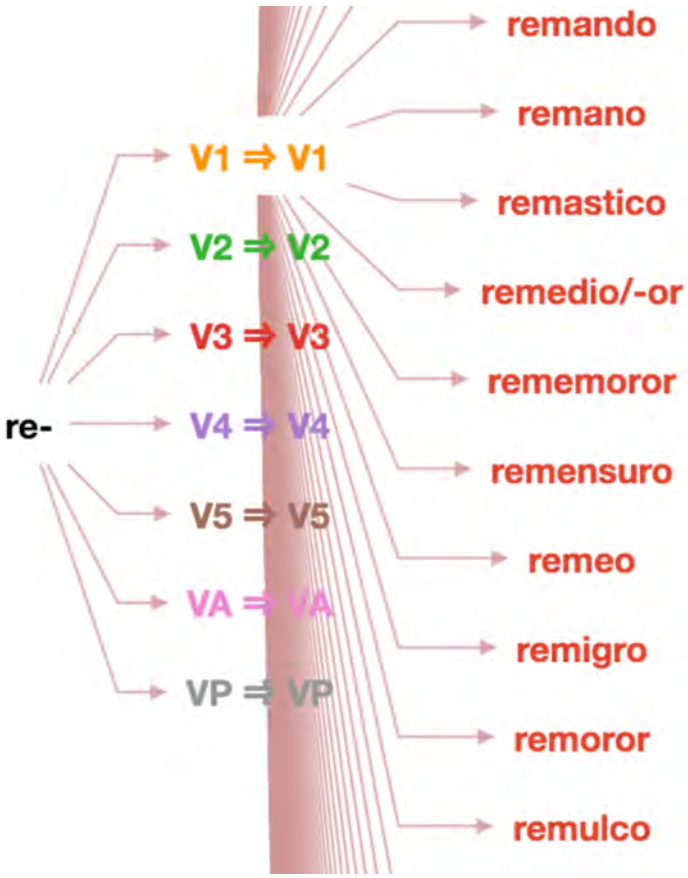


Figura 2: Derivational graph for a WFR

Phase 2 in the project workflow will be to integrate the information extracted from the resulting derivational morphological lexicon into the morphological layer of annotation of the *Index Thomisticus Treebank* (IT-TB). The *Index Thomisticus* (IT) is considered a pioneer in digital humanities: started by father Roberto Busa SJ in 1949, it is a database retaining the *opera omnia* by Thomas Aquinas (118 texts), plus 61 works by other authors related to Thomas.

The integration of the word formation-based information into the IT-TB will be executed through the embedding of the lexicon data within the morphological layer of annotation of the treebank, using TEI P5 conformant XML encoding to favour data exchange and linking to other lexical resources.

The final resource will be both a standalone lexicon and tool accessible through its own website, and interconnected with the IT-TB. Finally the whole resource will be made available through the CLARIN infrastructure (www.clarin.eu).

Bibliografia

Forcellini, A. 1940. *Lexicon totius latinitatis ab Aegidio Forcellini seminarii Patavini alumno lucubratum, deinde a Iosepho Furlanetto eiusdem seminarii alumno emendatum et auctum, nunc vero curantibus Francisco Corradini et Iosepho Perin seminarii Patavini item alumnis emendatius et auctius melioremque in formam redactum*. Vol. 1. Patavii: Typis Seminarii.

Fruyt, M. 2011. «Word-Formation in Classical Latin». *A Companion to the Latin Language*:157–175.

Glare, P. G. 1982. *Oxford latin dictionary*. Clarendon Press. Oxford University Press.

Gradenwitz, O. 1904. *Laterculi vocum Latinarum: voces Latinas et a fronte et a tergo ordinandas*. S. Hirzel.

Jenks, P. R. 1911. *A manual of Latin word formation for secondary schools*. DC Heath & Company.

Oniga, R. 1988. *I composti nominali latini: una morfologia generativa*. Vol. 29. Pàtron.

Passarotti, M. C. 2004. «Development and perspectives of the Latin morphological analyser LEMLAT». *Linguistica computazionale* 20 (A): 397–414.

Passarotti, M., e F. Mambrini. 2012. «First Steps towards the Semi-automatic Development of a Wordformation-based Lexicon of Latin.» In *LREC*, 852–859. Citeseer.

Ševcıková, M., e Z. Zabokrtský. 2014. «Word-Formation Network for Czech». In *Proceedings of LREC*.

Talamo, L., C. Celata e P. M. Bertinetto. 2016. «DerIvaTario: An annotated lexicon of Italian derivatives». *Word Structure* 9 (1): 72–102.

Zeller, B. D., J. Snajder e S. Padó. 2013. «DErivBase: Inducing and Evaluating a Derivational Morphology Resource for German.» In *ACL (1)*, 1201–1211.

Indice degli Autori

Index of the Authors

Autori / *Authors*

ASSOCIAZIONE PER
L'INFORMATICA UMANISTICA
E LA CULTURA DIGITALE

ISBN 978-88-942535-0-4