FEDERICO GIUSFREDI / ALFREDO RIZZA
(Hethitisches Wörterbuch, Institut für Assyriologie und Hethitologie, Ludwig-Maximilians-Universität München –
Dep. of Linguistics, UCB, Rotary International Ambassadorial Scholar)

Zipf's Law and the Distribution of written Signs

The purpose of the present paper is to extend the analysis of the rank-frequency distribution of linguistic elements (started with the analysis of word-distribution by Zipf, 1949, *Human Behavior and the Principle of Least-Effort*) to the domain of elements of writing systems.
By concentrating on linguistic units that we will call "Interpreted Graphic Signs" (IGS), corresponding to a functionally interpreted set of glyphs (where for instance an uppercase *A*, marking a proper name or the beginning of a sentence, results different from a lowercase *a*), we will analyze some different language-script couples (for instance Sumerian-Cuneiform or Italian-Latin) in order to demonstrate that the extreme regularity of the Zipf's distribution stating that $r \propto f^{-1}$ is generally absent in IGS distribution; nevertheless, depending on the language and script typology, the data sets will tend to a higher "zipfian" regularity in cuneiform Sumerian, an agglutinative language written with a logo-syllabary and characterized by the fact that every IGS on a graphic level almost corresponds to a morpheme (in some instances a lexeme) on a linguistic one ($I : M \approx 1$). This phenomenon, observed in experimental data-sets, has being tested through two different mathematical estimations, and it seems to point to the evidence that the regular distribution of linguistic elements is actually a semantic(-related) fact.
We pursued an attempt to estimate the rank-frequency distribution also with Uninterpreted Graphic Signs (UGS) in order to formulate a hypothetical fill to a need in a general algorithm we proposed at the ECAL conference in Prague 2007 in the aim of an integration of artificial intelligence in the constitution of electronic corpora for ancient and unknown languages.

MANUELA ANELLI / MARTA MUSCARIELLO / GIULIA SARULLO
(Istituto di Scienze dell'Uomo, del Linguaggio e dell'Ambiente, Libera Università di Lingue e Comunicazione IULM, Milano)

The Digital Edition of Epigraphic Texts as Research Tool: the ILA Project

In October 2007, the ILA Project (Iscrizioni Latine Arcaiche – A Digital Corpus of the Archaic Latin Inscriptions) directed by Professor Giovanna Rocca of the IULM University in Milan was presented at the 32nd Congress of the Società Italiana di Glottologia.
This project originates from the awareness that the Web is a particularly suitable place for the edition of epigraphic *corpora*; it represents a novelty in the field of digital epigraphy in that it will be the first publication of the entire *corpus* of the archaic Latin inscriptions from *Latium vetus* dating from the VII to the V century BC; moreover, the ILA Project will be the first one in Italy using the EpiDoc encoding, conceived to meet the peculiar requirements of digital epigraphy. Actually, it will be necessary to dwell upon these specific requirements, in that working with this *corpus* – made up by approximately eighty inscriptions and fragments – is problematic because of the frequent fragmentary nature of the materials, sometimes in bad repair, the plurality of the alphabets, the archaic character of the language and the *lacunae* of the texts.
These problems, that emerged during the process of encoding the archaic Latin inscriptions according to the EpiDoc specifications, are not present in later epigraphy; they will be discussed in this occasion – together with the procedures through which we tried to overcome them – by means of a few practical examples.
Finally, we will illustrate the structure of the website which will host the digital publication, both as regards the single epigraphic charts and the overall framework, drawing the attention to the most important characteristics for research.

BIBLIOGRAFIA

Articoli e relazioni sul lavoro della Commissione "*Epigraphie et Informatique*" dell'AIEGL (1997-2003) disponibili anche sul sito www.edr-edr.it :
"Epigraphica" 60, 1998, pp. 316-317.
"Epigraphica" 61, 1999, pp. 311-313.
"Epigraphica" 65, 2003, pp. 350-355.

ATTENNI, LUCA – MARAS, DANIELE, *Materiali arcaici dalla collezione Dionigi di Lanuvio*, in "Studi Etruschi" 70, 2005, pagg. 68-78.
HARTMANN, MARKUS, *Die frühlateinischen Inschriften und ihre Datierung*, Bremen, 2005.
MARAS, DANIELE, *Novità sulla diffusione dell'alfabeto latino nel Lazio arcaico*, in *Theodor Mommsen e il Lazio antico. Giornata di Studi in memoria dell'illustre storico, epigrafista e giurista*, a cura di F.MANNINO - M.MANNINO - D.MARAS, Roma 2009, pp. 105-118.
MARAS, DANIELE, *Interferenze culturali arcaiche etrusco-latine: la scrittura*, in "Annali della Fondazione per il Museo «Claudio Faina» XVI (2009), pp. 309-331.
MUSCARIELLO, MARTA, *Iscrizioni latine arcaiche: a Digital Corpus of the Archaic Latin Inscriptions*, in "Alessandria", 2, 2008, pp. 213-217.
PROSDOCIMI, ALDO LUIGI, *Studi sul latino arcaico*, in "Studi Etruschi" 47, 1979, pagg. 173-183.
PROSDOCIMI, ALDO LUIGI, *Considerazioni su un libro recente di epigrafia romana*, in "Epigraphica", 46, 1984, pp. 252-263.
SUSINI, GIANCARLO, *Epigrafia romana*, Roma, Jouvence, 1982.
TISSONI, FRANCESCO, *EpiDoc e l'epigrafia latina sul web. Il progetto Iscrizioni Latine Arcaiche*, in "ACME", 2008 pp. 29-49.

MARGHERITA FARINA
(Dipartimento di Scienze Storiche del Mondo Antico, Università di Pisa)

Electronic analysis and organization of the Syro-Turkic Inscriptions of China and Central Asia

The Syro-Turkic inscriptions of China and Central Asia are a corpus of about 1000 inscriptions, dating between 708 and 1378 A.D. ca.,[1] found in an area including Kazakhstan, Kyrgyzstan, China (Xinjiang, Inner Mongolia, Quanzhou, Yangzhou). The inscriptions are written in an eastern variety of the Syriac alphabet, in Syriac and Turkic language. These inscriptions have become known to the Western world in the second half of the 19th century, thanks to the curiosity and study of a number of Russian scholars. Ever since, they have been studied almost uninterruptedly, both by Western and Chinese scholars. A number of publications were issued in the course of time, among which the impressive collection made by Chwolson (1886, 1890, 1897). However, the available material is still scattered in an number of journal articles and partial publications, while a comprehensive edition is still needed. In 2009 Pier Giorgio Borbone and Margherita Farina have elaborated an electronic database and the concordances of the entire corpus. This paper will describe the structure of the data, the functioning of the program *Obelix*[2] that was used to elaborate the concordances and the perspectives that this techniques offers for the philological studies and to corpus linguistics. A short sketch will be also given of other analogous applications of the system to other domains of the Semitic philology (such as Biblical Hebrew and Aramaic studies).

References

Borbone, P. G. and Mandracci, F. (1989). "An other way to analyze Syriac texts. A simple powerful tool to draw up Syriac computer aided concordances". *Proceedings of the II Conference Bible and Computer, Jerusalem, 9-13 June 1988*. Paris-Genève: Champion-Slatkine: 135-145.
Chwolson, D. (1886). *Syrische Grabinschriften aus Semirjetschie* (St. Petersburg).
Chwolson, D. (1890). *Syrisch-nestorianische Grabinschriften aus Semirjetschie* (St. Petersburg).
Chwolson, D. (1897). *Syrisch-nestorianische Inschriften aus Semirjetschie. Neue Folge* (St. Petersburg).

[1] The dating in the inscriptions is expressed according both to the Seleucid era (312-311 B.C.) and to the animal cycle of the Chinese constellations.
[2] For a description of the program cf. Borbone and Mandracci (1989).

MARIACHIARA PELLEGRINI / ALFREDO TROVATO
(Laboratorio del Lessico di Linguistica - Dipartimento di Linguistica, Letteratura e Scienze della Comunicazione, Università degli Studi di Verona)

Analisi informatica dei fenomeni di interferenza grafematica nelle iscrizioni di Selinunte

This paper aims to present the preliminary results of a linguistic study carried out on a *corpus* of greek inscriptions from Selinunte, thanks to the use of a dedicated software of analysis (B.A.S.P.). We attempt to show the possibilities opened up by the use of new technologies not only within Epigraphy but also Historical Linguistics. B.A.S.P. as a seriation and clustering tool enables to collect the graphic features of the signs composing the texts, in order to define a chronological order based on the graphic features themselves. The employment of this tool implies a preliminary work to individuate the typologies of allographic variants, based on a detailed analysis and classification of the signs. The analysis will be supported by a linguistic approach to the texts: this enables morphonological features to be highlighted as well as aspects of diastratic and diamesic dimensions belonging to the texts themselves, neither of which could be accounted for only on the basis of a merely epigraphic approach.
For the purpose of the case study presented here, it will be taken into consideration some inscriptions not only in Greek language but also in other languages of Italic substrate, which pertain more closely to the analysed *corpus*, namely Elimo.
The starting point of the research will consist in evaluating the interaction between different but co-existing scripts, which may imply, from a graphematic point of view, phenomena of *code switching*. The signs will be analysed in the perspective of multilingual dimension by means of a contrastive method drawing on the evidence from attested languages.

References

Adrados, F.R. (1990). *Nueva sintaxis del griego antiguo*, Gredos, Madrid.
Adrados, F.R. (2005). *A history of the Greek language*, Brill, Boston.
Arena, R. (1992). *Iscrizioni greche arcaiche di Sicilia e Magna Grecia, Iscrizioni di Gela e Agrigento*, Edizioni universitarie di lettere economia e diritto.
Bertolini, F. (2005). *Dialetti e lingue letterarie della Grecia arcaica*, Ibis, Studia ghislieriana, Pavia.
Christidis, A. (2007). *A history of ancient Greek: from the beginnings to late Antiquity,* Cambridge.
CIG *Corpus Inscriptionum Graecarum*, (1828-1877), Preußische Akademie der Wissenschaften, Berlin.
Consani, C. (2005). "Dialettalità genuina e dialettalità riflessa nel quadro delle più antiche attestazioni dei dialetti greci", in *Lingue e dialetti della Grecia arcaica*, Atti della IV giornata di filologia classica, (2004, Collegio Ghislieri), ed. F. Bertolini & F. Gasti, Pavia, Ibis Edizioni, (pp. 45-95).
Dubois, L. (1989). *Iscriptions Greques dialectales de Sicile*, École Française de Rome, Palazzo Farnese, Roma.
Dunbabin, T.J. (1948). *The Western Greeks*. Oxford.
Ghinatti, F. (1998). *Profilo di epigrafia greca: gli orizzonti della ricerca attuale*, Rubettino.
Giacomelli, R. (1988). *Achaea Magno-Graeca: le iscrizioni arcaiche in alfabeto acheo di Magna Grecia*. Studi grammaticali e linguistici. Paideia, Brescia.
Guarducci, M. (2005). *L'epigrafia greca dalle origini al tardo impero*. Libreria dello Stato, Roma.
Hondius J. J. E. (1923-1996). *Supplementum epigraphicum graecum,* Giebe, Amsterdam.
Jeffery, H. (1961) *The local scripts of archaic Greece*, Oxford Clarendon Press.
Meiggs, R. - Lewis, D. (1989). *A selection of Greek historical inscriptions*, Clarendon press, Oxford.
Pellegrini, M., Trovato, A. (2008) *Nuovi strumenti per un'indagine epigrafica runica: il caso del B.A.S.P*. Atti del IX seminario avanzato in Filologia Germanica, ed. Dell'Orso, Alessandria. (In corso di stampa).
SEG: *Supplementum Epigraphicum Graecum*.
Wachter, R. (2002). *Non-attic Greek vase inscriptions*, Oxford University Press. *Non-attic Greek vase inscriptions*, Oxford University Press.

FEDERICO BOSCHETTI
(Centro Interdipartimentale Mente/Cervello [CIMeC], Università degli Studi di Trento)

Modello collaborativo per migliorare l'accuratezza dell'OCR del Greco antico

Questo studio ha lo scopo di illustrare un modello collaborativo di correzione semiautomatica dell'OCR applicato allo studio della classicità.
Le grandi iniziative di digitalizzazione di testi non più coperti da copyright hanno reso disponibili ai filologi un grande numero di edizioni critiche, commentari, riviste e monografie. Le opere (o le parti di opera) scritte in caratteri latini sono fruibili tramite motori di ricerca, in quanto il testo, creato dall'OCR, è mappato sull'immagine della pagina. Al contrario, nella quasi totalità dei casi, le opere in Greco antico di cui siano disponibili scansioni dell'edizione originaria, sono fruibili soltanto come immagini.
Attualmente alcuni software open source, come Tesseract e Ocropus, ed alcuni software commerciali, come FineReader ed Anagnostis, sono in grado di fornire livelli di accuratezza superiori al 95% nel riconoscimento dei caratteri greci. Come è ampiamente dimostrato, almeno tre fattori possono far aumentare l'accuratezza dell'OCR:

1. un significativo training dei singoli software di OCR per adattarsi a condizioni specifiche, legate ai tipi di font, alla qualità della carta, etc.;
2. l'applicazione di algoritmi di allineamento ai diversi output dei singoli software per l'OCR;
3. l'impiego di un correttore ortografico automatico che ordini le proposte di correzione secondo la loro probabilità.

In primo luogo, la creazione di file di training, basati sulla correzione manuale degli errori prodotti dall'OCR su un campione di testo, richiede costi elevati in termini di tempo. In secondo luogo, l'installazione e la manutenzione dei software per l'OCR utilizzabili tramite API può richiedere delicati interventi tecnici. In terzo luogo, gli algoritmi di allineamento e i correttori ortografici da applicare possono essere variati.
Per questa ragione il modello collaborativo qui proposto prevede la creazione di webservices:

1. per conservare e distribuire i file di training fra diverse unità sulle quali sono installati i software per l'OCR;
2. per caricare sulle unità selezionate i file d'immagine su cui deve essere applicato l'OCR;
3. per applicare gli algoritmi di allineamento e la correzione ortografica agli output dell'OCR prodotto su diverse unità, ottenendo come risultato un testo digitale sempre più accurato da mappare sull'immagine digitale.

MATTEO ROMANELLO
(Centre for Computing in the Humanities, King's College London)

L'edizione critica digitale di frammenti: problemi teorici e soluzioni tecniche

Il modo in cui i testi fammentari sono attualmente rappresentati all'interno delle collezioni digitali di testi non solo risulta poco adeguato alla natura stessa dei fammenti, ma corre anche il rischio di falsare i risultati di analisi prodotte a partire da tali testi. Infatti, testo del testimone e testo del fammento allo stato attuale possono comparire più volte in una stessa collezione costituendo pertanto dei duplicati. Lo studio di una soluzione tecnica per un'adeguata rappresentazione digitale di testi fammentari si è rivelato una preziosa occasione per rifettere sulla natura stessa di questi testi, e in particolare sull'importanza della componente interpretativa nell'individuazione di un fammento (Berti et al. 2009).
Se consideriamo l'individuazione di un fammento come il futto di un atto flologico, e perciò interpretativo, una collezione digitale di testi che contenga anche dei fammenti deve poter rifettere la molteplicità delle interpretazioni prodotte dagli studiosi senza tuttavia produrre erronei duplicati.
Il risultato di tale studio è stato un modello a due livelli per la rappresentazione dei fammenti come oggetti digitali che renda giustizia dell'intrinseca natura ipertestuale e interpretativa dei fammenti (Romanello et al. 2009). Il primo livello contiene il testo dei "testimoni" codifcato in TEI, un formato di codifca di testi che si è stabilito negli anni come standard de facto nell'ambito delle Digital Humanities. Il secondo livello contiene invece i metadati sui testi, tra cui anche le interpretazioni formulate dagli studiosi sui testi (individuazione e attribuzione di fammenti, varianti etc.). Per poter collegare tra loro i due livelli è necessario un sistema di linking con un livello di granularità tale da consentire i riferimenti alla parola in testi di cui possono esistere molteplici edizioni. Tale sistema è stato implementato basandosi sul protocollo CTS (Canonical Text Services) (Smith 2009).
Alcuni esempi di codifca di fammenti nei Deipnosofsti di Ateneo saranno presentati per illustrare il funzionamento del modello proposto.

Bibliografia

M. Berti, M. Romanello, A. Babeu, and G. Crane. 2009. Collecting fragmentary authors in a digital library. In *Proceedings of the 9th ACM/IEEE-CS joint conference on Digital libraries*, 259- 262. Austin, TX, USA: ACM. doi:10.1145/1555400.1555442.

M. Romanello, M. Berti, F. Boschetti, A. Babeu, and G. Crane. 2009. Rethinking Critical Editions of Fragmentary Texts By Ontologies. In *Proceedings of 13th International Conference on Electronic Publishing: Rethinking Electronic Publishing: Innovation in Communication Paradigms and Technologies*, ed. S Mornati and T Hedlund, 155-174. Milano, Italy.
http://conferences.elpub.net/index.php/elpub/elpub2009/paper/view/158/66.

N. Smith. 2009. Citation in Classical Studies. *Digital Humanities Quarterly* 3, no. 1 (Changing the Center of Gravity: Transforming Classical Studies Through Cyberinfrastructure).
http://www.digitalhumanities.org/dhq/vol/003/1/000028.html.

ALESSANDRO BAUSI
(Asien-Afrika-Institut, Universität Hamburg)

Il progetto COMst (Comparative Oriental Manuscript Studies); Etiopistica e filologia digitale

A. *Il progetto COMSt.* – Conformemente alla sua natura di "networking project", COMSt – (Comparative Oriental Manuscript Studies), ESF (European Science Foundation), Research Networking Programme – nasce dalla convinzione che i ricercatori in settori di studio vicini e talvolta contigui per l'oggetto e il contesto storico-culturale affrontato – nel caso specifico, con evidente centralità delle culture del "codex" nell'area mediterranea e limitrofe – debbano condividere, o almeno discutere, decisioni strategiche su metodi e obiettivi di studio dei manoscritti come oggetto materiale (tradizionalmente: paleografia e codicologia); metodologie e "requisiti minimi" di edizione e interpretazione dei testi (filologia "globale", ma con particolare riferimento alla critica del testo in senso stretto, o ecdotica); esperienze, risultati, prospettive, standard di codifica e tecnologie dell'applicazione digitale; criteri catalografici previa definizione delle esigenze, spesso assai diverse nella prospettiva – per semplificare – degli utenti e delle istituzioni bibliotecarie; e in un contesto più ampio siano anche informati, ed eventualmente prendano posizione, sui problemi relativi all'accesso a – e conservazione e tutela del – materiale manoscritto. Su ciascuno dei punti evocati, nel rispetto delle regole e indicazioni della ESF, COMSt si è strutturato in gruppi di lavoro autonomi, con la finalità di pervenire a una prima sintesi nell'arco dei 5 anni del progetto (2009-2014).

B. *Etiopistica e filologia digitale.* – L'etiopistica – qui intesa nella sua accezione più tradizionale di studio di una delle diverse culture letterarie (con quella siriaca, copta, armena, georgiana e arabo-cristiana) componenti lo spettro dell'Oriente cristiano – presenta spunti per la discussione di problematiche comuni ad altre discipline orientalistiche di carattere prevalentemente filologico. La presentazione di alcuni primi, recenti tentativi di "approccio digitale" – come il nome di uno dei gruppi di lavoro del progetto COMSt ("Digital Approaches to Manuscript Studies") suggerisce – allo studio dei manoscritti e dei testi etiopici, si presta a qualche riflessione sulle perduranti ambiguità di metodo, sulla definizione degli obiettivi comuni, e sullo stato generale della ricerca, in cui i "desiderata" prevalgono di gran lunga sulle acquisizioni.

MANUEL BARBERA
(Dipartimento di Scienze letterarie e filologiche, Università di Torino)

Intorno a Schema e storia del "Corpus Taurinense".

Il volume *Schema e storia del "Corpus Taurinense": linguistica dei corpora dell'italiano antico*, frutto di una ricerca decennale, è di notevole ampiezza quantitativa (circa 1.300 pagine, con 4.195 citazioni tratte da 254 testi e 510 query CQP) e funzionale, assolvendo a pratico manuale di riferimento ed accurata documentazione dell'innovativo *Corpus Taurinense*, storia di una ricerca e vademecum dell'aspirante costruttore di corpora, irrinunciabile punto di riferimento sulla linguistica dei corpora dell'italiano antico, rilevante contributo ai rapporti tra linguistica teorica, storica e computazionale, *ubi consistam* in materia della linguistica italiana, romanza e computazionale, ecc.
Il *Corpus Taurinense* (257.185 token, 18.876 type, 8.325 lemmi), a sua volta oggetto principale del volume, è costituito da ventidue testi fiorentini della seconda metà del XIII secolo, annotati e completamente disambiguati per parti del discorso, categorie morfosintattiche, genere letterario, caratteristiche filologiche ed articolazione paragrafematica del testo, portando le esperienze e le tecniche più avanzate della linguistica dei corpora dalle lingue moderne a quelle antiche. Costruito, infatti, secondo specifiche EAGLES>ISLE compatibili nel formato CWB (Corpus Work Bench, sviluppato dall'IMS Stuttgart), e rilasciato sotto licenza Creative Commons Share Alike, è liberamente consultabile con CQP (Corpus Query Processor) alla sua homepage http://www.bmanuel.org/projects/ct-HOME.html.
La presente breve presentazione intende fornire un primo orientamento in questo vasto affresco, offrendo al contempo (nei limiti del tempo disponibili) l'esemplificazione di alcuni problemi significativi affrontati.

MARCO TOMATIS
(Dipartimento di Scienze letterarie e filologiche, Università di Torino)

Aspetti computazionali e metodologici della disambiguazione del 'Corpus Taurinense'

La disambiguazione del Corpus Taurinense è stata la fase conclusiva di un progetto di ampio respiro nato con lo scopo di mettere a disposizione a filologi e storici della lingua una base di dati testuali dell'italiano del '200. Nel corso dello sviluppo del progetto ci si è accorti del ruolo importante che il corpus in questione poteva avere in veste di *training corpus*, ossia corpus di riferimento per sistemi stocastici di annotazione morfosintattica.
Trattandosi di una lingua ancora vergine dal punto di vista del trattamento automatico, la soluzione più adeguata consisteva nel completo sviluppo di una serie di regole di disambiguazione unitamente al relativo sistema di gestione, affiancando ai modelli più generali un gruppo di regole *ad-hoc* capaci di gestire l'enorme carico di eccezioni. Oltre a ciò, il corpus presentava tre diverse tipologie di ambiguità su cui intervenire: la transcategorizzazione esterna, relativa a parti del discorso diverse tra loro; quella interna, relativa a caratteristiche inerenti la singola categoria morfosintattica (es. genere, numero); infine quella intra-POS, relativa a elementi distintivi appartenenti alla stessa categoria morfosintattica (es. modo, tempo). Tutto questo ha fatto sì che le regole, organizzate secondo uno schema a mutua esclusione, venissero distribuite su sei moduli distinti attivabili in cascata. Per quanto riguarda più in particolare gli aspetti computazionali, il sistema di gestione delle regole si avvale di un motore di scansione del testo incentrato sull'utilizzo di puntatori mobili che permettono una ben precisa definizione della sezione di testo sulla quale la regola andrà ad operare.
Per concludere è importante aggiungere che gli aspetti metodologici non sono stati presi in considerazione solamente durante le varie fasi di realizzazione del programma, bensì anche durante la progettazione delle regole stesse. Infatti, al fine di evitare potenziali errori di applicazione di una data regola, è stato predisposto l'uso preventivo di un sistema di simulazione denominato PEX (Pattern EXtractor) capace di fornire allo sviluppatore tutti i dati necessari per confermare la validità o meno di un dato modello prima della sua implementazione definitiva all'interno del programma.

Bibliografia

Manuel Barbera - Elisa Corino - Cristina Onesti, *Cosa è un corpus? Per una definizione più rigorosa di corpus, token, markup*, in *Corpora e linguistica in rete*, a cura di Manuel Barbera, Elisa Corino, Cristina Onesti, Perugia, Guerra Edizioni, 2007 "L'officina della lingua. Strumenti", pp. 25-88

Manuel Barbera, *Un tagset per il Corpus Taurinense. Italiano antico e linguistica dei corpora*, in *Corpora e linguistica in rete*, a cura di Manuel Barbera, Elisa Corino, Cristina Onesti, Perugia, Guerra Edizioni, 2007 "L'officina della lingua. Strumenti", pp. 135-168.

Manuel Barbera - Carla Marello, *Linguistica dei corpora per l'italiano antico. Annotazione morfosintattica di testi fiorentini del Duecento*, a cura di Manuel Barbera e Carla Marello, Alessandria, Edizioni dell'Orso, 2001 "Gli argomenti umani" 6.

Michael Brennan, *GAWK: Effective AWK Programming: A User's Guide for GNU AWK*, 2nd edition, Free Software Foundation Inc., 2000

Marco Tomatis, *La disambiguazione del Corpus Taurinense. Problemi teorici e pratici*, in *Corpora e linguistica in rete*, a cura di Manuel Barbera, Elisa Corino, Cristina Onesti, Perugia, Guerra Edizioni, 2007 "L'officina della lingua. Strumenti", pp. 169-181.

Marco Tomatis, *La disambiguazione. Trattamento finale*, in Manuel Barbera, *Schema e storia del "Corpus Taurinense". Linguistica dei corpora dell'italiano antico*, Alessandria, Edizioni dell'Orso, 2009, pp. 171-191.

ODD EINAR HAUGEN
(Institutt for lingvistiske, litterære og estetiske studier, Universitetet i Bergen)

Do we need all these characters? On the transcribing and encoding of medieval vernacular manuscripts

In the transcription of a text it is essential that all characters in the source are correctly rendered. Until fairly recently, characters outside Basic Latin created problems, and it was not unusual to see transcriptions using the digit '3' for yogh, or w for the wynn character. However, the increasing support of the Unicode Standard and the development of large Unicode compatible fonts have been a breakthrough in this area. Now, many transcribers of medieval texts find all characters they need in this standard. Through coordinated efforts such as the Medieval Unicode Font Initiative, a large number of additional characters are available in the so-called Private Use Area. These characters can be encoded and displayed in several fonts, in word processing applications, in PDF files as well as by almost all browsers on the Internet.

Unicode has a very strict interpretation of a character, defining characters as "the smallest components of written language that have semantic value" (ch. 2.2 of the Standard). Most characters have a number of variants, in written as well as in printed form, but Unicode is in general not interested in this variation. For example, there is an obvious graphic difference between the straight ‹r› and the round one, ‹ɾ›, but they are commonly regarded as variants of the same character, ‹r›, representing the singlephoneme /s/. On the other hand, there is only a minute difference in most fonts between the standard ‹f› and the tall ‹s› – in the first, the horizontal stroke crosses the stem, in the second it does not. Even so, they are universally regarded as separate characters, in most languages representing the phonemes /f/ and /s/.

There seem to be two major criteria for deciding on the character status of a specific letter form:

1. The graphonomic criterion: Characters are letter forms with a distinguishable form and distribution. Thus, Latin ‹f› and ‹s› are separate characters, as well as straight ‹r› and round ‹ɾ›.

2. The phonological criterion: Characters are classes of letter forms with similar phonological properties, meaning that replacing one character with another will lead to a change in meaning (the minimal pair test). From this point of view, there is a distinction between ‹f› and ‹s›, but not between the round ‹s› and the tall ‹s›, nor between the straight ‹r› and the round ‹ɾ›.

In my presentation, I shall examine these two criteria and discuss where the line should be drawn between characters and character variants, both with respect to a linguistic analysis and concerning editorial practice.

MATTHEW JAMES DRISCOLL
(Den Arnamagnæanske Samling, Københavns Universitet)

Mapping the manuscript matrix

The Text Encoding Initiative's *Guidelines for Electronic Text Encoding and Interchange* ('TEI P5') provide extensive facilities for the encoding of data pertaining to persons, whether actors in history or those living today, using the `<person>` element. Such data include:

- physical characteristics such as sex and eye colour;
- cultural characteristics such as socio-economic status, ethnicity and religion;
- information on occupation and education, and the events in people's lives such as birth, marriage or appointment to office;
- the various names by which a person may be known.

Information about places can be encoded in a similar way, using the `<place>` element, including:

- the physical location of the place, for example as a street address or a set of geographical co-ordinates;
- information on population, climate and terrain;
- descriptions of events associated with a place;
- the various names by which a place may be known, either in different languages or over time.

The `<relation>` element can then be used to describe any kind of relationship between a specified group of people, between people and places or between people, places and objects, such as manuscripts.

Using these mechanisms one can create an extensive system of authority files, associated for example with descriptions and/or transcriptions of primary source materials such as a collection of manuscripts. This not only helps to prevent repetition of information and minimise the possibility of error, it also allows one to map the relationships between the artefacts and the people who produced, disseminated and consumed them, to show in a dynamic way how the 'manuscript matrix' worked. It is this latter aspect which will be the focus of my paper.

Bibliography

TEI P5: Guidelines for Electronic Text Encoding and Interchange, ed. L. Burnard & S. Bauman (Oxford — Providence — Charlottesville — Nancy, 2009), esp. Cap. 10, 'Manuscript Description', and Cap. 13, 'Names, Dates, People, and Places'.

Driscoll, M. J.: 'P5-MS: A general purpose tagset for manuscript description', *Digital Medievalist* 2.1 (2006), http://www.digitalmedievalist.org/article.cfm?RecID=12.

Driscoll, M. J.: 'XML markup of biographical and prosopographical data', *Proceedings of the TEI Day 2006 in Kyoto: Toward an overall inheritance and development of Kanji culture*, ed. Christian Wittern (Kyoto, 2006), pp. 75-83.

MARINA BUZZONI
(Dipartimento di Scienze del Linguaggio – Università Ca' Foscari Venezia)

The 'Electronic Hêliand Project': theoretical and practical updates

The 'Electronic *Hêliand* Project' was started in June 2006 at the University of Venice (see Buzzoni 2009a; Buzzoni 2009b). Its main aim was to show how the electronic medium is capable of capturing the often disregarded differences amongst the witnesses of the ninth-century Old Saxon poem, i.e. its inner *mouvance.* As against a printed edition which offers us a static text, an electronic edition presents the text in a variety of forms and permit users to choose between visualizing only one tag scenario or several (cf. Ciula and Stella 2006; Burnard, O'Brien O'Keeffe and Unsworth 2006).
This paper will focus on the theoretical and practical updates of the aforementioned Project. Attention will be paid to the linguistic and cultural features that a close scrutiny of the witnesses has brought to the surface. Furthermore, the strategies used in editing the text and annotating the *lectio variorum* will be thoroughly analyzed. Finally, the process and progress of manuscript digitization will be critically considered.

Works cited

Burnard, Lou, Katherine O'Brien O'Keeffe, and John Unsworth, eds. 2006. *Electronic textual editing*. New York: Modern Language Association of America.
Buzzoni, Marina 2009a. "*Uuarth thuo the hêlago gêst that barn an ira bôsma*: towards a scholarly electronic edition of the *Hêliand*". In: Saibene, Maria Grazia and Marina Buzzoni, eds. 2009. *Medieval Texts – Contemporary Media. The art and science of editing in the digital age*. Pavia: Ibis, pp. 35-55.
Buzzoni, Marina 2009b. *Edizioni elettroniche e valorizzazione della storicità del testo…* In: Ferrari, Fulvio and Massimiliano Bampi, eds. 2009. *Storicità del testo, storicità dell'edizione*. Trento: Università degli Studi di Trento.
Ciula, Arianna and Francesco Stella, eds. 2006. *Digital philology and medieval texts*. Pisa: Pacini editore.

STEFANO MINOZZI
(Dipartimento di Linguistica, Letteratura e Scienze della Comunicazione, Università degli Studi di Verona)

Latin WordNet: una rete semantica per il latino

Questo progetto di costruzione di un *database* di conoscenza semantica per la lingua latina nasce con l'obiettivo di poter fornire lo *specimen* di uno strumento in grado di permettere l'implementazione, su testi in latino, di nuove tecniche d'analisi derivate dagli studi di *Natural Language Processing*.
Una rete semantica è uno strumento che vede la riscrittura del dizionario tradizionale in una struttura dove le parole sono collocate mediante una gerarchia di concetti e relazioni, fornendo una base di conoscenza che ampia le potenzialità dell'analisi computazionale. Un testo, con il supporto di una rete semantica, può essere marcato in modo tale da poter essere processabile non come mera sequenza simbolica, ma come insieme di concetti, ricostruendo un miglior modello di testualità per l'analisi assistita dal calcolatore.
La rete semantica *Latin WordNet* è realizzata in conformità con il modello di *Multi WordNet* (Pianta, 2002), con il quale è pienamente compatibile e integrabile.
Per la costituzione di *Latin WordNet* sono stati sviluppati vari metodi automatici che hanno permesso di riorganizzare la conoscenza semantica contenuta in dizionari digitalizzati. In particolare è stato messo a punto un sistema di assegnazione che, servendosi di fonti multilingue, ha contribuito a una più rapida creazione dei nodi (Minozzi, 2008).
Il database contiene attualmente 9.378 parole assegnate a 8973 *synset* (gruppi sinonimici) che individuano 143.701 archi di relazioni. Nella tabella si mostra la consistenza relative a ciascuna parte del discorso.

	Sostantivi	**Verbi**	**Aggettivi**	**Avverbi**
SYNSET(gruppi sinonimici)	5621	2283	775	294
LEMMI	4777	2609	1259	479
CONCETTI	13060	10062	2054	732

L'implementazione di *Latin WordNet* apre all'occasione di sperimentare nuove opportunità di applicazione del computer allo studio dei testi:

- *Information Retrieval*: le relazioni di sinonimia sono usate per l'espansione delle *query* per migliorare i risultati delle ricerche; inoltre si profila la possibilità di applicare ricerche multilingua grazie alla struttura della rete (*Cross Language Information Retrieval*);
- *Semantic tagging:* i testi possono essere marcati attraverso concetti identificatori a partire dalle parole che li costituiscono, rendendo possibile la loro catalogazione automatica;
- *Processi di disambiguazione automatica:* le relazioni semantiche descritte attraverso la rete permettono di oggettivare la *distanza semantica*, offrendo dati quantitativi per l'applicazione di algoritmi di disambiguazione;
- *Costruzione di ontologie*: la collocazione gerarchica delle parole in rete costituisce la base per la realizzazione di tassonomie specifiche e glossari, ulteriormente utilizzabili per operazioni di *Natural Language Processing*.

FRANCO D'AGOSTINO / MATTEO SCALZO
(Dipartimento di Studi Orientali, Università La Sapienza)

Toward a Knowledge Based Approach to the Sumerian Culture

In recent years, we investigated a knowledge based approach to the the study of the Sumerian culture: the results have been the creation of an ontology of a Sumerian grammar, and the design and implementation of a knowledge based catalogue system.

The ontology of a Sumerian Grammar is an original attempt to produce a formal description of a non formal language and to directly test such an approach for the ancient Sumerian language itself, aiming at detecting (on very large corpora, thanks to computer aided annotation) some regularities and recurring patterns that may point out to scholars clues to discover "rules" of Sumerian grammar.
The ontology identifies the two main elements in the structure of the sentence in Sumerian: the Verbal Chain and the Nominal Chain, from which are also identified all the grammatical components. In particular, the order of appearance of the elements in the chain is meaningful, since there are some specific positions according to the syntactic role the components have to play.

As mentioned above, we also designed a innovative cataloguing system for the Dhi Qar -University and Heritage Project.
The system provides the opportunity to systematically integrate and combine various categories that are approached by different disciplines, in different ways, even if the object of their observation is the same and even if the way of investigating it is more or less specific.
This versatility is possible thanks to the use of a knowledge base, a powerful expressive system offering a more precise characterization of data relations.
This system is able to help us in the representation of scientific knowledge about archaeology and epigraphy. Both categories specification and observational criteria are defined by scientists of all contemplated disciplines. These specifications are codified in a specific form suitable for the computer usage (a formalism belonging to the Description Logics family).
The system is based on a specific software (the so called reasoner) that is able to follow the connections among data trough the relations across the information structures.

Bibliography

S. Alivernini. Progetto ME: l'ontologia di una grammatica sumerica. DOI 2006: 10.1683/ab0002 ; available on line at http://dx.doi.org/10.1683/ab0002 (in Italian).

Kiengi - Dhi Qar Project http://www.kiengi.org/dhiqar/

The Description Logic Handbook: Theory, Implementation and Applications. Cambridge University Press, 2002. Edited by F. Baader, D. Calvanese, D. McGuinness, D. Nardi, Peter Patel-Schneider.

ENRICA SALVATORI
(Dipartimento di Storia, Università di Pisa)

Umanista esperto di informatica o informatico umanista? Ragionamenti su discipline, ricerche e professioni a cinque anni dalla nascita di Informatica Umanistica all'Università di Pisa.

Nonostante l'ormai unanime riconoscimento del peso che il digitale ha acquisito nell'ambito delle così dette "scienze umane", la didattica universitaria si sta adeguando con lentezza e in maniera non coordinata alle nuove esigenze che il cambiamento richiede. Nel processo di trasformazione possiamo individuare due correnti principali: la prima tesa a mantenere una solida formazione tradizionale a cui viene aggiunta qualche nozione superficiale di informatica; la seconda che punta a fornire una preparazione ibrida, bilanciata tra informatica e scienze umane. I mondi del lavoro e della ricerca, come stanno rispondendo a queste due diverse figure di professionisti del sapere? L'intervento è mirato a evidenziare pregi, difetti e problematiche insite nelle due scelte, anche alla luce di quanto fatto in questi anni nel Corso di laurea di Informatica Umanistica dell'Università di Pisa.

Bibliografia

Fabio Ciotti e Gino, Roncaglia, *Il mondo digitale. Introduzione ai nuovi media,* Roma, 2008

McNamara, Billie R., *The Skill Gap: Will the Future Workplace Become an Abyss*, in "Techniques: Connecting Education and Careers", 84 n. 5 (2009), pp. 24-27

Gino Roncaglia, *Informatica umanistica: le ragioni di una disciplina*, in "Intersezioni**"** , XXIII n. 3 (dicembre 2002), pp. 353-376

Nicola Rossignoli, *Appunti di cultura digitale : informazione, comunicazione, tecnologie*, Milano 2008

E. Salvatori, *Didattica della storia e nuove tecnologie: opportunità, problemi e scenari plausibili nelle Università italiane*, in "Reti medievali"- Didattica, 2008 - <http://www.storia.unive.it/_RM/didattica/corsi/salvatori2.html>

Bani, M., Ciregia E., Genovesi F., Rapisarda B., Salvatori E., Simi M., *Learning by creating historical buildings in Second Life,* in *Virtual Learning and Teaching in Second Life*, a cura di J. Molka-Danielsen e M. Deutschmann, Trondheim, Tapir Akademisk Forlag, 2009

E. Salvatori, *Hardcore history: ovvero la storia in podcast*, in "e Ricerca", XVII, Nuova Serie, n. 30, gennaio-aprile 2009, pp. 171-187

E. Salvatori, *Interventi* su: parte 1a) *Una questione di definizioni. I rapporti tra discipline umanistiche e informatica*; parte 2a) *Quantità e qualità. I testi, le biblioteche e l'accesso alle informazioni*; parte 4a) *Cultura, didattica e ricerca*, in "Informatica Umanistica", Volume 1, Anno 2009, http://www.ledonline.it/informatica-umanistica/

Francesco Varanini e Walter Ginevri (a cura di), *Il project management emergente*, Guerini e Associati, 2009

Hanna, Donald E., Higher Education in an Era of Digital Competition: Choices and Challenges, Atwood Publishing, 2000

ROBERTO ROSSELLI DEL TURCO
(Dipartimento di Scienze del Linguaggio, Università di Torino)

Filologia digitale: ragioni, problemi, prospettive di una disciplina

Nel mondo anglosassone e, con un certo ritardo, anche in quello italiano cominciano a diffondersi edizioni digitali di testi antichi e moderni. La loro diffusione appare alquanto limitata, tuttavia, e la loro produzione sembra essere esclusivo appannaggio di quelli che sono stati battezzati "tecno-filologi", ovvero una ristretta cerchia di filologi esperti nell'uso degli strumenti informatici necessari per la creazione di edizioni digitali. Quali sono i benefici che la filologia digitale può apportare al lavoro del filologo? si tratta unicamente di benefici materiali, legati alla forma elettronica di produzione e distribuzione dell'edizione, o vi sono anche progressi sul piano teorico e metodologico? perché un filologo "tradizionale" dovrebbe avvicinarsi agli strumenti della filologia digitale? L'intervento si propone di fare il punto sulla situazione attuale della disciplina e di sottoporre all'attenzione del pubblico alcune riflessioni in merito a queste domande.

Bibliografia

Faulhaber, Charles B. (1991). "Textual Criticism in the 21st Century", *Romance Philology* , pp. 123-48.
Fiormonte, Domenico (2003). *Scrittura e filologia nell'era digitale*, Torino, Bollati Boringhieri.
Fiormonte, Domenico (2007). "Parole online. Quale editoria e filologia nell'era di digitale?", *Nuova Informazione Bibliografica*, vol. 2, pp. 355-362.
McGann, Jerome (2001). *Radiant textuality. Literature after the World Wide Web*, New York and Basingstoke, Palgrave.
Mordenti, Raul (2001). *Informatica e critica dei testi*, Roma, Bulzoni.
Robinson, Peter (2005). "Current Issues in Making Digital Editions of Medieval Texts—or, Do Electronic Scholarly Editions have a Future?" Digital Medievalist1.1 [Online journal]. http://www.digitalmedievalist.org/article.cfm?RecID=6.
C. M. Sperberg-McQueen - L. Burnard, eds. (2009). *TEI P5: Guidelines for Electronic Text Encoding and Interchange [v. 1.5.0]*. Oxford, Providence, Charlottesville, Nancy: Text Encoding Initiative Consortium. La versione più aggiornata è sempre disponibile sul sito http://www.tei-c.org/.

Paola Cotticelli / Alfredo Rizza / Alfredo Trovato
(Laboratorio del Lessico di Linguistica - Dipartimento di Linguistica, Letteratura e Scienze della Comunicazione, Università degli Studi di Verona)

Lessico di Linguistica On line: A Linguistics Lexicon Archive

The aim of this presentation is to outline the metalinguistic project *Linguistics Lexicon Archive*, started in July 2009, within the *Laboratorio del Lessico di Linguistica* under the direction of Prof. Dr. Paola Cotticelli. It consists of a comprehensive database of metalinguistic records, containing a briefly definition of the linguistic items as well as a reference list of the main related titles.
In the first step, the project started out to collect a subject-indexed bibliography (15.000 entries), providing a list of published literature on different topics. For this purpose, the project team made use of a reference management software package, which allows us to store bibliographic entries in a standard format as well as to retrieve them quickly by searching under keywords.
In the second step, the data stored will be converted into a relational database, designed with SQL, which properly will represent the metalinguistic archive hosted in a website. The *Linguistics Lexicon* will offer an open access to the linguistic items recorded, providing scholars and students with a helpful and innovative research tool in the field of linguistic studies.
In this presentation, we will illustrate the preliminary results of the project (still in progress), sketching the next phases of the work plan.

References

Atkins, B.T.S. and Zampolli, A. (eds.), (1994). *Computational approaches to the lexicon*. Oxford.
Atzeni, P. et alii (1996). *Basi di dati, concetti, linguaggi ed arichitetture*. Milano.
Ciotti, F. (1995). "Testi elettronici e biblioteche virtuali: problemi teorici e tecnologie, Schede Umanistiche II n.s., n. 2, pp. 147-178.
Cotticelli, P. (2007). *Lessico di linguistica* (Traduzione italiana, adattamento e revisione sulla base della 3° ediz. originale rivista ed ampliata). Alessandria.
Lorenzi, F. (1993). *Sul linguaggio e informatica*. Alessandria.
Lorenzi, F. (ed.), (2002). *DLM. Dizionario generale plurilingue del Lessico Metalinguistico*. Roma.
Riordan, R.M. (1999). *Progettare database relazionali*. Milano.
Spina, S. (1997). *Parole in rete. Guida ai siti Internet sul linguaggio*. Firenze.
Vallini, C. (ed.), (2000). *Le parole per le parole. I logonimi nelle lingue e nel metalinguaggio* (atti del convegno, Napoli, Istituto Universitario Orientale, 18-20 dicembre 1997). Roma.

ADELE CIPOLLA / FEDERICA GORIA
(Dipartimento di Anglistica, Germanistica e Slavistica Università degli Studi di Verona – EdiText Torino)

Open BMS: a New Software for a Snorri's Edda Annotated Bibliography

Users' interaction with the huge bibliographical database about the *Snorra Edda* is made difficult by a typical *mouvance*, since actually it seems to move from the work to the literature devoted to it. The history of the editorial practice around Snorri's work reverberates through the secondary literature. Since the *Edda*, because of its complex composition (four thematically distinguished parts that in the course of time, sometimes cut off from each other, had different fates) and of its tradition (four main manuscripts that, independently, summarized the 'basic text', or interpolate other materials in it), was edited, translated and acknowledged in various forms of, often unnoticed, rewording and reworking (the first striking instance being the *editio princeps* of 1665, which actually edited none of the extant manuscripts but a drastic 17th century reworking, that updates the medieval encyclopedic criteria of the work to the modern alphabetical ones).
So aiming at creating an on-line version of the *Snorri's Edda Bibliography 1665- onwards*, we felt the need for an application which, by means of keywords, can, apart from the usual thematic contents of the annotated bibliographical texts, give back the required paratextual information. This could allow for the understanding of which version of Snorri's *Edda* each title refers to, and so offering to the user the possibility to delineate a history of Snorrean studies (e.g. in pointing out the countries, the languages, the genres of publications). Nevertheless, because of the continuous increase of this bibliographical *corpus*, we felt the need for an easy, constantly improvable tool, which could make the most of the interaction between administrator and users.
To achieve this, the project plans the development of specific software, OpenBMS (Open Source Bibliography Management System). This application allows for literature searches using the most modern means of interaction with the user, such as virtual keyboards for easy insertion of special characters and advice for completing the research in Ajax. The application consists of a back-end to populate the database and a search interface.

The bibliographical entries are uploaded to the database through the use of a graphical interface, so as to enable users without technical skills in this field to populate the database.
The search interface allows for different types of research through the use of a form. The use of JavaScript libraries of widgets and asynchronous communication makes possible the achievement of two fundamental objectives: to facilitate access to UTF-8 characters that are not mapped on normal keyboards, and communication with the server to suggest search parameters against approximate searches or with data similar to those found in the data base.

This solution allows, on the one hand, making real-time changes to the bibliography, on the other, it facilitates navigation within the vast literature. Access to the portal is possible for two different types of users: unauthenticated users can perform searches in the bibliography, using the parameters "title", "author", or "keywords". However, administrators can modify the database via the web interface, so as to make updates available in real time. The web application is developed in Java/Javascript and the source code is released under the GPL license to encourage the development of additional functionality and reusability.