A Preliminary Analysis of a Nuzi Scribal Family in view of an application in the ENEA-GRID

Paola Negri Scafa - Daniela Alderuccio - Giovanni Bracco - Silvio Migliori
ENEA - Rome (Italy)

Assyriology uses new technologies for computer encoding of cuneiform texts, to support the analysis and management of a great deal of data coming from tablets. At the present, integrating e-tools in the GRID virtual environment opens new perspectives in the analysis of ancient texts and enables ENEA-GRID for e-Humanities, with a particular attention devoted to Assyriology. Therefore, after a preliminary study, an experimental application of new technologies to the Nuzi Corpus has been carried out, using Multilingual Text Mining in ancient languages, in view of an integration into the digital environment of the ENEA GRID. This innovative approach into the study of cuneiform corpora may influence software developments and collaborative working in order to match the specific needs of scholars communities.

Introduction

The Project's goal is to demonstrate that:
- accessing the collaborative environment of GRID computing
- sharing its computing resources distributed on many different computers around the world,
- integrating and sharing text mining software, lexical resources and data representation tools in the GRID environment

can open new ways of approaching the study of cuneiform *corpora* and enable ENEA GRID[1] for e-Humanities. (Fig. 1)

WHAT IS THE GRID?

The **ENEA GRID Digital Environment**

The GRID Technology

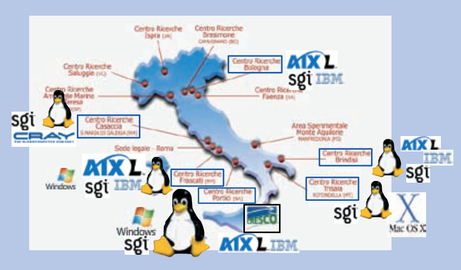
The availability of powerful computing systems, distributed over a wide area and connected by high speed networks, has lead to the development of the concept of computational GRID. The GRID provides a unified approach to heterogeneous computational resources located in distant sites so that thousands of computers can act in concert on research or engineering problems, accessing large amounts of data and performing heavy duty computations.

Characteristics of ENEA GRID

ENEA - The Italian National Agency for new Technologies, Energy and sustainable economic Development - has substantial experience in GRID computing technologies

ENEA GRID (www.eneagrid.enea.it) mission started in 1999 and now it provides a unified user environment and a homogeneous access method for all ENEA researchers and their collaborators, irrespective of their location, optimizing the utilization of the available computational resources.

ENEA GRID infrastructure covers 12 Research sites and is designed and managed by the ENEA ICT Unit (www.utict.enea.it) with 6 computer centers, providing multi-platform resources for serial & parallel computation and graphical post processing. The main computational resources consist of about 4600 cores for a total of more than 40 Tflops (10^15 floating point operations per seconds), equivalent to the computing power of several thousands of standard PC.



ENEA GRID offers a fast and secure access to software and hardware platforms implementing the ENEA e-Science[2] approach for high performance computing[3]:

- sharing computing resources across collaborative projects;
- attracting, engaging and supporting a wide range of users and researchers from science and industry, providing them with a production service supported by extensive technical and training support.

ENEA GRID is now integrated with other GRIDs: IGI, EGI and PON Projects/GRISU as a result of interoperability activities performed on previous GRID projects, as DATAGRID, EGEE, BeInGRID.

Fig. 1

GRID and Assyriology

GRID can offer an important technical support in Assyriology mainly thanks to the great power of its computing and storage infrastructure. Because of the easy access to its resources, scholars can widely get or add information.
Nevertheless some problems exist: Text Mining and Lexical Analysis Programs are oriented to modern languages and require adaptation in order to deal with problems related to cuneiform texts, like, for example:

- Graphemic ambiguities and inconsistencies, including a special use of caps and lower case letters;
- Use of more than a language in the same text;
- Questions connected with fonts and the use of particular characters;
- Other peculiarities to be taken into consideration, like lack of punctuation and/or space and carriage return, that are the only separators of graphic forms.

Methodological Steps

Therefore preparatory phases and methodological steps are necessary, (A) partly connected with the Text Mining Methodology and (B) partly related to the graphemic and linguistic characteristics of the documents.

A) Methodological steps connected with the Text Mining Methodology

1st step: Preparatory phase
- Preliminary definition of a general Text Mining methodology applied to texts

2nd step: Selection of software for
- Multilingual Text Mining, in order to extract hidden knowledge from texts
- Network Analysis and Visualization of hidden text relations

B) Methodological steps related to the graphemic and linguistic characteristics

1st step: Preparatory phase
- Graphic notation of the transliterated texts due to the (present) lack of linguistic quantitative software able to process adequately fonts employed for the texts
- Identification of writing and stylistic elements that make possible to know the authorship of the texts.

2nd step: Corpus selection and tagging
- Choice of a test-bed corpus of transliterated e-texts to be analyzed
- Pre-processing rules for text segmentation, based on text position in the tablet (recto, verso, left/right margin, etc), and on text sections (formulas, witness lists, seals lists, etc).
- Grammatical and semantic tagging of texts (Fig. 2)

Operational Steps

3rd step: Extraction of preliminary results:
First results give an opportunity to enhance methodology and to improve text mining strategies following an iterative process.

Future Steps

4th step: Finding and Applying the final Text Mining Methodology

5th step: Extraction of Final results

The Present Analysis - A Case-Study

Text Corpus
Among the whole *corpus* of Nuzi texts, a small group of documents, belonging to the scribal family of Šeršiia, his son Ḫupita and his grandson Muš-teššup, has been analyzed as a case-study, in order to put into light innovative and conservative stylistic elements in the redaction of texts through three generations. The reduced dimension of the corpus allows an exhaustive control of the results in the application of the software.

Šeršiia worked as a scribe in particular in the town of Temtenaš, mainly for the archives of Tulpun-naia, of Kurpa-zaḫ and Teḫip-tilla; his son Ḫupita also worked as a *mār šipri* for the Palace; the horizon of Muš-teššup seems to have been mainly Nuzi.
The texts of this small group of documents are mainly contracts, declarations in court or trials, and have been analyzed according to their authors.

For the purposes of this presentation a complex of programs (Fig. 3) has been employed on the best preserved texts of the three scribes [11 texts of Šeršiia, 2 texts of Ḫupita and 5 texts of Muš-teššup], for the sake of a correct evaluation of the number of occurrences of any word within the texts, but the problem of words occurrences in fragmentary texts will be taken into consideration in a very next future.
The documents have been tagged in order to put in to light information of different kinds: graphic and graphemic, grammatical, prosopographical data, and so on. (Fig. 4 and Fig. 5)
Preliminary results on personal names compounded with *-(š)enni* and logograms are shown in Figg. 6-7.
Moreover prosopographical data have been preliminarly investiged; some results of a tentative match of five texts (AASOR XVI 23, 26, 31; JEN 124, 331) are shown in Fig. 8.

Substitutive Characters in Encoding Texts

Awaiting for new programs that enable the use of UNICODE fonts, the following characters have been employed in encoding texts:

š	→	$
ṭ	→	v
ṣ	→	c
ḫ	→	h

Moreover:

=_= indicates an erasure without text

=abc= indicates a text over erasure

/abc/ indicates a text omitted by the scribe and added by the scholar

//abc// indicates a text mistakenly added by the scribe

Akkadian is written in lower case and Sumerian logograms in caps.

Fig. 2

e-Tools

Text Mining Software: TALTAC2 - "Trattamento Automatico Lessicale e Testuale per l'analisi del contenuto di un corpus "[4] - Dipartimento di Metodi e Modelli per l'Economia, il Territorio e la Finanza (MEMOTEF) - Università "La Sapienza" di Roma -
Data Representation Tools: Software PAJEK[5] for Network Analysis and Visualization. (University of Ljubljana - Slovenia). Pajek runs on Windows, Unix or Mac and is free for non-commercial use.
Lexical e-Resources for the analysis of Personal Name lists with lemma and alternative graphics; Verb Lists with verb conjugation, etc.

Fig. 3 - e- Tools

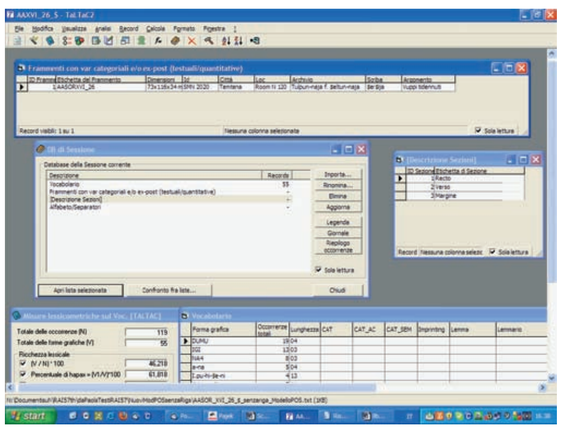


Fig. 4 – Table: TALTAC[2] snapshot - AASOR_XVI_26_$

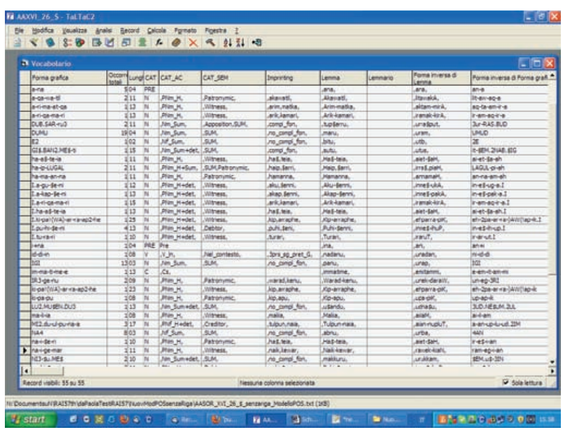


Fig. 5 - Table: TALTAC[2] Vocabulary snapshot – AASOR_XVI_26_$

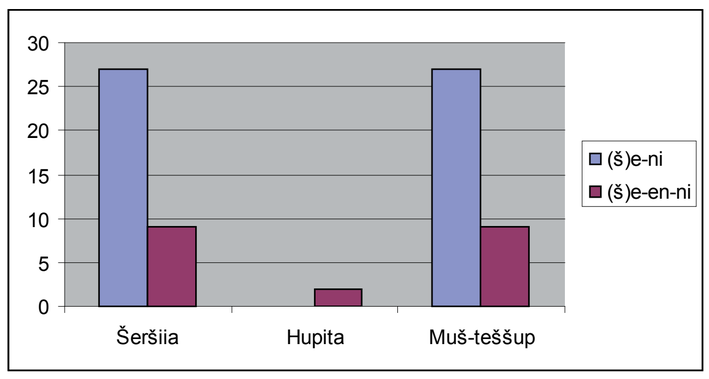


Fig. 6

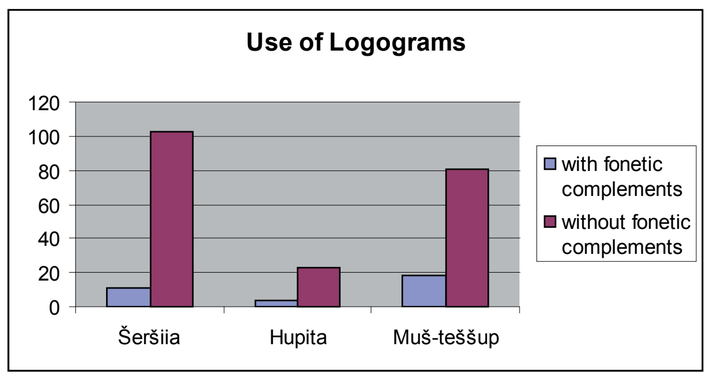


Fig. 7

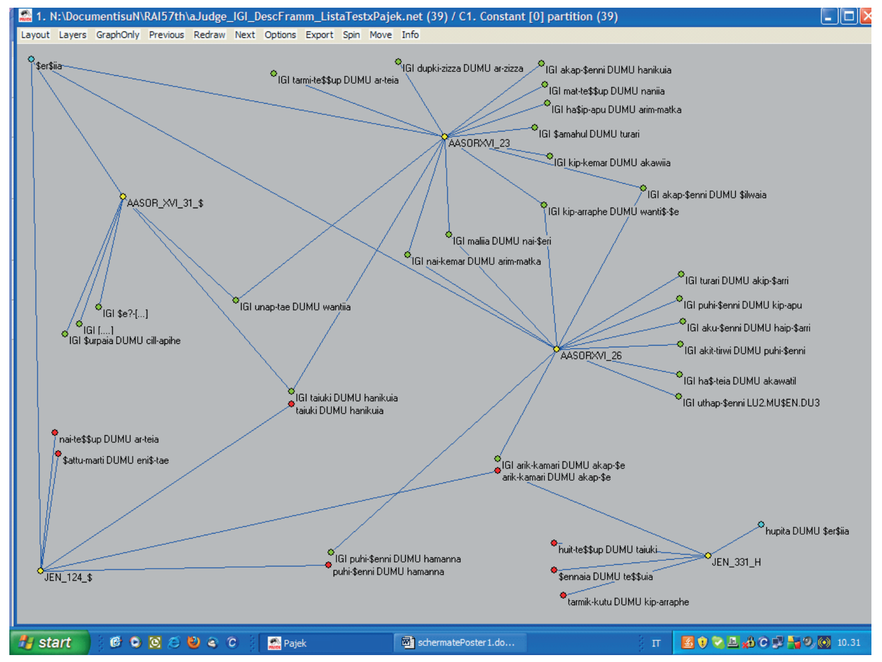


Fig. 8 - PAJEK snapshot: Comparison between Judges (red) and Witnesses (green)

Results

As it is known, the writing system in the Kingdom of Arrapḫa knew marked, subtle changes in usage, so that, it developped in the distinctive Nuzi writing system from a formative stage. Is it possible to observe traces of this process in the Šeršiia family? It is necessary to analyze many elements to trace the modus operandi of a scribe: uses in orthography (for example, logographic and syllabic writing), uses of verbal tenses (for example preterit or perfect), uses of clauses, and so on. Because the texts of the three scribes are different in nature and number, attestations are hardly comparable. Nevertheless, also if it is of course not possible to draw general conclusions in a so preliminary phase of the project, some indications can be reached. As seen in Fig. 6 the variants in *(š)e-ni* tend to disappear; Fig. 7 shows that the three scribes use fonetic complements in the same way; moreover no traces of the occasional assyriasms of Šeršiia are found in the texts of his son and grandson. As far as other elements, it seems that the use of clauses is more consistent in Šeršiia and in Muš-teššup than in Ḫupita, but investigations are being carried out.

Conclusions

GRID and e-Humanities
The GRID enables scholars to perform quantitative and comparative studies on text corpora, by using linguistic software for computer-aided analysis: a) The GRID technology offers an innovative insight in the study of ancient texts. b) The GRID infrastructure provides access to computational resources for the storage and processing of large textual corpora (from the INTERNET, from Digital Libraries, Data Bases, Archives, etc.).

GRID in Assyriology
In this virtual environment, the integration of text mining software, lexical resources and data representation tools opens new ways of approaching the study of cuneiform corpora. In the domain of Assyriology ENEA GRID is a digital environment for the storage and processing of rare and fragile sources materials. Furthermore, it offers a collaborative digital environment, to share knowledge and digital resources: electronic language resources (grammar, multilingual dictionaries, lexica, thesauri, ontologies, ecc.), digital textual archives, database, software, fonts, research drafts, etc. In this project Text Mining and Network Analysis have been successfully used to acquire new knowledge about the Nuzi scribal systems. Preliminary results are promising and encourage the pursuit in this research direction. The integration of Language Technologies and other e-tools in the ENEA GRID makes it possible to extract hidden knowledge from large textual archives, performing semantic research, which can be used to map relations among texts. Enabling ENEA GRID for e-Humanities opens new perspectives in the analysis of ancient texts.

Short- and Long-term Prospects
This innovative approach into the study of cuneiform *corpora* may influence software development, specially in order to match the specific needs of scholars communities. Adaptations and improvements of programs will be carried out in view of an integration into the GRID environment. Also some practical elements, like the use of an Assyriological-friendly font, will be taken into consideration and attempts with the use of UNICODE fonts are being carried out. Therefore suggestions, proposals and forms of discussion are in anyway welcomed.

1 http://www.eneagrid.enea.it/
2 e-Science: science using immense data sets requiring grid computing: social simulations, particle physics, earth sciences and bio-informatics, etc.
Sometimes e-Science includes technologies such as the Access Grid (an ensemble of resources, including multimedia large-format displays, presentation and interactive environments,) enabling distributed collaboration between groups of people in different locations.
3 Protein folding, financial modelling, earthquake simulation, and climate/weather modelling.
4 http://www.taltac.it/it/index.shtml
5 http://pajek.imfm.si/doku.php?id=pajek