58th Rencontre Assyriologique Internationale - Private and State

Data Mining Tools and GRID Infrastructure for Assyriology Text Analysis (an Old-Babylonian Situation Studied Through Text Analysis and Data Mining tools)

Giovanni Ponti, Daniela Alderuccio, Giorgio Mencuccini,
Alessio Rocchi, Silvio Migliori, Giovanni Bracco, Paola Negri Scafa

Abstract

The fifty letters found at Tell Harmal can provide a significant evidence on the relationships among citizens, local authorities and the Palace in the kingdom of Ešnunna. They also offer important elements for a study of the Ešnunna old Babylonian grammar. For these reasons these texts have been selected as a first step in the e-Shnunna Project, that is part of the most complex TIGRIS Project. In this project, beside the assyriological methodologies, ENEA-GRID facilities are employed: they include an innovative approach to the study of cuneiform corpora by a software application for Text Analysis (TaLTaC2) and Data Mining tools (ASTEC).

In this paper, some preliminary results, obtained mainly thanks the experimental analysis, are presented. A detailed description of procedures and systems employed in studying documents is also given and the general project is synthetically presented.

Introduction

Tell Harmal, in the suburbs of Baghdad City, where in the forties excavations were carried out by the Iraq Antiquities Department, corresponds to the ancient walled town of Šaduppûm. There, several buildings have been found, that include templar, palatine institutions, artisan workshops and residential quarters. Also if it is possible that the site was occupied in XXIV century B.C., only five levels have been excavated. The most recent levels (II-III) covers a period of about one hundred years (about XIX-XVIII century B.C.), when the town was a relevant centre of the Kingdom of Ešnunna; during this time, the architectural structure of the town remained practically unchanged. Thousands of tablets have been found there and offer an important evidence of the organization, social structure and

economy of the town. In particular, the administrative structure of the town is of great interest: it is known that the major authority of the town was the šakkanakkum, who represented the king of Ešnunna in the town, where he governed with the help of his co-operators. On the other side, a relevant role was also played by the Elders of the town.

In order to investigate thoroughly this organizational structure, the procedures and the hierarchized administrative structure of the town, the 50 letters found at Tell Harmal and published by Goetze in 1958 have been resumed in this paper. These 50, or better 51[1], letters are important not only from a sociological, but also from both a historical and a linguistic point of view: they belong to the same period (king Ibāl-pî-El II, 1779-1765) and this is useful in order to examine more closely procedures and more or less official ways of doing; moreover, they represent a consistent linguistic evidence, because they belong to the same period, and therefore can offer a useful comparison with other letters of that Kingdom[2]. For the variety of terms and forms and for the consistency of the language employed, these texts can become a first step for a study of the Ešnunna old Babylonian grammar. It is well known that the scribal school of Šaduppûm was excellent, and therefore it is important to investigate on it.

In these investigations, beside the most usual assyriological methodologies, we intend to make use of the opportunities that advanced informatic tools can offer, either in clustering texts according to their most internal meanings, and in collecting and gathering grammatical, lexical and syntactical forms. Also in this case, the limited dimension, the chronological and typological consistency and the richness of the corpus offer the opportunity of tuning informatic tools to the documents at the best, so that these methodologies could be employed in the future on larger corpora.

After all, since the 1970s, Assyriology used new technologies for computer encoding of cuneiform texts, to support the analysis and management of a great amount of data extracted from tablets. Important projects in digital Assyriology have been developed in the last decades, such as, for example, CDLI (Cuneiform Digital Library Initiative) and BPS (Berkeley Prosopography Service) with his open-source prosopographical toolkit, *ETANA* (Electronic Texts and Ancient Near Eastern Archives) and *eTACT*, *ORACC* (the Open Richly Annotated Cuneiform Corpus) with lemmatized corpora, used as a basis to extract glossaries on Sumerian and Akkadian, *DCCLT* (Digital Corpus of Cuneiform Lexical Texts), *SAAO* (State Archives of Assyria On-line), and the *Portal* Mainz with Hittite and Nuzi Texts.

In order to enhance access to text knowledge extracted from cuneiform tablets, the integration of Language Technologies (Multilingual Text Mining) and GRID Technologies offers a new perspective in the analysis of ancient texts in Assyriology. The GRID environment exploits supercomputer sharing computational resources distributed on

[1] The letters from Tell Harmal published by Goetze are fifty; nevertheless, in the commentary of IM 51198 (= n. 16) at p. 39 he has also quoted a large part of IM 51269, which has been taken into consideration in the present paper.
[2] Cf the letters from Tell Asmar, that cover a longer period of time [Whiting 1987].

many computers in the same or different site. The GRID Infrastructure 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.

ENEA, the Italian Agency for New Technologies, Energy and Sustainable Economic Development, has 12 different research sites located in northern, central and southern Italy. In particular, 6 of them host computational facilities to allow the execution of serial and parallel jobs and advanced graphical applications. Those resources consist in a multitude of hardware and software platforms, composed by libraries, compilers, remote applications, cloud computing environment, and virtual labs. This environment composes ENEA-GRID infrastructure. ENEA-GRID allows to share computing resources across collaborative projects, attracting, engaging and supporting a wide range of users and researchers from science and industry communities. For these reasons, ENEA-GRID facilities can be used in several contexts and fields of interest, such as weather forecasting, fluid dynamics, 3D applications, etc.

ENEA activities and research on the integration of language technologies (e.g., text mining) in the ENEA-GRID infrastructure provide a novel approach for studying and analyzing cuneiform corpora. In the domain of assyriology, ENEA-GRID offers a digital collaborative environment, to share knowledge and digital resources with the integration of Multilingual Text Mining Software, Lexical Resources and Data Visualization tools for Network Analysis. The GRID enables researchers to perform quantitative and comparative studies on transliterated cuneiform texts, providing access to computational resources for the storage and processing of large textual corpora. For this purpose, the TIGRIS Project (Toward Integration of e-tools in GRid Infrastructure for e-aSsyriology) started in 2008, and preliminary studies by P. Negri Scafa and D. Alderuccio [Negri-Alderuccio, 2009 and 2011] provided the basis for the activity for integrating and analyzing assyriological texts. Here, a first experimental application of new technologies to a corpora of Babylonian texts has been carried out using Multilingual Text Mining in ancient languages, in the prospective of an integration into the digital environment in ENEA-GRID. The main goal consisted in offering an innovative approach for the study of cuneiform corpora, from which software developments and collaborative working could coexist to match the specific needs of scholar communities [Negri-Alderuccio, 2009 and 2011].

In this paper, we start from the past experience on assyriological texts and propose a new strategy to analyze such data. In particular, our latest study consists in exploiting data mining algorithms tuned on assyriological transliterated corpora from cuneiform tablets [Ponti et al., 2012], supporting document clustering of assyriological e-texts. For this purpose, we developed ASTEC - *ASsyriology TExt Clustering*, a tool integrated in ENEA-GRID able to analyze transliterated e-text form assyriological corpora. Cuneiform texts have been pre-processed, and then clustering task has been executed employing the K-means clustering algorithm, which partitions the corpus into groups (i.e., clusters) according to the number chosen by the analyst.

As said above, we performed experiments on a corpus of 51 letters from Tell Harmal/Šaduppûm. A two-stage analysis has been performed: a quantitative analysis accessed the quality of the clustering task exploiting quality-based indexes known in literature, and a qualitative analysis aimed at describing data relations and affinities discovered by the clustering algorithm involving assyriologists. Our ASTEC tool and ENEA-GRID infrastructure allow to achieve high quality results in terms of quantitative analysis and, from a qualitative viewpoint, are able to highlight interesting hidden relations in the data.

The rest of the paper is organized as follows. We start describing the first stage activities performed in the TIGRIS project, and continue describing the integration of Computer science and Assyriology, illustrating the KDD process for text data, the clustering task and our ASTEC tool for analyzing ~~t~~ransliterated e-text. Then, the experimental section has been discussed, where we describe the analyzed corpus, the preprocessing phase, and the clustering results. At this point, starting from these clustering results an examination of some results will be made in order to understand more about these results and their relevance in the study of the administrative procedures of Šaduppûm. We conclude the paper illustrating the TIGRIS Web Portal and providing conclusions and future works.

The TIGRIS Project and e-Shnunna

The project "e-ŠNUNNA" (e-Shnunna, in the following) exploits Text Mining tools, in particular Clustering, to discover homogeneous groups and hidden relations in the data coming from that small homonymous Kingdom. At the moment, it concerns the small corpus of the letters from Tell Harmal. The focus of this case-study has been the definition of a methodology to analyze transliterated corpora from Ešnunna cuneiform tablets, exploiting informatics.

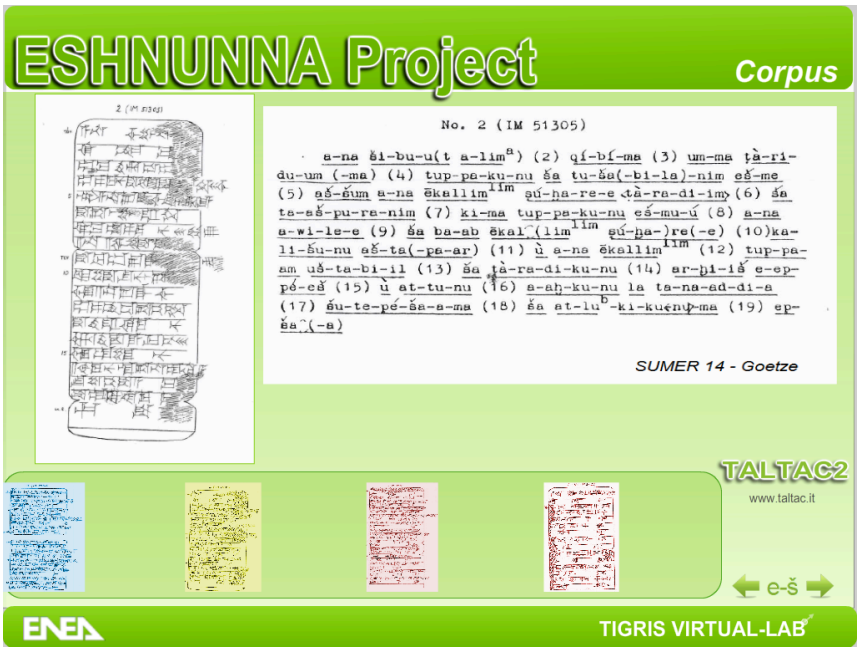

Figure 1 – The TIGRIS portal and e-Shununna letter visualization

A former experience in this field has been carried out. This first case-study is referred to the texts of the town of Nuzi [Negri-Alderuccio et al., 2011], belonging to the small Kingdom of Arrapḫa and located East of the Tigris River (XV-XIV centuries B.C.). Among the whole corpus of Nuzi texts, the best preserved texts belonging to the scribal family of Šeršiia (with his son Ḫupita and his grandson Muš-teššup) has been analyzed, in order to highlight innovative and conservative stylistic elements in the redaction of texts (administrative texts and contracts) through three generations. Documents have been tagged in order to highlight graphic and graphemic, grammatical, and prosopographical data. The small dimension of the corpus allowed for an exhaustive control of the results in the application of the software, that is TaLTaC2, which has been developed starting from a research carried out at the Universities of Salerno and Rome "La Sapienza", under the supervision of Prof. Sergio Bolasco (University of Roma "La Sapienza"). Preliminary results invite to continue with the application of these methodologies and to extend the scribal and textual analysis to the whole corpus of the Nuzi texts. In particular, studies performed on such a data provide relevant intuitions and guides to attempt also on old Babylonian documents. In particular, the geographical position of Ešnunna, East of the Tigris River, made that Kingdom a preferential objective for further investigations.

The ENEA TIGRIS Virtual Lab in e-Humanities is a research activity within the IT@CHA Project [IT@ACHA Project]. ENEA-GRID in e-Assyriology is an ENEA contribution to Safeguarding Material and Immaterial Knowledge from Ancient Mesopotamia.

Computer Science and Assyriology

Knowledge extraction from data

Computer-based systems are widely used to help users in organizing and managing everything produced by applications and/or business tasks. Within this view, computers and workstations are largely employed in industries and companies that need to use and analyze data generated by their processes and activities. However, although systems that archive and integrate data have improved their effectiveness and their efficiency in the last few years, it is difficult to explore this huge and heterogeneous amount of data to produce new relevant and strategic concepts. Analysts and system users need new instruments and techniques that support them in the extraction of patterns and relations hidden in the data.

Indeed, it results to be very important to distinguish between *data* and *information*. Usually, we refer to "data" as something that contains features and values related to a fact and that are used to describe an event by means of a set of variables. Within this view, *raw data* are useful only for descriptive purposes. Instead, with the term "information"' we refer to something that has crucial value and that provides indications; in substance, while data are fundamental to describe a phenomenon, information is everything that comes from the interpretation and the analysis of such description. Moreover, information is the only thing that can be used by analysts to understand the data and to outline strategies.

The entire process that aims to extract useful information and knowledge starting from raw data takes the name of *Knowledge Discovery in Databases* (KDD) [Fayyad et al, 1996]. The process starts from raw data and consists in a set of specific phases that are able to transform and manage data to produce models and knowledge.

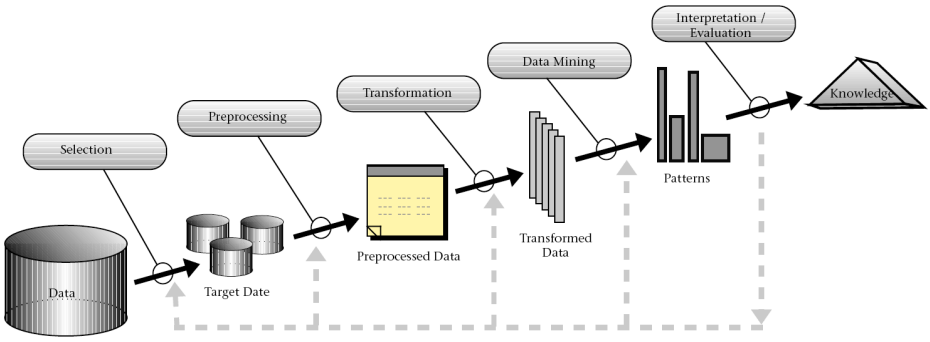

Figure 2 – The KDD process

Figure 2 shows the main steps involved in the whole KDD process. The first three steps (i.e., selection, preprocessing and transformation) substantially operate on raw data and perform filtering and transformation tasks. These phases are particularly useful to solve all the problems regarding data representation and data integration (as discussed above), and to prepare data for the next analysis processes. Data mining techniques (i.e., the fourth step of the KDD process) furnish a collection of procedures and algorithms to find out relations and hidden patterns within the data. This step has a great impact in the entire KDD process since it is responsible of producing good models that can summarize main data features and that are able to underline new trends, patterns and relations. The last step of the KDD process consists in the interpretation and validation of the results. It is important to note that this step takes, as an input, the output of data mining tasks, and not the raw data and/or the transformed ones; therefore, it exploits relations and models produced by data mining techniques and infers new information and knowledge.

Data mining and Clustering

Data mining is the non-trivial process that aims to elaborate data in order to extract hidden patterns and relations within the data. Data mining algorithms can be divided into two main families, that comprise predictive and descriptive task. Predictive tasks aim to build a model useful for predicting future behaviors or values of certain features. Predictive data mining algorithms are typically supervised, as data class labels are known and training dataset is used to build the predictor schema. Typical predictive algorithms are *classification* and *association rules*. On the other hand, descriptive task aims to build a model able to describe the data in an understandable, effective, and efficient form. The most relevant descriptive task is *clustering*, which identifies groups in the data in such a way that objects within the same cluster are similar and/or highly-related each other, whereas objects belonging to different groups are quite dissimilar.

Clustering algorithms are *unsupervised*, as they are applied to unclassified data (i.e., *unlabeled*). In such a scenario, the obtained clusters extract homogeneous groups and data analyst can underline hidden patterns and relations.

Text mining algorithms

Data mining algorithms are traditionally applied to structured data, which are typically in a relational scheme. However, in recent years there has been a large diffusion of digital information, especially in the contexts of World Wide Web (WWW) and Internet. This process leads to a massive proliferation of many communication channels, such as e-mail, forum, and digital libraries. Typically, data generated in these scenarios are *textual*, since they are

composed by a set of words (or terms), phrases, and paragraphs. Therefore, term *text* is often used to refer to *document*, which denotes a single unit of textual information.

The process of extracting new information and patterns from text data is known as *Knowledge Discovery form Text* (KDT). It is possible to easily draw a parallel between KDD and KDT, since KDT follows the same tasks of KDD, but it is applied to text data. It is straightforward that KDT has to face with several specific domain issues of handling text data. The main problems are related to the absence (or the poor presence) of any kind of structure within these data. The main difference between KDT and KDD regards the way data are represented. Indeed, preprocessing phase plays a crucial role in KDT, and it usually relies on experiences and results obtained by other related research areas, such as Information Retrieval (IR), Natural Language Processing (NLP), and Information Extraction (IE).

The way text data are represented is crucial in KDT. However, it has been shown how this step is challenging and how many aspects have to be considered to produce a structured while high-informative representation. In fact, since various levels of text structuring can be identified, i.e., *non-structured* (plain) texts and semi-structured texts (e.g., HTML and XML), there are many techniques that try to equip document modeling with structural and semantic aspects. Nevertheless, traditional approaches use only syntactics, since they are based on the *bag-of-words* model, which represents a text document by means of a set of words. The bag-of-words model needs a proper preprocessing phase to filter out from the text elements that are not relevant and to properly transform tokens to exploit all their informative content.

Text Mining for Old-Babylonian texts – ASTEC tool

In this paragraph, will be described the main aim of this paper, that is the application of data analysis technique on a transliterated corpus from cuneiform tablets coming from Ešnunna old-Babylonian kingdom. We developed a tool, called ASTEC - *ASsyriology TExt Clustering*, which offers a set of clustering algorithms and features to analyze and discover homogeneous groups of transliterated texts. Performing algorithms and setting up measure in the tool, it is possible to underline hidden relations within the Ešnunna corpus, in such a way that the analyst (i.e., the domain expert) is eased in discovering new patterns and information from the data.

The tool is customizable by the user, which is able to choose several settings, as the clustering algorithm, the relevance measure, and the clustering quality evaluation measure. In particular, the tool is currently equipped with two clustering algorithms, that are the well-known K-means [McQueen, 1967] (a partitional based algorithm) and the UPGMA [Jain et al., 1988] (a hierarchical agglomerative based algorithm), three relevance measures for terms, that are the DF (Document Frequency), the TF (Term Frequency), and the TF-IDF (Term Frequency – Inverse

Document Frequency), and two clustering evaluation measures, that are the Inter-Intra distance *Q* (internal criterion), and the F-measure (external criterion).

The tool is written in JAVA [JAVA], which makes it executable in every system as it is independent from the execution platform and makes particularly suitable for web environments. Moreover, it is modular and projected to be easily extended with other algorithms and relevance measures.

Experiments

In this section, we describe the goal of our experiments and the obtained results. In particular, we describe the main characteristics of the corpus, the preprocessing phase, and the tool settings. Then, we illustrate the advantage ENEA-GRID, of our execution platform, in terms of data accessibility and computational power offered by the CRESCO HPC systems. We continue with the cluster analysis, which lead to new interpretations and to discover new hidden relations in the data.

Data Description and Preprocessing

The 51 letters from Tell Harmal/Šaduppûm, object of our experiment are well-articulated prose texts, showing a more varied language and different widely ranging argument, and for these reasons they are particularly suitable for text analysis, lexicon and grammar extraction. Senders and Receivers were explicitly identified, and this will be useful for further prosopographic and geographic analysis and studies. The corpus has 2741 terms, with a mean of 53 terms per letter. The vocabulary in composed by 500 distinguished graphical forms identifying terms.

Row-cuneiform texts need to be *preprocessed* to be represented and analyzed by our tool. The first step consisted in transliterating cuneiform texts. In this phase of the project, we resorted to the work of Albrecht Goetze, in his “Fifty Old-Babylonian letters from Hermal” (1958), here slightly reviewed.

In order to apply the software, the transliterated texts must be in a ASCII character set; therefore we have adopted the following substitutions:

š --> $

ṭ --> v

ṣ --> c

ḫ --> h

For the second step of the preprocessing phase, we resorted to the TaLTaC2, a software application for the automatic analysis of texts according to the logics of Text Analysis (TA) and Text Mining (TM). TaLTaC2 is

particularly suitable for quantitative linguistic analysis. After the transliteration, each text becomes a file, and all files have been collected in a unitary Corpus, converting it into a TaLTaC2-machine-readable form and tokenizing it. The process of tokenizing segments text unit into "graphical forms" separated by white spaces or carriage return and line feed, whereas punctuation marks were not assumed as separator, due to their different function according to the transliteration methodologies in Assyriology. For Text Mining purposes, some variables (i.e., Id of tablets, Editor, Subject, Sender, Receiver, Time) have been added before each tablets text.

The corpus was then part-of-speech tagged and lemmatised by a domain expert (assyriologist). Then TaLTaC2 "Corpus Reconstruction by Lemma" transforms the text corpus into a word list for Data Mining. At the base of an Ešnunna grammar, there are the following steps:

- grammatical and semantic tagging,
- lemmatization,
- concordances extraction by lemma.

Tool settings

ASTEC offers several possibility of customization and setting, in order to execute fine-tuned experimental sessions. For our experiments, we resorted to the K-means algorithm, and employed two relevance measure for terms, that are TF and TF-IDF. In particular, we executed experiments by varying the number of clusters from 2 to 15, as that are a reasonable range of analysis for a corpus of 51 not-too-long texts.

As regards the representation technique, we resorted to the well-known Vector Space Model [Salton et al., 1968], widely employed in text mining context. Such a model starts form the bag-of-words assumption and represents each text as a vector in an V-dimensional space (where V is the vocabulary of the corpus): in this view, a document is seen as a vector of its terms, and each term is "weighted" according to the relevance measure employed. This model allows make text data suitable to be treated by a computer science algorithms and, in particular, by our ASTEC tool.

ENEA-GRID and CRESCO HPC Systems

ASTEC tool has been executed in our ENEA-GRID infrastructure. ENEA-GRID infrastructure [ENEA-GRID] provides a unified and homogeneous environment for the computational resources in ENEA, in such a way that researchers and their collaborators can perform their work despite of the location of the specific machine or of the specific hardware/software platform that is employed. ENEA-GRID manages heterogeneous resources and multi-platform systems in an optimized and transparent way. The ENEA-GRID infrastructure permits to access to all these

resources as a single virtual system, with an integrated computational power of more than 50 Tflops, provided by several multiplatform systems, i.e., AIX SP5 (~256 CPU), Linux x86_64 (over 6000 cores provided by CRESCO HPC clusters [CRESCO]), and special dedicated systems (e.g., GPU), and offers two different file systems, that are OpenAFS (distributed file system for ubiquitous access), and GPFS (parallel file system for high performing calculus). Distributed resources are handled by a resource manager LFS Musticluster, whereas resource monitoring has been demended to Zabbix software.

ENEA-GRID infrastructure provides a novel approach for studying and analyzing cuneiform corpora. In the domain of assyriology, ENEA-GRID offers a digital collaborative environment, to share knowledge and digital resources with the integration of Multilingual Text Mining Software, Lexical Resources, Data Mining tools, and Data Visualization tools for Network Analysis. The GRID enables researchers to perform quantitative and comparative studies on transliterated cuneiform texts, providing access to computational resources for the storage and processing of large textual corpora. Moreover, ASTEC tool is integratend in the GRID and exploits its computational power and accessibility features.

Cluster analysis

In the following, we describe the results obtained form the experimental sessions on the 51 transliterated texts. The clustering solutions produced by our experiments have been analyzed in a twofold aspect: form a *qualitative* viewpoint, we aimed to describe data relations and affinities discovered by the clustering algorithm involving a domain expert (i.e., an assyriologist), whereas from a *quantitative* viewpoint, we aimed to assess the quality of the clustering exploiting quality-based indexes known in literature.

It is important to note that the Ešnunna corpus is composed by letters referring to activities and actions performed by public officers of a provincial town in relation to the capital. It is expected that the main aggregating aspect in clusters is something related to the procedures, citizens and goods. In fact, exploring the results of our experiments, it is possible to notice that texts regarding the same officer are often grouped together by our tool, and that relations between officer and palace are dominant in many clusters. It is important to say that when we refer to “palace”, this is the Ešnunna palace (i.e., the capital palace, which can be considered, under several aspects, the palace of the king), whereas when we refer to “city”, “citizens”, and “Council of the Elders”, these are related to the city of Šaduppûm, that is Tell Harmal, where the 51 texts come from. As regards goods treated in the letters, the majority of them are barley, and this is underlined by the lots of uses in which it is involved (e.g. food, drink constituent, seeding element, trading element). Clusters discovered by our ASTEC showed frequently that the different ways in which barley is employed are important for aggregating letters.

We have performed a thorough experimental session by varying the number of clusters and the relevance measure for terms. It is clear that here, for the sake of the brevity, we provide only some global remarks and a little insight into the most relevant specific aspects of some cluster. For global remarks, some interesting relations come into light from clusters, that are officers and procedures, palace/king and the official documentation, citizens and the Council of the Elders, work manners and activities, barley managing and employing strategies.

From a different viewpoint, it is interesting to analyze clusters observing the subject of the letters. More specifically, when analyzing a cluster discussing barley, there are also letters referring to bovines and to agricultural works; in particular, it is interesting to observe that bovines are highly related to the palace, whereas the most common ovines are only marginally cited.

As regards the relevance measures for terms, TF and TF-IDF provided different manners of cluster identification. In fact, TF discovers clusters the are more "raw", in the sense that cluster interpretation and discovering relationships among letters is a more trivial task w.r.t. the one in the case of TF-IDF. This is due to the different nature of the two measures, as TF is more sensitive to conjunctions, adverbs, and all the other terms that frequently occur in the majority of the letters, whereas TF-IDF tends to give most importance to highly-specific terms that occurs only in few letters (called *hot topic terms*) and, consequently, to exclude terms not relevant for topic identification.

From a quantitative viewpoint, we analyze clustering results of our ASTEC in terms of compactness of the cohesiveness of the achieved clusters. For this purpose, we resorted to the analysis of the Q index, that is based on inter-similarity and intra-similarity criteria. Q ranges within [-1, 1], as -1 is for lowest clustering quality and 1 for highest one. Results showed that clustering produced by ASTEC on Ešnunna achieves quality results form 0.3 to 0.8 (by varying the cluster number), which indicate high quality clustering results. In particular, it is interesting to note that quality results are independent from the relevance measure employed (i.e., TF and TF-IDF): this results is expected, as, in general, cohesiveness is related only to the corpus analyzed and to the clustering algorithm employed, whereas relevance measures are crucial for qualitative analysis and for discovering relationships.

In this context, it might be interesting to draw a parallel between two clusters: it is an example of how we can make use of the indications provided by this complex procedure; for reasons of space it is not possible to offer more than this example, but additional studies on the clusters obtained are scheduled.

The two clusters have been obtained with two different applications of the measure TF-IDF. As it not possible to show and enumerate all the clusters and all the algorithm applications, we simply refer to the first cluster as A, and to the second as B. The two clusters, despite they derive from two different applications of the measure, are very similar, since they are composed by the same texts: IM 51114, IM 51192, IM 51229, IM 51238b, IM 51240, IM 51270, IM 51305, IM 51365, they are different only for IM 51180 and IM 51503 (Cluster A) and IM 51193, IM

51311 and IM 51376 (Cluster B). From the point of view of content, we can see that the key points of these clusters are the fields: as a place to work barley, for the problems of irrigation, as a form of payment of a *šangûm*. This latest point introduces another element of connection: officials busy in solving problems, however, require the intervention of the palace; the palace finally is the recipient of messengers and needs of animals (cattle) of quality. If the use of these measures highlights the points of contact, it is also interesting to underline which are the elements of the texts that differ in the two clusters. Even in this case, beyond the obvious differences, there are, however, some contact points. There are specifically relevant legal issues (expressed by tablets: *ṭuppum*), or problems regarding personnel.

But of particular interest are IM 51311 and IM 51503: each text belongs to a different cluster, but both of them show an interesting affinity: problems and procedures in assigning responsibility to local authorities. In one, well-known, case, the Council of Elders moves and solicits an official to assume the power on the town; that official sends them to the king. This is one of the Mesopotamian texts, that concern a discussion on "democracy in Ancient Near East"[3]. In the other text the essential involvement in the solution of practical problems arising from the death of a *šakkanakku* (the main authority in a town) is taken into consideration. It is evident that two different situations are concerned, but the two texts show significant points of contact. These aspects related to the procedures, detected and shared in the documents of the two clusters, confirms the importance of this tool.

TIGRIS web portal

The Virtual Lab is a tool available in the ENEA-GRID for users, projects and activities in ENEA that could need external visibility and remote access to the systems. In essence, it is a web portal containing information and access tools to the ENEA-GRID computing platform used for the project's aim. Particularly, in the Tigris VLab case, the web portal is configured this way:

- the Home page, containing information and details about the project
- informative pages about people and groups involved
- access portal to computing resources
- thematic areas containers
- links to other projects and studies connected to VLab activity
- developing tools
- past and future events relevant to the project.

[3] S. Viaggio 2000 with bibliography.

Figure 3 – ENEA VLab for TIGRIS project

Figure 3 shows the home page of the ENEA VLab for TIGRIS project [Tigris Portal]. All the data contained in the portal, as well as many services that it provides and all the software in GRID, are stored in the ENEA geographically distributed file system, named AFS. This system allows the data to be accessible everywhere, in any worldwide workstation online, inside or outside the ENEA network, and the only user requirement is the authentication mechanism, based username and password and on Kerberos5 security layer.

Another feature of this system is that also the computing machines, and all software running inside, can access same data area, as well as the ENEA web server for publishing information; this implies the possibility, if such is what desired, to publish the elaboration data results immediately after the software ends.

Conclusion

As noted already elsewhere, the present Project, with software application for Text Analysis (TaLTaC2) and Data Mining tools (ASTEC), offers a significant example of integration between Humanities (Assyriology) and Computer Science. Moreover, features offered by ENEA-GRID infrastructure and the computational power of ENEA CRESCO HPS systems provided a perfect environment to execute experiments and data analysis tools in an efficient effective way, and offer also the possibility of ubiquitous access to our research by means of the ENEA Web Portal for TIGRIS Project and e-Shnunna.

What we saw with our former experiences appears in this phase reinforced: therefore, results achieved in this phase of our activities, just compared with the progress from our former experience, invite us to carry on with the Project.

Several steps are urging for future work and to continue our research:

1. To start off with the grammatical and syntactical tagging of the texts and to extend our research to the other texts of the Kingdom of Ešnunna.

2. To go on with the clusters analysis and utilization: the analysis will suggest eventual improvements, while the utilization will offer stimulating points of view in texts analysis. On the ground of these experiences further implements in the algorithm application can be attained.

3. To continue to develop and to improve the Portal. Particular attention we intend to give to the Portal, with further implementations and developments, in order to make use of it either as an information and knowledge tool and as a communication tool, opening to possible exchanges of experiences within other scholars groups.

4. Activities on middle Babylonian Nuzi texts will be resumed: this study section will certainly enjoy of the progress in grammatical and syntactical tagging of the texts that originates from e-SHNUNNA Project.

Therefore, we can conclude that the integration between Humanities (Assyriology) and Computer Science facilities in ENEA-GRID we referred to above have a synergic characteristic, by which each discipline enriches of tools and experiences, with perspectives of significant progresses either in Text Analysis, in Text Mining and Data Mining, and in Assyriology.

Bibliography

CRESCO ENEA HPC system: http://www.cresco.enea.it/

ENEA-GRID: www.eneagrid.enea.it

Fayyad, U.M. et al.

1996 From Data Mining to Knowledge Discovery: An Overview. In Advances in Knowledge Discovery and Data Mining, pages 1–34. AAAI/MIT Press

Goetze, A.

1958 Fifty Old-Babylonian letters from Hermal, *Sumer* 14

IT@CHA Project. The IT@CHA Project in ENEA: http://utict.enea.it/it/progetti/utict-e-i-progetti/it-cha

Jain, A.K., Dubes, R.C.

1988 Algorithms for Clustering Data. Prentice-Hall

JAVA Programming Language http://www.java.com

McQueen, J.B.

1967 Some methods for classification and analysis of multivariate observations. In Proc. Berkeley Symposium on Mathematical Statistics and Probability, pages 281–297, 1967.

Negri Scafa, P., Alderuccio, D.

2009 A new experimental approach to text computer analysis applied to the Nuzi texts, 55th RAI 2009, "Family in the Ancient Near East: Realities, Symbolisms, and Images", Paris, France

Negri Scafa, P., Alderuccio, D., Bracco, G., Migliori, S.

2011 A preliminary analysis of a Nuzi scribal family in view of an application in the ENEA-GRID, 57th RAI 2011; "Tradition and Innovation in the Ancient Near East", Rome, Italy

Ponti, G., Negri Scafa, P., Alderuccio, D., Mencuccini, G., Rocchi, A., Bracco, G., Migliori, S.

2012 Toward the integration of informatic tools and GRID Infrastructure for Assyriology Text Analysis, 58th RAI 2012; "Private and State (Privat und Staat) in the Ancient Near East", Leiden, The Netherlands, July 16th-20th

Salton, G., Lesk, M.E.

1968 Computer Evaluation of Indexing and Text Processing. Journal of the ACM, 15(1):8–36

Saporetti, C.

1998 *Epigrafia di Tell Harmal*, Pisa

2000 Siti storici nella Valle della Diyāla. Passato e presente, *Geo-archeologia*, 1

2002 *La rivale di Babilonia. Storia di Ešnunna ai tempi di Hammurapi*, Roma

Viaggio, S.

2000 Tell Abu Harmal / Šaduppûm, in Saporetti, 2000, *Geo-Archeologia*, 1: 11-26

Whiting, R.M.

1987 *Old Babylonian Letters from Tell Asmar*, AS 22 Chicago