ANDREA BOZZI

MS. BOZZI:

Thank you very much for your invitation to present my activities at the Institute for Computational Linguistics in Pisa **[slide 1]**. I would like to show you some important aspects about the management of old documents in digital format. You know that digital libraries are very useful not only for general purposes (data navigation, information retrieval, etc.), but also for more specific applications, in particular for historical linguists and philologists.

Therefore, I will be speaking about the computer-aided, philological and linguistic analysis of digital sources. My current activities are concerned with computational philology, of which I am going to present five different aspects.

Firstly, I am going to show you a European-supported system for Greek papyrology, then a special application for the browsing and searching of documents on *ostraka*, followed by a philological workstation for digital medieval manuscripts. We are then going to see a cultural heritage language technology system - called LEMLAT - for lemmatization of Latin texts, and finally the possibility of integrating all these modules in a web-based open source application **[slide 2]**

1. An EU supported system for Greek papyrology **[slide 3]**.

   The system, not yet available in the web, is a desktop application which allows scholars to study, analyze and transcribe Greek papyri **[slide 4].** The digital image on the left-hand side of the screen can be enlarged and enhanced by filters in order to facilitate reading and transcription of the text contained in the papyri. At the present state of technology it is not yet possible to perform optical character recognition, which we hope we will be able to do in the future using artificial intelligence technologies, such as the neural networks. A very interesting feature allows users to link each word zone within the image with the corresponding word in the transcription **[slide 5]**. This operation can be done in part automatically, and in part semi-automatically, correcting the mistakes. It is also possible to obtain the concordance of all the word-zones in which a specific word appears and the corresponding word in the transcription: such a concordance is very useful to correct any erroneous text interpretation. Some years ago we tried to reconstruct words contained in a number of fragmentary papyri using statistical data processed on the basis of the *Thesaurus Lingua Graecae* implemented by Ted Brunner at the University of California, Irvine. This methodology was also applied to papyri of comedies and tragedies, with satisfactory results.

[slide 6] At the bottom of the screen we can add some information regarding different types of apparatuses (papyrological, prosopographical, variants), as well as annotations for words dealing with special semantic fields (proper names, technical terms, etc.).

[slide 7] All this information can be indexed by the system, and if we select, for example, the “T” button, all the wordforms present within the text are shown in alphabetical order making it possible for the user to browse the contexts in which each wordform appears. By selecting the “PA” button, the index of the words as they appear in the image, with no accents and spirits, can be read in alphabetical order. “PR” stands for proper name: the indexation system is able to produce the list of this lexical group as long as the scholar has marked it up during the transcription phase with a specific flag. By “ME” we intend all the medical terms, while “V” refers to the variants. All these words are highlighted in the text by different colours.

[slide 8] This slide shows the first attempt made to move from the original desktop application for papyrology to the web-based application: the technological activities have been rather complex with particular regard to the software which produces the automatic linkage between words in the images and corresponding words in the transcription. However, this web application is now being tested at the Vitelli Papyrological Institute in Florence with restricted use, because these documents have not yet been published and are copyrighted. The documents are related to medical pharmaceutical prescriptions written in the fourth century A.D.

2. **[slide 9]** I would now like to present a specific adaptation of the same philological computer assisted workstation managing demotic documents written on *ostraka*. This application is very important also to teach and train people who want to understand the demotic script.

   [slide 10] These are three pieces of ceramic documents which need to be interpreted. This is particularly difficult because the writing moves from right to left; moreover, it is really complicated to understand the correct combination of the signs, because it is possible to link them in a certain number of different ways obtaining different words and different textual meanings.

   [slide 11] In order to partially overcome these problems, we designed and realized a virtual keyboard (you can see it on the right of the slide) with the possibility of segmenting the part of the demotic text with one of its most suitable meanings and of assigning to it the correct symbol available on the keyboard. At the end of the segmentation process of the whole *ostrakon* archive (more or less one hundred documents) it is possible to retrieve all the places **[slide 12]** in which a specific symbol selected on the keyboard appears in the demotic documentation so as to check whether your interpretation has been correct or not. In fact, you can easily retrieve all the words within the entire *ostrakon* archive which contain the selected sign or sign combination. This procedure looks like a very special concordance program. The system is used in the Antiquity Department in Pisa to teach the demotic script to students, so that they can better study and understand the important archive discovered in Egypt by our colleague Edda Bresciani.

3. **[slide 13]** The third part is referred to another specific use of my philological workstation for digital medieval manuscripts. This application has been tested on old Provençal manuscripts considered as different witnesses transmitting the same text, but with some variations (mistakes, variants). The final aim of the project is to produce the electronic critical edition of the text (scholarly edition).

   [slide 14] The position of the data on the screen is quite similar to the one I have shown so far: the image on the left, and the transcription carried out manually by the physiologist on the right. At the bottom, the list of the other witnesses of the same text is available. In this case, only a Princeton manuscript has been collated. If you click on the button placed immediately before the name of the collated manuscript, you obtain **[slide 15]** the visualization of the image of the Princeton manuscript in which it is easier to find all the text variants. In this case, we have the form “eixens”, which is a graphical phonetical variation to be recorded in the specific field of the critical apparatus **[slide 16]**; it could be used, alongside other information, to reconstruct the original text - as it was written by its author - which has gone lost but handed down on us by copies realized during the medieval ages.

As I said before, also in this case the system is able to produce the automatic index of all the wordforms of the transcription and of all the variants in the critical apparatus. In addition, the link between variants, text and image is always provided. So far, we have tested this system for no more than 10 witnesses.

[slide 17] The same application has also been used for the philology of ancient printed books. In this case the transcription has been done automatically using not a commercial OCR (Optical Character Recognition System), but a Neural Networks engine which we implemented some years ago. The difference between our system (LaperLA: Lettore Automatico per Libri Antichi) and the commercial ones is that we link the segmentation and interpretation of each word with a Latin dictionary considered as a sophisticated and statistically controlled spelling checker. The thesauri adopted by the commercial products are poor while we have a very rich look-up table of wordforms and a great many statistical combinations between Latin characters. In this way it is quite easy to check the mistakes and propose the substitution of the misinterpreted words with the correct ones.

On the bottom left-hand size of the screen the user has written the names of four different printed editions of the same work with the aim of finding eventual variants. If he selects the digital images of the volume printed in Marburg **[Slide 18]**, the application shows the corresponding pages, where it is possible to read any textual differences between the *Editio Princeps* and the others. In the same way described for the manuscripts, the user can record the variants within the specific field of the apparatus. The program produces automatically the linkage between word in image / word in transcription. This means that the selection of whatever word in the image of the *editio princeps* produces the visualisation of the same word in the transcription, and viceversa. The same happens if the selection is done on whatever wordform in the alphabetical index.

4. **[slide 19]** Let us now see a morphological automatic analyser of Latin texts able to produce the lemmatization of Latin works. This is another module which we intend to link to the other ones I spoke about before. Some years ago, with the cooperation of the Classical Department of Turin University, we prepared an appropriate dictionary to be converted in machine-readable form and a detailed list of grammatical rules. Originally installed on mainframes, then updated to run on PCs and now active on our Web site (www.ilc.cnr.it/lemlat), such dictionary is able to perform a very satisfactory linguistic analysis with particular regard to lexicographical projects. In the next slide **[slide 20]** I would like to show the results obtained on the work *De coniuratione Catilinae*, 1-2, by Sallustius. *Aestumo*, for example, appears in the text with only one occurrence. This means that this form is not a homograph. In the next slide **[slide 21]** the segmentation process is shown with its morphological analysis; but it is interesting to note that the proposed lemma is *aestimo*, which is in the base dictionary: as a matter of fact our system is able to manage the graphical variation.

The same occurs in the case of *inmutatur* lemmatized as *immuto*, because our rules have been designed to overcome the assimilation and dissimilation problems.

5. **[slide 22]** How can we integrate all these modules in a web-based open source application? The aim is really ambitious but difficult, because the software already running on stand-alone machines is not automatically transportable in a web architecture. Moreover, the use of the international standards has to be strictly considered. As a possible solution, we found that the Pinakes 3 system **[slide 23]**, developed at the Institute and Museum of the History of Science in Florence, could represent the ideal solution to our purposes. In fact, the aim of Pinakes 3 is the creation of a web-based open source application managing cultural heritage historical data in digital format.

   [slide 24] The technology we have adopted is described briefly in this slide: as you can see, none of the components are proprietary.

   The last slide **[slide 25]** lists all the standards we are using strictly.

   In a few months' time philologists will be able to apply this open-source system for text criticism, since it has been tailored to the needs of these scholars.

   For any further information about the methodological aspects of the system, I suggest to refer to the Proceedings of a conference on digital technology and philological disciplines that I organized for the European Science Foundation.

   Thank you for your attention.

(Applause)

MR. BOBLEY: Okay. Any questions before we move on to the next speaker? Arne.

MR. FLATEN: I wonder whether there are any plans to use your applications to digitize material coming from the cadastre and from the Archive of State, in Florence.

MR. BOZZI: We have never used this system for Archives of State, but we have gained used a lot of experience working with huge amounts of documents stored in the Archives of Cultural Foundations, for example the Vieusseux Cabinet in Florence, or the Primo Conti Foundation in Fiesole, especially important for the history of Futurism. A total of ca 300,000 pieces has been managed so far. We hope we will be able to apply the final version of the "Pinakes Text" to Galileo's manuscripts describing his telescope on the occasion of the International Year of Astronomy in 2009.

MR. FRISCHER: This is a wonderful achievement and you deserve a lot of credit for it. It's very difficult to go from manuscripts and to be able to transcribe them quickly and at the same time to provide the information that editors need. This is really wonderful.
I'm wondering if you've thought about a next step, especially for Latin; now the lemmatizer could make it possible, which would be syntactic analysis, which would be so important for especially quantitative linguistics. And possibly start lexicometrics and authorship studies for example.

MR. BOZZI: Thank you very much for this question because obviously also for the automatic disambiguation of the homographs, it is very important to have a morpho-syntactic analyser. In this respect, I have convinced a young colleague of mine, who is now working at the Catholic University in Milan, to use a dependency grammar available in Prague, specifically designed for the construction of syntactic trees (Treebank). He is now testing this system on the Latin Index Thomisticus archive. This means that in the next future, we will have not only the morphological analysis of a Latin text, and not only its lemmatization, but also the syntactical position of each word within the frame of the sentences. Consequently, it will be possible to retrieve linguistic information not only from a linguistic or semantic, but also from a syntactical point of view. Thank you for your question because it allows me to emphasize the linguistic part of my activities, which I intend to develop in the future. Of course we hope we will be able to continue our research, considering the lack of funding especially in the field of the Humanities, due to the stagnant economic situation in Italy.