Report on the visit to the CEL (Vila Real, Universidade de Trás-os-Montes e Alto Douro) 17th and 18th of November, 2011

Masayuki TOYOSHIMA

Research Institute for Languages and Cultures of Asia and Africa (ILCAA),
Tokyo University of Foreign Studies, Tokyo, Japan
mtoyo@joao-roiz.jp

I was honored to be invited by the CEL of the UTAD on the 17th to the 18th of November, 2011, to conduct a joint research on the development of databases of the resources on the Missionary Linguistics, especially those resources of the Portuguese tradition, e.g. orthographies, grammars, and glossaries (dictionaries) by or on the Portuguese language.

1. Current state of the development of databases of the resources on the Missionary Linguistics of the Portuguese tradition

Since November, 2009, UTAD and ILCAA have been in close relationship on building distributes databases of the resources on the "Missionary Linguistics" in the "Grand Voyage Era", especially of the Portuguese tradition. Since last year (2010), part of the resources developed by the University of Aveiro has also been incorporated, e.g. "Vocabulário portuguez e latino" by Raphel Bluteau (1712 – 1728).

UTAD has already compiled almost all the grammars and orthographies of Portuguese before early 20th centuries, e.g. "Gramática da língua portuguesa" by João de Barros (1540). ILCAA has compiled some of the grammars of Asian languages written in Portuguese, such as "Arte da língoa japoa" by João Rodriguez (1605), and dictionaries of Asian and Portuguese languages, e.g. Dictionarium latino-lusitanicum ac japonicum (1595).

A part of these resources are already available as an on-line database publicly accessible in http://joao-roiz.jp/LGR/ . There are also other resources by the University of Aveiro, the University Nova de Lisboa, the CEDOCH of the University of São Paulo, and the "Corpus do português" by Mark Davies, etc.

During my stay in the UTAD, Prof. Dr. Carlos Assunção, Prof. Dr. José Barbosa Machado and I examined these resources and their interfaces, and found that in most of these resources, lemmatization and normalizations, such as unifying variant spellings (e.g. effeyto/effeito/efeito, cousa/couza/coisa), are insufficient or inadequate, and thus the lexical statistics thereof are not always reliable. The "Lexicon5" software developed by

Prof. Dr. Machado has almost overcome this problem by preparing an extensive internal dictionary of lexical variants, but this program works only as a stand-alone PC program, and cannot be on-line unfortunately.

2. Further development of databases and future plans

Discussion on these problems of the current state led us to the conclusion that we will enlarge our project into a new one, whose objects are :

1. to further exchange resources between UTAD and ILCAA, and other existing projects
2. to ameliorate the quality of these resources, some of which are still under proof-reading process (e.g. "Vocabulário da língoa de Japam com a declaração em portugues", 1603)
3. to create editions of still un-edited resources, such as "Dictionarium annnamiticum lusitanum et latinum" by Alexandro de Rhodes (1651) and "Dicionário português-chinês" by Michele Ruggieri and Matteo Ricci (1588?).
4. to enhance the on-line databases by constructing dictionaries for lemmatization and normalization (*canonicalization*) out of the resources, which are already available, and which are going to be created as above.

The resources of UTAD have been handed over to ILCAA, and the resources of ILCAA are going to be transfered to UTAD according to the conclusion above. These resources are going to be transformed into multi-lateral on-line databases in early 2012, in order to create a *canonicalization* dictionary.

The un-edited resources (Rhodes' and Ruggieri's dictionaries) are planned to be edited starting in this year, and will be completed in 2012. For editions of Vietnamese and Chinese, further cooperative actions are planned to be taken.

In order to facilitate the project, we have agreed on a rough sketch of a future project, possibly funded by external foundations, titled "Tesouro da língua portuguesa com as fontes asiáticas", a cooperative research environment formed by several Portuguese universities led by UTAD, University of São Paulo, and ILCAA.

3. Conclusion

The stay in UTAD was not long, but the resources which were already prepared by Prof. Dr. Carlos Assunção and Prof. Dr. José Barbosa Machado enabled me to achieve quite efficiently a fruitful result in development of the linguistic resources of the Missionary

Linguistics. I would like to extend my sincere gratitude to CEL, UTAD, that kindly enabled this visit, and would like to thank especially Prof. Dr. Carlos Assunção, for having prepared and conducted the whole program of this opportunity.

Nov. 29, 2011,   Tokyo, Japan
豊島正之 (TOYOSHIMA, Masayuki)