Development of the interdisciplinary master program on Computational Linguistics at Central Asian universities/ CLASS

Co-funded by the
Erasmus+ Programme
of the European Union

CLASS Master Program Implementation Risks

WP1.2

Report 2.1

Zygmunt Vetulani (AMU CLASS coordinator)

23.03.2018

Adam Mickiewicz University in Poznań, Faculty of Mathematics and Computer Science
Dept. of Computer Linguistics and Artificial Intelligence
Contact: AMU CLASS Coordinator, Prof. Dr. Zygmunt Vetulani, +48 6-1777296, vetulani@amu.edu.pl

Introduction

In this study we consider the CLASS project risks in the context of the following two assumptions:
- it is assumed that teaching should be at the quality level represented by good EU universities or higher,
- the proposed teaching organisation will strictly observe the results of Bolonia Process.

The third assumption is related to the teaching objectives that we define as follows:
"formation of well trained professional staff apt to undertake a collective work in creation of language technologies (both software and language resources) as a part of professional teams of language technology developers (supervised by NLP and HLTs experts); preparation for professional career in the sector of language industries. Alumni will constitute the core of future highly competent human resource to launch language industry for the concerned languages."
(cf. our report "*Analysis of international master programs by AMU - with recommendations*")

Our observations are based on the analysis of several European master programs in Computational Linguistics. We analysed several MA/MSc CL programs from Europe (7) and the USA (2) elaborated/implemented in the following countries: Australia, China, Czech Rep., France, Germany, Italy, Malta, Netherlands, Spain, Sweeden, UK, USA. These observations were confronted with our personal teaching experience.

We explored the followin web sites:
https://lct-master.org/ (this master program was implemented in several countries)
http://www.brandeis.edu/departments/computer-science/comp-linguistics/
https://flov.gu.se/english/education/masters-second-cycle/mlt/programme-syllabus
http://www.uni-heidelberg.de/courses/prospective/academicprograms/computerling_ma_en.html
http://www.sas.rochester.edu/lin/graduate/MS.html
https://www.uni-stuttgart.de/en/study/study-programs/program/Computational-Linguistics-M.Sc./
http://www.sfs.uni-tuebingen.de/en/courses-of-study/courses-of-study-at-the-sfs/international-studies-in-computational-linguistics/international-ma-programme-iscl.html
http://www.uu.se/en/admissions/master/selma/program/?pKod=HSP2M
http://courses.wlv.ac.uk/course.asp?code=WL050P31UVD

Risk categories

We have identified three main risk categories related to:
- language resources and tools,
- teaching staff,
- students.

Risks related to language resources and tools

Risks identification

The curriculum must meet two challenging requirements:
- offer high level formation in General Computational Linguistics (what means that students, in order to become educated consumers of CL-based top technology should get familiar with the state-of-the-art of CL for leading CL languages /first of all English/),

- provide students with practical know-how of development of the CL-based technology for the CA languages of the project: Uzbek and Kazakh.

This second challenge requires, besides the good familiarity with linguistic description of the concerned languages, **availability of basic resources and language engineering tools** for these languages, as text/speech corpora (mono and multilingual), digital computer-processible dictionnaries, grammars, parsers, tree-banks, language ontologies (e.g. in form of wordnets). Because of specific linguistic features of both Uzbek and Kazakh languages, these resources are of first importance if one attempts to form highly competent alumni well prepared to face challenges of pushing forwards development of the Uzbek and Kazakh language industries.

It is well known that public availability of the required resources and tools for the two concerned languages is low. E.g. they are wery rare in the reputable repositories of ELDA and META-NET (META-SHARE). This situation generates a major risk to slow-down the expected effect of rapid growth of language technologies for the concerned languages.

Suggested counter-measures
The first step to be done is the possibly complete listing of existing language resources and tools for the project CA languages. Two subcategories are of the first concern:
- resources and tools publically available (at least for teaching and research),
- resources and tools that are owned by the CLASS project universities (ready to use or in development or testing phase).
On the basis of this study, it is recommended to the CA consortium partners to agree on common use of all these tools, or of a selection of them, for teaching purposes (including their further development within the student projects being part of their studies). It is important to notice that all intellectual property and ownership issues must be solved at the initial stage of the suggested collaboration. A common CLASS project repository for language data and tools will be useful.

The second remedy measure is to mobilise students of the master program to produce or to enhance the already existing resources and tools. There is a broad spectrum of possibilities. For example, students, organised in mini-teams of 2-4 people, may contribute to creation of language resources under suprevision of teachers within the (mandatory) student projects.

Another possibility consists defining and proposing to the students themes of the master-thesis-projects generating desired language resources (subgrammars, dictionary modules, specialised corpora, wordnet fragments, etc.) to be further re-used in teaching. Also, students may contribute to the enhancement of disposable linguistic resources within the Erasmus+ internship mobility program. Internship at an external university may (but not needs to) be related to the student's master thesis project. We recommend involvement in the internship program first of all the CLASS consortium partner universities. (In that case special inter-university arrangements may be necessary).

Teaching staff related risks
Risk identification
Launching master programs requires top level teaching staff. It is an usuall promotional practice to attract potential students via publication of the teaching staff lists, possibly including international worldwide known professors. It is a generally observed standard to recruit the

teaching staff with research experience in the subjects they are supposed to teach. To comply with these standards may be hard, especially for young faculties with no long date research experience in the CL field, as formation of the university's own teaching staff is time consuming. (It requires minimum 4 years of PhD studies plus further acquisition of academic teaching practice.) Recruitment of highly experienced external teachers is very costly (cf. tuitition fees at the universities that we have listed in the Introduction).

Suggested counter-measures

To cover the broad spectrum of subjects from linguistics, computers science and related fields, involvement of external staff may appear useful. An exchange of teaching services between the CLASS partners (between CA partners as well as between CA and EU partners) is to be considered. The existing Erasmus+ staff mobility exchange program may appear very helpful to reduce the staff problems. This issue requires further analysis within the CLASS project.

Risks related with student recruitment

Risk identification

Computational Linguistics (CL) is an interdisciplinary research and industrial domain belonging both to linguistics and computer science. The CL community is heterogenuous. The initial background of many of today's experts is either linguistic or computer-scientific. Nowedays more and more BA/MSc CL student candidates have already BA/MSc degree in CL. Consistently, BA or MSc in CL is very often a formal recruitment requirement at the analysed EU univeristies. This will however not be the rule for the CLASS CA partners. Although one may expect many candidates to start CL master studies as defined in the CLASS project, this group certainly will be heterogenous in what concerns the initial background. There is a risk of withdrawal of students during or after the first semester already.

Suggested counter-measures

To face risks of mass withdrawal special attention must be payed to the compensatory classes of the first semester (e.g. *Introduction to Programming for NLP* and *Introduction to Computational Linguistics* in our exemplary master program defined in the report *Analysis of international master programs by AMU - with recommendations*). These classes must be carried out by lecturers experienced both in linguistics and computer science and being able to anticipate and to solve problems that student may have with assimilation of the content distant from their initial bacground.