ERASMUS+ PROJECT EXPERIENCES ON THE COMPUTATIONAL LINGUISTICS MASTER DEGREE IN THE CENTRAL ASIA UNIVERSITIES

N. Abdurakhmonova[1], M. Aripov[2], A. Duarte[3], K. Georgouli[4], C. Gómez-Rodríguez[5], Y. Kandalina[6], B. Maia[7], A. Sharipbay[8], U. Tukeyev[9], G. Urazboev[10], Z. Vetulani[11], O. Yusupov[12]

[1]Tashkent State University of the Uzbek Language and Literature (UZBEKISTAN)
[2]National University of Uzbekistan (UZBEKISTAN)
[3]University of Santiago de Compostela (SPAIN)
[4]University of West Attica (GREECE)
[5]University of a Coruña (SPAIN)
[6]Kostanay State University named after Baitursynov (KAZAKHSTAN)
[7]University of Porto (PORTUGAL)
[8]Eurasian National University named after Gumilyov (KAZAKHSTAN)
[9]Al-Farabi Kazakh National University (KAZAKHSTAN)
[10]Urgench State University (UZBEKISTAN)
[11]Adam Mickiewicz University in Poznań (POLAND)
[12]Samarkand State Institute of Foreign Languages (UZBEKISTAN)

Abstract

The paper presents the experience of implementing the ERASMUS+ project “Development of the Interdisciplinary Master's Program in Computational Linguistics at the Universities of Central Asia (CLASS)” No. 588545-EPP-1-2017-1-ES -EPPKA2-CBHE-JP-CLASS in seven universities in Central Asia, namely, in three universities of Kazakhstan (A. Baitursynov Kostanay State University, Eurasian National University named after Gumilyov, Al-Farabi Kazakh National University) and four universities in Uzbekistan (Urgench State University, Samarkand State Institute of Foreign Languages, Tashkent State University of the Uzbek language and literature, National University of Uzbekistan in Tashkent). The European participants in the CLASS project are: University of Santiago de Compostela, Spain; University of a Coruña, Spain; University of West Attica, Greece; University of Porto, Portugal; Adam Mickiewicz University in Poznań, Poland. The relevance of training in computational linguistics in connection with the active development of artificial intelligence and its use in various industries is one of the top priorities at present. Within the framework of the project, the joint interdisciplinary master's program in computational linguistics was developed, and includes the development and teaching of such relevant disciplines as: Machine Learning in NLP, Language Resources, Language Analysis, Speech Processing, Machine Translation Technology, Understanding Natural Language, Ontologies and Semantic Technologies, Statistical Methods in NLP, Formal Grammar, and Python Programming. In the process of preparing the educational process, modern pedagogical technologies will be implemented, such as blended learning, and project based learning. During the implementation of the project, an analysis of the needs of the labor market was conducted, during which more than 50 companies were interviewed. Tools are being developed for processing the Kazakh and Uzbek languages that will be used in organizing and conducting the educational process for this educational program. The impact of this educational program both at the individual level, at the level of teachers involved in the implementation of this educational program, and at the institutional level is very significant.

Keywords: Erasmus+ project, Interdisciplinary educational program, computational linguistics, master’s degree, Central Asia universities.

1 INTRODUCTION

Computational Linguistics, and more generally Human Language Technologies, is crucial for the development of sophisticated information technologies, absolutely necessary for industrial, social and civilizational progress. These technologies are in progress almost everywhere in the world but at different speeds in the countries concerned. Although there is a lot of software for world languages

such as English, Mandarin, Russian, and Spanish, the Kazakh and Uzbek languages are still under-represented. Thus it is practically impossible for the world community to retrieve any information in Uzbek or Kazakh. Though the policies of both countries, Uzbekistan and Kazakhstan, aim at mastering English as the way of entering the world economy,[1] English proficiency is still low. The need for access to the world's information, as well as the need to do so in the UZ and KZ languages, are still very high. Besides, the world interest in information on the Turkic languages is growing and the demand for quality machine translation into\from Kazakh and Uzbek is evident. Although Google started this service about two years ago, it still needs much improvement. The Computational Linguistics master's program will improve the situation by training students and stimulating research in computational linguistics. The objectives are to understand the nature of linguistic representations and linguistic knowledge, and how linguistic knowledge is acquired and deployed in the production and comprehension of language. These objectives include the computing of the relation between form and meaning, the facilitation of document processing and information retrieval, and the provision of new resources and tools for learning languages (Kazakh, Uzbek, English). Computational Linguistics as an interdisciplinary educational program did not exist previously in the CA partner countries, though as a research area it has been developed since 2000.

The project "Development of the interdisciplinary master's program in Computational Linguistics at Central Asian universities / CLASS" began on 10/15/2017 [1]. Section 2 presents the methodology behind the preparation of the new interdisciplinary master's program in computational linguistics the novelty of which lies in a systematic approach; Section 3 describes analyses of existing educational programs in computational linguistics; Section 4 presents the results of a needs analysis for the specialty Computational Linguistics; Section 5 describes the development of a new educational program in Computational Linguistics for the universities in Kazakhstan and Uzbekistan; section 6 describes a particularity of teaching technology in the new interdisciplinary master program; Section 7 describes the development of NLP tools for the Kazakh and Uzbek languages with in the framework of the project; Section 8 provides conclusions and suggestions for further work.

2 THE METHODOLOGY BEHIND THE PREPARATION OF THE NEW INTERDISCIPLINARY MASTER'S PROGRAM IN COMPUTATIONAL LINGUISTICS

The methodology used in preparing a new interdisciplinary master's program in computational linguistics was systematic and included the following steps:

1 Analysis of existing educational programs in computational linguistics;

2 Needs analysis for an interdisciplinary educational program in computational linguistics in the CA partner countries;

3 Development of the content of a new interdisciplinary educational program in computational linguistics;

4 Development of training technology for a new interdisciplinary educational program in computational linguistics;

5 Development of tools to support practical training in the new interdisciplinary educational program in computational linguistics for the Kazakh and Uzbek languages.

Features of the development of the interdisciplinary educational program in computational linguistics take into account existing educational programs in computational linguistics in the world, the needs of the specialty in computational linguistics in the CA partner countries using specially designed survey forms, effective modern teaching technologies such as blended learning and project-based learning, and the need for language resources and tools during the educational process in computational linguistics.

[1] See: Kazakhstani Cultural program «Trinity of languages», the decree of the President of Uzbekistan № 1875 "On measures to further improve of foreign language learning system", December, 10, 2012.

3 ANALYSES OF EXISTING EDUCATIONAL PROGRAMS IN COMPUTATIONAL LINGUISTICS

Within Work Package 1 of the program CLASS we were supposed to define the needs and competency of alumni taking as a model the universities recognized as leading in the world in teaching Computational Linguistics. This was one of the initial steps of the project with the intention of serving further development as a reference.

We analyzed the methodology and approaches of the teaching and the contents of computational linguistics master's programs in order to articulate recommendations for developing curricula and syllabuses for the master's program. We presented two reports: the analytical study [2], and risk analysis [3].

On the basis of accessible Internet sources, we analysed 9 examples of MA/MSc programs from Europe (7) and the USA (2). These programs have been elaborated/implemented in the following countries Australia, China, Czech Rep., France, Germany, Italy, Malta, Netherlands, Spain, Sweden, UK, and the USA. In 8 (out of 9) cases, programs have been designed and implemented by the universities concerned. The only exception is the case of the "European Master's Program – Language and Communication Technologies" co-funded by the Erasmus+ Programme and involving a consortium of 9 universities from Australia, China, Czech Rep., France, Germany, Italy, Malta, Netherlands, and Spain. We focused on aspects that are crucial for the CLASS project. We considered it appropriate to take into account both formal and substantive features of curricula such as the degree awarded, duration of studies, teaching language or languages, requirements concerning students' profile, recruitment prerequisites, curriculum structure (courses, projects, exams), students' mobility and flexibility issues, and – last but not least – tuition fees.

In the analytical part, we summarized our observations concerning the typical structure of CL Master's Studies – usually composed of three modules: common module of mandatory courses with a relatively rigid structure, the more flexible module of elective courses, and the third one, composed of the obligatory items imposed by the university concerned in order to evaluate the master's candidates' acquired skills and expertise. For all three modules, we presented our findings in the form of lists of particular activities, typically identified by course names or types of evaluation measures.

The analytical part of this study concluded with the general recommendation for further work within the project. Our final recommendation formula takes into account specific requirements of the CLASS project already articulated in the project proposal document and discussed at the Kick-off-meeting of the project confronted with observations made while analyzing CL master programs. In particular, our recommendation concerning objectives postulates the formation of well trained professional staff prepared to undertake a collective work in the creation of language technologies (both software and language resources) within professional teams of language technology developers (supervised by NLP and HLT experts); preparation for a professional career in the sector of language industries. Alumni will be expected to constitute the core of future highly competent human resources to launch the language industry for the languages concerned. Recommendations were illustrated by a curriculum proposal (not binding at that stage).

Analysis of several master programs in Computational Linguistics drew our attention to possible project risks. We have identified and described three categories of risks: first those related to disposable language resources and tools, second those related to the need for highly qualified staff prepared to serve the huge territory of Kazakhstan and Uzbekistan, third those related to student recruitment and mobility. The risks that we have identified appear higher than for highly digitalized countries and must be taken into consideration seriously. In the report [3] we have proposed counter-measures for all three cases.

4 NEEDS ANALYSES

To analyze the needs of the specialty Computational Linguistics, it was necessary to conduct a survey of employer companies in various fields, since Artificial Intelligence and text analysis are actively introduced in almost all areas of life. To conduct a needs analysis it became necessary to develop a special survey. The questionnaire included the following questions:

1 Where, in your opinion, can specialists in the field of Computational Linguistics work?

2 Where, in your opinion, is the greatest need for specialists in Computational Linguistics?

3 Specify the types of activities, the main tasks that these specialists should know and implement?

4 Indicate what problems you encounter in your enterprise regarding Computational Linguistics?

5 Indicate what tools (tools) you use in your enterprise for information processing (software and materials)?

6 In your opinion, what general competencies should a specialist in Computational Linguistics have?

7 In your opinion, what specific competencies should a specialist in Computational Linguistics have?

8 In your opinion, what subjects will allow these specific competences to be obtained?

9 In your opinion, what forms and methods of instruction are preferred for the acquisition of these specific competences?

10 As an expert, do you consider the need for professional training in Computational Linguistics?

11 Are you personally interested in obtaining professional training in Computational Linguistics?

12 Are you interested in taking interns of the specialty «Computational Linguistics»?

For each question, possible answers were offered.

A survey of more than 50 companies is presented below.

1 Where, in your opinion, can specialists in the field of Computational Linguistics work? 57.1% - Publishing companies; 71.4% - Telecommunications; 75% - educational institutions; 42.9% - State and administrative; 28.6 % - Small and medium-sized production service companies; 32.61% - In all companies where there are computers; 75% - Research laboratories; 28.6% - Public non-governmental organizations; 17.9% - Other

2 Where, in your opinion, is the greatest need for specialists in Computational Linguistics? 60.7% - Telecommunications; 46.4% - Publishing companies; 14.3% - Banks; 7.1% - Industrial enterprises; 42.9% - State administrative; 17.9% - Small and medium-sized production service companies; 25% - Small and medium enterprises manufacturing products. 35.7% - Large companies; 28.6% - In all companies where there are computers; 75% - Research institutes, universities; 7.1% - Other

3 Specify the types of activities, the main tasks that these specialists should know and implement? 39.3% - Design; 28.6% - Production and technology; 32.1% - Organizational and management; 75% - Research; 64.3% - Innovative; 50% - Service-operating; 21.4% - Other

4 Indicate what problems you encounter in your enterprise regarding Computational Linguistics? 25% - The problem of confidentiality of information; 14.3% - The problem of changing data; 17.9% - The problem of the replacement of personal data; 28.6% - Ignorance of the modern electronic language environment; 46.4% - Inability to navigate in computer tools of linguistic analysis and in Computational Linguistics; 35.7% - Ignorance of the main approaches to digital modeling of humanitarian data in the field of history and literature; 50% - Inability to design specialized linguistic databases; 35.7% - Ignorance of various methods of mathematical generalization of the results of linguistic research; 50% - Inability to conduct processing of linguistic data by modern means; 35.7% - Ignorance of the principles of constructing various linguistic resources, including the corpus of texts; 42.9% - Inability to use existing tools and linguistic resources for product development.

5 Indicate what tools (tools) you use in your enterprise for information processing (software and materials)? 60.7% - Packages of applied programs; 82.1% - Information retrieval systems; 39.3% - Software for character processing; 50% - Office programs; 10.7% - Other.

6 In your opinion, what general competencies should a specialist in Computational Linguistics have? 53.6% - Creativity; 39.3% - Efficiency; 28.6 % - Care for quality; 35.76% - Work in a team; 39.3% - the ability to perform and develop their intellectual and general cultural level; 42.9% - The ability to perceive mathematical, natural science, socioeconomic and professional; 42.9% - Ability to develop and study techniques for analysis, synthesis, optimization and forecasting; 25% - Ability to simulate processes and objects based on standard automation packages; 17.9% - Initiative and ability to take necessary actions; 46.4% -critical thinking; 39.3% - Ability to analyze and draw conclusions; 28.6% - Ability to communicate with a diverse

audience; 10.7% - Confidentiality; 21.4% - Project design and management; 35.7% - Ability to work in an international context; 3.6% - Other.

7 In your opinion, what specific competencies should a specialist in Computational Linguistics have? 75% - Ability to apply existing tools and linguistic resources for development; 78.6% - Knowledge of basic levels of analysis and synthesis of text in natural language; 71.4% - Knowledge of the principles of the construction of various linguistic resources, including the corpus of texts; 39.3% - Understanding the essential differences of natural languages from the artificial and the feature of Computational Linguistics; 60.7% - Ability to conduct processing of linguistic data by modern means; 50% - Knowledge of the understanding of the algorithms of primary processes for automatic processing of text and speech; 39.3% - Ability to understand which language tools are behind a particular local task; 42.9% - Ability to use statistical methods of analyzing language data and visualization tools; 50% - Knowledge of various methods of mathematical communication of the results of linguistic research; 35.7% - Ability to design specialized linguistic databases; 39.3% - The ability to correctly use the results of mathematical generalization and use the results obtained; 25% - Knowledge of basic approaches to digital modeling of humanitarian data in the field of history and literature; 32.1% - Ability to navigate in computer tools of linguistic analysis; 21.4% - Ability to evaluate the complexity of different solutions and the thresholds of acceptable solutions; 28.6% - Ability to program prototypes and decision models; 21.4% - Ability to design a chain of processing of language data and interpret results; 32.1% - Knowledge of database and service management systems (SQL Server); 39.3% - Ability to understand the current state of Computational Linguistics and information; 39.3% - Ability to conduct independent research and obtain new scientific results in the field of Computational Linguistics; 39.3% - Work in interdisciplinary work, ability to interact with experts in other subjects; 21.4% - Solving a wide range of known theoretical and practical problems of computer facilities.

8 In your opinion, what subjects will allow these specific competencies to be obtained? 64.3% - English language; 46.4% - Kazakh language; 60.7% - Digital signal processing; 64.3% - Automatic text recognition; 28.6% - Recognition of announcers; 64.3% - Syntactic analysis of texts; 71.4% - Morphological analysis of texts; 57.1% - Methods of transcription of sounds and speech; 60.7% - Methods of speech synthesis; 50% - Specialized linguistic databases; 67.9% - Intelligent data analysis 46.4% - Methods for processing the text corpus; 35.7% - Methods of processing the audio case; 32.1% - Programming in Python 39.3% - Application packages for processing speech signals; 53.6% - Statistical methods of natural language processing; 46.4% - Professional internship; 3.6% - Programming in the Java language.

9 In your opinion, what forms and methods of instruction are preferred for the acquisition of these specific competences? 60.7% - Theoretical; 67.9% - Case studies; 53.6% - Technical; 78.6% - Practical internships; 35.7% - Group work; 28.6% - Digital lessons; 3.6% - Other

10 As an expert, do you consider the need for professional training in Computational Linguistics? If so, what are your needs? Specialization in linguistics makes it easier to understand the principles of constructing and using formal controls for computer systems. For example, it simplifies the understanding of the paradigms of different programming languages, the transition between them and rapid adaptation; if necessary, the creation of their own language components with specialized syntax. Specialists in this field (with an applied bias in programming and system architecture development) are very much in demand. Syntactic and morphological analysis of the text, basics of data mining.

11 Are you personally interested in obtaining professional training in Computational Linguistics? If so, what would you like to learn? In what field? Development of programs for speech recognition; Phonetics of the language; In the field of building a computer model of natural language (Kazakh); Digitalization of speech programming languages; IT Text recognizing; In the field of education.

12 Are you interested in taking interns of the specialty «Computational Linguistics»? If yes, what type of internship? 35.7% - Research; 71.4% - Practical internship; 14.3% - Functional internship with subsequent employment; 7.1% - None.

If so, for what period would you accept the trainees? 42.9% - from 1 to 2 weeks; 28.6% - 1 month; 14.3% - 3 months; 7.1% - more than 3 months; 7.1% by mutual agreement; 7.1% - No.

The survey results showed that the need for this educational program is present in almost all areas of companies' activity.

5 DEVELOPMENT OF NEW INTERDISCIPLINARY MASTER PROGRAM IN COMPUTATIONAL LINGUISTICS IN THE CENTRAL ASIA UNIVERSITIES

The analysis of international CL programs described in Section 3 drew attention to the need to take the realities of the CA universities into account when developing the educational program for the CLASS Master's. The other important aspect was to consider the future employment of the graduates within the academic and research environment as well as in public and private institutions. CL may be key to sophisticated language processing and AI, but there is a need to develop better language resources and tools for Kazakh and Uzbek before such ambitions can be realized. Despite the ambitions of the largely computer science academics involved, the inclusion of departments of linguistics and languages would not only provide valuable theoretical and practical input on the CA languages for language resources and tools, it should also provide language graduates with the training and tools in areas like document and terminology management and translation technology. A realistic description of the variety of qualifications offered by the degree for future employment was essential to the planning of the educational program.

Once the background and future of the CLASS Master's degree were established, the next step was to develop a core curriculum and syllabuses that would allow for the conditioning factors described in Section 3, and provide the infrastructure and planning for future developments. The curriculum defined in the meeting in Tashkent in October 2018 focused on the more general obligatory elements essential to any Master's program, while also providing a longer list of elective subjects. Given the restrictions on the number of ECTS available for the specific CL elements in the program after taking the national requirements for Master's degrees in Kazakhstan and Uzbekistan into consideration, however, it became clear that the role of elective subjects would be minimal. The next step, therefore, was to provide well-coordinated syllabuses for the obligatory subjects. The elective subjects will need to be developed later, either as the Master's program matures or as part of future PhD courses.

The EU colleagues will remember the difficulties universities went through when the 'Bologna Process' demanded carefully planned syllabuses, which defined 'objectives', 'outcomes', and 'competences', and provided clear indications of the teaching/learning methodology, evaluation procedure, and academic honesty expected. Academics tend to believe that the need for their area of expertise is self-evident, and do not always question the validity of their own programs. They are also sometimes unaware of what is planned for other subjects in the Curriculum, and coordination at a more general level is needed, if repetition and overlapping are to be avoided. It should have come as no surprise, therefore, that the CA colleagues would have similar difficulties, particularly when required to write these syllabuses in English, a language of which most of them had a passive knowledge, but less experience in writing.

One must also recognize the fact that, since this Master's course is such a new initiative, some of the universities will need to provide training in teaching certain subjects for existing teachers, or contract others with the necessary expertise. The CLASS project expects English to be used in the documentation required, and in the teaching and research projects, and this will be a problem for many of the teachers involved. There are also other obstacles, such as evaluation through individual and group project work, which may need to be overcome. All this will require understanding and help from their administrative offices, not to mention 'the system', and these are not always forthcoming.

However, despite all these difficulties, we can report that, although a steep learning curve was required of both the EU and CA colleagues, and the work has taken rather longer than originally planned, the results have been generally positive and we can look forward to seeing the Master's curriculum and syllabuses being implemented in the near future. The curriculum itself, as well as the qualification description and the syllabuses that have been developed so far are available at http://erasmus-class.eu/out-course-curricula/.

6 TEACHING TECHNOLOGY IN A NEW INTERDISCIPLINARY MASTER'S PROGRAM (BLENDED LEARNING, PROJECT-BASED LEARNING)

The other aspect of the program, established by the CLASS project, is the obligation to encourage and provide technical access to, 'blended learning', which places a strong emphasis on the effort needed for students to learn autonomously, rather than rely largely on teachers in the traditional classroom. This can be helpful in any coordinated effort to establish degrees involving a variety of universities, as it facilitates sharing efforts rather than duplicating them, and is particularly important in the context of the CA universities, not all of whom have staff with expertise on every subject in the curriculum. The

exchange of teachers is probably only viable in certain cases, as, for example, between the universities in Tashkent. However, distance can be overcome by facilitating teaching between universities using several forms of communication technology available.

A key technological asset for this purpose is learning management systems (LMSs), such as Moodle and Blackboard, now common in many universities. These systems are comprehensive platforms for blended learning that facilitate teacher<>student communication (in the form of messaging systems and forums), autonomous learning (by supporting multimedia content, interactive content and web links), and the general management, assessment and tracking of a course involving students in different universities (by online tests and attendance management).

For the purpose of the project, we decided to settle on Moodle as our reference LMS, as it is popular, feature-complete, familiar to the European project partners, and crucially, open-source and free to use. While it is arguably more difficult to set up than the well-known alternative Blackboard, the paid nature of the latter would be problematic, especially in the context of a project with multiple universities involved which would need to purchase the system.

As most of the staff involved in the CA universities had no or little previous contact with blended learning, LMSs and Moodle, several sessions were provided on these subjects during the training at Universidade da Coruña that was held from 18th June to 3rd July 2018, with the participation of 33 CA colleagues that will in turn train staff in their home institutions. The sessions covered both specific materials about Moodle 3.0 (general use and administration, activities, resources, grading, tutoring tools, etc.) and more generic knowledge about virtual teaching (pedagogical bases of e-learning, instructional design models, virtual course planning and mentoring). It is also worth noting that the training sessions themselves (both those about blended learning and other subjects relevant to the project) used blended learning and were managed and imparted using Moodle, a further way to improve the familiarity with the system and methodology.

Beyond technology and methodology, a challenge that the CA institutions will face when implementing blended courses for the CLASS Master's is the choice of language. It would be desirable for Kazakh students to benefit from courses based in Uzbek institutions, and vice versa. For this purpose, either all the contents have to be translated from Uzbek to Kazakh and vice versa, or a common language needs to be used. While English was proposed as such a language in early discussions, Russian may be more accessible for the teachers and prospective students.

Apart from the more or less structured and curated content that can be provided in a blended course using an LMS, students are probably as aware of the teachers of the availability of information over the Internet and through other digital sources. The role of the teacher is increasingly that of the expert who facilitates access to information that is reliable and encourages students to evaluate and use this information properly.

It is in order to learn more about the use of technology in education that we are so anxious to learn from the experience of others during the INTED 2020 conference.

7 DEVELOPMENT NLP TOOLS FOR THE KAZAKH AND UZBEK LANGUAGES IN FRAME OF THE PROJECT

One of the key project objectives is developing Natural Language Processing (NLP) tools for the Kazakh and Uzbek languages. The developed NLP tools will serve academic as well as research purposes and the needs of stakeholders. The number of NLP tools is growing every day, but it is hardly possible to find open NLP tools that serve the Kazakh and Uzbek languages. CLASS project provides for the better quality of the designed tools through staff training in NLP design organized by AMU (Poland). By the outcome of educational curriculum in the partners' university, capabilities and training seven Central Asian partners initiated designing of NLP tools: Morphological analyzer tool for the Uzbek language (Urgench State University); Morphological analyzer tool for the Kazakh language (Urgench State University); the Ontology tools for the Uzbek language (Tashkent State University of Uzbek language and literature); the Kazakh Language Processing Ontology Tools and Speech Recognition tools (L. Gumilev Eurasian National University); Monolingual Kazakh language corpus (Al-Farabi Kazakh National University); Kazakh language parallel bilingual corpora (Al-Farabi Kazakh National University).

Analysis of available software products for this profile showed that currently there are no text processing tools for text analysis and there is still no adaptation of existing NLP tools for working with

the Kazakh and Uzbek languages. The morphological analyzer NLP tool for the Uzbek and Kazakh languages analyzes a given text and generates morphological information, such as gender, number, case, and so on, as an output.

Collecting tool of the Monolingual Kazakh language corpus. This tool works as a complex of several processes: a collection of text data with meta-information. Data collection from official state portals: news and comments. The tool is implemented using Spring Boot technology. The application.properties file plays a key role. To change global configuration data, it needs to put application.properties with the new data next to the executable file. Data is stored in the NoSQL database MongoDB. Data preprocessing: tokenization, clearing of stop words, morphological analysis. Text with morphological markup. Indexing data using Elasticsearch to quickly search by parameters and create an API for queries. The user interface of the web application to interact with the Elasticsearch API. The tokenization algorithm is implemented using regular expressions. Tool for removing stop words. The dictionary of stop words is compiled. The list of stop words can be expanded, depending on the pre-processing conditions and the initial type of text. Lemmatization tool. Based on the Porter algorithm, according to the grammatical rules of Kazakh, affixes are sequentially cut off from the end of the word: that is how stemming is performed. Furthermore, normalization is performed on the basis of results of stemming; synthesis of normal form is carried out. Morphological analyzer tool. It produces an analysis of the word.

Bilingual NLP tools collect URLs of pages in the language in which the most news is published. The tool collects resources from a list of URLs in .xml format. Each file is saved by the corresponding numbering and the corresponding markup of the language, for example, 1000.kk, 1000.en. The file consists of a section, article title, and publication date and article text. The data cleaning tool clears and replaces erroneous characters, such as Greek letters, incorrectly opened brackets, removing extra spaces, etc.

Sentence splitting tools break down sentences into lines, takes into account that the sentences may consist of contraction, abbreviations, quantitative listing as a list. Sacremoses tool is used to normalize punctuation and tokenization. Normalization adds spaces between punctuation. Tokenization is another key concept. This includes taking the sentence and breaking it into the smallest separate parts. For example, “Kassym-Jomart Tokayev held a meeting with the founder of Alibaba Group Jack Ma.”. The sentence will be divided into the following tokens: “Kassym-Jomart”, “Tokayev”, “held”, “a”, “meeting”, “with”, “the”, “founder”, “of”, “Alibaba”, “Group”, “Jack” “Ma”. Bilingual Frequency Lexicon is exploited to form a bilingual dictionary. The frequency of the word in the source part with the frequency of occurrence of the translation in the target part is taken into account. For example, the word “state” occurs 27 times, it may have translations like “мемлекет(memleket)”, “жағдай(zhagdai)”, “үкімет(ukimet)”. Translation as “үкімет” is more common than the others, therefore it will be added to the frequency vocabulary. Adapted Hunalign tool is for aligning source texts with the target language. Using the above tools, Hunalign has been adapted to align parallel bilingual corpora. Morphological segmentation tools: bpe - split into frequent segments, does not take grammar into account; segmentation based on ending systems: takes into account the rules of word formation based on types of endings of the Kazakh language.

Natural Languages Processing Ontology Tools (TSUULL; L.Gumilev ENU) are used for formal and specialized concepts and relations that belong to the exact domain. Having the advantage of ontology in NPL to create metalanguage in the sphere of machine translation (mainly, rule-based machine translation) or other purposes (information retrieval system, text analysis, annotation of text). Thanks to ontology, creating the structure of information based on systematical and hierarchical data aids to ease the computational processing of the natural language. Effective way to create ontology is representing OWL by means of Protégé software. What is interesting to note is that ontological as ready devices like other sketch engines over and beyond investigation based in order to build data structure in different spheres. As similarities and distinctions of grammatical features between Uzbek and Kazakh languages compared with ontological models, it can ease to classify subclasses of relations of the words by different syntagmatic properties. In our point of view, for agglutinative languages, morphology is strong side for all-natural language processing by computer. We used the editor Protégé (http://protege.stanford.edu) to input grammatical classes and relations of Turkic languages (Uzbek, Kazakh, Tatar, Kyrgyz, and Turkish) into the ontological framework. It is a free open source ontology editor and a framework in order to generate knowledge bases so far. The ontological model of the Uzbek parts of speech allows us to work easier with programming languages like Java. It includes the morphological rules and the relationships of categories as a hierarchic system. Ontology gives to the opportunity to map all chain with relations, properties, subclasses also.

The semantics of ontological models in Excel format and their structure were analyzed in the Uzbek language in independent and useful word genres. The morphological categories of the Uzbek language include subcategories, explanations, grammatical questions, and related examples.

The next tool is a morphological analyzer by FST technology. According to two types of script Uzbek (Latin and Cyrillic), FST technology is a handful to create an analyzer. Ubuntu using FST technology to represent morphological features of the Uzbek language is convenient for lemmatization and tokenization of the text. In this case, the formal morphology of Uzbek language was written in both graphemes: alphabet, sets, definition, rules, and lexicon.

In the lexicon, we included POS (40000 units) in Uzbek both graphemes Latin and Cyrillic. Grammar rules are also adjusted according to alphabet requests. In spite of different files in data, we had to unify them into a single morphological analyzer. Lexicon and grammatical rules were the very important points in text analysis because some orthographical rules and letters did not match each other. Consequently, we came to the conclusion to adjust in some content by converting letters to Cyrillic.

Ambiguities are corrected slightly by hand in each level of analyzing text. In perspective, our tool is to analyze the semantic and pragmatic levels after correcting words in the texts. A special program intersecting composition was developed in order to facilitate the combining of the lexicon transducer and the two-level rule transducers (TWOLC-two-level compiler) and to avoid excessively large intermediate results.

8 CONCLUSIONS

The article describes the current development experience of the Erasmus + project “Development of the Interdisciplinary Master's Program in Computational Linguistics at the Universities of Central Asia (CLASS)” No. 588545-EPP-1-2017-1-ES -EPPKA2-CBHE-JP-CLASS in seven universities in Central Asia. The novelty of the work presented is a systematic review of the project development process, which is relevant not only to the partner countries participating in the project but also to the field of Computational Linguistics, which is at the cutting edge of artificial intelligence. The systematic approach used in the project included the stages of analysis of existing educational programs in computational linguistics, an analysis of the need for computational linguistics in the CA partner countries, the development of the contents of the educational program of Computational Linguistics, the development of student teaching methodology using modern teaching technologies, such as blended learning and project training, the development of NLP tools to support the practical training of the educational process of the program in Computational Linguistics.

At the level of individual participants and team members, the project provides an opportunity for master’s students to follow an interdisciplinary educational program, thereby expanding their field of activity; by providing mobility it allows students to gain experience in communication and learning in a new academic environment, with new approaches to the learning process. The project provides an opportunity for team members to improve their qualifications in a new, very promising, and developing field of computational linguistics, which is at the intersection of artificial intelligence, machine learning, linguistics, and big data. The project provides an opportunity for project participants to expand collaborative ties with foreign partners from European and Central Asian universities. The students of this educational program will have the opportunity of scientific internships at partner universities for this project.

At an institutional level, the project offers an opportunity to develop a new area of study in Computational Linguistics, to improve the skills of the personnel involved in the project, and to establish and strengthen international relations.

At the national level, the project opens up the possibility of creating and developing resources and tools for processing natural languages, especially the CA national languages. The project will also become an important component of the “Digital Kazakhstan” program, especially in terms of training the teaching staff in the field of computational linguistics, artificial intelligence, machine learning, and big data.

The results of the project actively influence the development of relationships with companies in the fields of natural language processing, large text data, and sentiment analysis. So, KazNU has established relations with Alem Research company, the Transition Group Company.

Project CLASS development duration is 15.10.2017-14.10.2020. Within the framework of the CLASS project, the development of the following tools for the Kazakh and Uzbek languages is foreseen:

- tools for creating an acoustic enclosure for speech recognition based on a transcription of audio files;
- text graphemic analysis tools;
- tools for creating ontological models of morphological rules, allowing to formalize the structure of words taking into account the semantics of morphemes and automate morphological analysis and synthesis of words;
- tools for creating ontological models of syntactic rules, which allow formalizing the sentence structure taking into account the semantics of sentence members and automating parsing and synthesis of sentences.

ACKNOWLEDGEMENTS

The project “Development of the interdisciplinary master program on Computational Linguistics at Central Asian universities. Number 585845-EPP-1-2017-1-ES-EPPKA2-CBHE-JP (15/10/2017 to 14/10/2020)” funded by the Education, Audiovisual and Culture Executive Agency of European Union.

REFERENCES

[1] Development of the interdisciplinary master program on Computational Linguistics at Central Asian universities. Number 585845-EPP-1-2017-1-ES-EPPKA2-CBHE-JP (15/10/2017 to 14/10/2020) *(unpublished)*

[2] Z. Vetulani, M. Kubis, J. Marciniak, “Analysis of international master programs by AMU with recommendations“, *CLASS Project Report (unpublished)*, March,2018.

[3] Z. Vetulani, “CLASS Master Program Implementation Risks“,*CLASS Project Report (unpublished)*, March,2018.