Abstract

The paper presents complex teaching approach used for the course "Corpus Linguistics" which is a part of masters study program in Computational Linguistics currently tough at Sofia University "St. Kliment Ohridsky". We propose a useful combination of theoretical issues, exercises and problem-solving tasks, and practical course work as a successful teaching strategy with multilingual application. The course uses extensively e-Learning tools for searching electronic text corpora. It is based on “learning by doing” methodology allowing students to discover linguistic phenomena and interpret them semantically. The final course work improves students’ ability to do independent research work by fostering their curiosity, discovery and developing their ability for reasoning.

© 2014 Elsevier Ltd. This is an open access article under the CC BY-NC-ND license.

Peer-review under responsibility of the Organizing Committee of CY-ICER 2014.

Keywords: Corpus Linguistics, Language Learning and Teaching, e-Learning, Technology-based Learning;

1. Introduction

The traditional view on Corpus Linguistics is that it is a language-dependent area which can be tough with respect to specific language. However, there are very many general language features and some standards related to corpus research with multilingual applications. The results of teaching approach we used (a successful combination of theoretical knowledge, problem-solving tasks and practical independent research) reveal that students have performed standardized problem-solving tasks and achieved good results doing work for several languages.

First of all, the final results were very successful in discovering linguistic phenomena and their further interpretation (both formal and semantic). The reason for that is general distribution and optimal combination of courses included in the program which are offered by both the Faculty of Slavic Philology and the Faculty of

* Corresponding name: Velislava Stoykova. Tel.:+ 359 2 979 2953; fax: +359 2 970 23 02.
* E-mail address: vstoykova@yahoo.com
Mathematics and Informatics. Also, the Computational Linguistics masters program is open for students from bachelors programs in Linguistics, Mathematics, and Computer Science bachelors. The program also uses different ICT applications and e-Learning tools for searching on-line electronic text corpora and offer program-specific exercises and problem-solving tasks oriented to multilingual applications.

Generally, the problem-solving tasks are oriented to resolve logical, phonological, morphological, grammar and lexical ambiguity. Problem-solving tasks were especially designed in line with theoretically transposed topics and are oriented towards practical applications and improvement of individual research ability and discovering. The educational approach used is related and indebted to that known as learning by doing (Aldrich, 2005) as it was used in teaching mathematics (Stoykova, 2009). The problem-solving tasks were created and developed in accordance with the general course outline.

2. Course outline - teaching content

The introduction of Corpus Linguistics course is included in the masters program in Computational Linguistics during its last semester after all introductory courses. The course is distributed into three combined parts - theoretical issues, exercises, and practical course work - all they are related during the teaching process. Students are allowed to be from different bachelors’ programs like different philology programs, mathematics, computer sciences, etc.

The theoretical content of the course is designed to consist of different topics in Corpus Linguistics including general issues of historical perspective, key concepts, techniques and methodologies for creation and working with electronic text corpora, etc. It presents in historical perspective the evolution and tradition in development electronic text corpora both as a grammar structures and as a lexical data base resources.

The theoretical topics also include description of principles and techniques for annotations which are illustrated with related exercises so to give students insight how related annotation schemes reflect search results. Different corpora search systems are described - both web-based and desktop. The corpora search systems are compared with respect to the principles of their design, to the query search language and to the limitations of the search procedures. That part of course outline deals both with universal and multilingual principles of semantic representations.

There are also theoretical topics which include principles, methodologies and techniques for design of electronic text corpora for different purposes. Several topics describe electronic text corpora classifications divided into different types according to various principles among which are mono- or multilingual, synchronic or diachronic, general-purpose or specialized, etc. The topics are supported by exercises which demonstrate both electronic text corpora of related type and methodology to create electronic text corpora for different purposes.

Special topics include presentation of electronic text corpora of Bulgarian language (generally Bulgarian National Corpus) – the history, creation, search systems and recent improvements.

3. Exercises and problem-solving tasks

The exercises and problem-solving tasks included in the course outline support both studied theoretical topics and prepare students to accomplish their independent research work. They are part of specific teaching course methodology aimed mostly at discovering semantic and logical connections between linguistic data, and their further theoretical and formal interpretations.

In general, exercises and problem-solving tasks are thematically related to a particular theoretical problem and support its acquisition by using electronic tools for different corpora. The exercises and problem-solving tasks are performed by using both available and newly established on-line educational resources - mostly electronic text corpora. The problem-solving tasks were especially designed to allow students to perform different standardized search procedures which illustrate related theoretical topic.

The exercises are focused mostly on interpretation of ambiguity - a phenomenon which is typical in working with electronic text corpora. They consist of on-line problem-solving tasks with multilingual application like disambiguation of language phenomena which express different semantics or grammar features (generally, identifying grammar category) or different lexical semantics (different meaning) (BNC, 2000).
Thus, problem-solving task demonstrates two basic approaches to electronic text corpora use - rule-based and statistical, however in many corpora search systems they can be combined. A useful example of that which is also with multilingual application is the use of the Sketch Engine (SE) software (Kilgarriff, Rychly, Smrz, & Tugwell, 2004) which allows both grammar disambiguation tasks and lexical semantic ambiguity tasks. Examples of grammar disambiguation tasks using SE (SE, 2013) software can be performed for different languages like for flective and for not flective languages. The search systems use Corpus Query Language (CQL) for encoding inflectional grammar features which present related grammar category.

The problem-solving tasks presented at Fig. 1 and Fig. 2 give opportunity to define lexical meanings of the word good both in English and in Bulgarian by supporting them with language usage examples and illustrating the use of electronic text corpora for lexicographic purposes by generating concordances of related word (Sinclair, 1991).

Lexical semantics disambiguation tasks are of various kinds including identification of terminological relations (Stoykova & Petkova, 2012). The tasks are based on the use of statistical approaches incorporated in the SE software (SE, 2013) based on estimation of different types of probabilities and focused on generation of concordances, collocations and co-occurrences of related word.
Fig. 2. The concordances of Bulgarian word *good* from Bulgarian National Corpus - SE version

The problem-solving tasks for generation of concordances of Bulgarian word for *good* with its inflected forms are given at Fig. 2 - examples from Bulgarian National Corpus (BulNC, 2010) - SE version (Koeva, Blagoeva, & Kolkovska, 2010). They illustrate examples for both grammar disambiguation tasks - *good* as a modifier and as a different feature.

The problem-solving tasks for generation of collocations and co-occurrences use statistically-based search (Oakes, 1998) and can help to exercise discovering and identification of more complex lexical semantic relations like synonymy, hyper- or hyponymy, etc. The SE software maintain a mechanism to present semantic features like hierarchy, typology, synonymy which is used for universal language-independent lexical semantic representation (Kilgarriff, Reddy, Pomikalek, & Pvs, 2010). Generally, it reveals semantic relations and can demonstrate electronic corpus-based technologies to produce dictionaries of synonyms, antonyms, collocations, etc. (Stoykova, 2013).

They also can be used to relate terms to a particular domain by building conceptual hierarchies or identify phrasal by generation collocations. For lower grade students, these tasks develop general ability for reasoning and similarity. Such problem-solving tasks are with multilingual applications since they relate the very general semantic phenomena existing in almost all languages. The search software uses statistically-based ICT approaches with multilingual applications (Frydryhova Klimova & Semradova, 2012).

4. Practical tasks

At the end of the course students are expected to perform independent research on a chosen topic by investigating it in already existed corpora or in designed by their own electronic corpora. The students’ achievements from final exam result in creation of several newly established students’ electronic text corpora in different European languages. They present original research insights interpreting various linguistic phenomena among which are
investigation of stylistic phenomena like language of specific author, social and sub-cultural language use as slangs or dialects.

Some students practical works resulted in a creation of newly established electronic text corpora, terminological resources in military domain using AntConc software (AntConc, 2013) and in the domain of journalism using EuroTermBank system (ETB, 2013).

5. Conclusion

The teaching methodology used for the course Corpus Linguistics which includes theoretical topics combined with exercises and problem-solving tasks was very successful. It gives an insight to students and state-of-the-art experience how to perform linguistic research and to develop ability for enquiry and discovery.

References