Scaffolding for CLIL in Computer Science Courses: Data Driven Learning Approach*

Xenia Piotrowska1 Tamara Alekseeva2
krp62@mail.ru ktv@gukit.ru

1Herzen State Pedagogical University of Russia,
2Saint-Petersburg State University of Film and Television
St. Petersburg, Russian Federation

Abstract

Prospects of organizing of Content and Language Integrated Learning scaffolding provided at the Herzen University of Russia, are under discussion. Scaffolding methodology, based on data-driven learning and well-known corpora tools as Sketch Engine, AntConc, Skell and LexTutor, is constructing for groups of Math and Computer Science students are discussed. First results, gathered through online post-course questionnaire that learners answered voluntarily, showed that, overall, students perceived the CLIL-scaffolding that was provided as rather effective, some shortcomings were identified.

Keywords: data-driven learning, corpora, content and language integrated learning, scaffolding, computer simulation, data-mining, concordansers, second language skills

Introduction

Archaic form of educational content in the field of second language learning for special purposes in Russian universities leads to traditional lags in the formation of second language skills. This content is very different from the current level of development of scientific and technical speech. From another point of view, it is a well-known fact that knowledge of more than one language is also found to be linked to enhanced learning, problem-solving and communication competencies – all of which are ought to be fundamental 21st-century transversal skills.

In this context, we will focus on the description of Content and Language Integrated Learning (CLIL) techniques based on well-known corpora technologies both concordansers and computer lingua didactic tools. We’ll try to show how, with the help of a comprehensive methodology based on scaffolding tools and Data-Driven Learning technologies, it is possible to support second language acquisition skills among students of mathematical and informational areas of training.

Similar approach, named as intellectual-cognitive learning in the field of computer lingua didactic, was proposed in Russia in the 1990s. This approach was based on the potential (data and programs) of machine translation systems [Piotrowska, 1991][Piotrowska, 2005]. Ideology of this approach was to provide students with the opportunity to independently extract language knowledge that meet individual cognitive needs. It was opposed to the traditional technology of “transferring of finished knowledge”. Today, with the development of corpus technologies employing digital support for didactic content, all opportunities have been created for implementing these ideas. Due to the

*Copyright © 2020 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).
practice of DDL, it is becoming possible to make the transition from simulating a foreign language situation in training lessons to organizing an almost authentic foreign language environment in a professional field.

This poses before bachelor’s, master’s and postgraduate training the task of the maintenance of foreign language communicative skills, not only in language classes in the professional field, but directly in profile classes (while working in laboratories with foreign language software in the field of mathematics, computer science, applied physics, chemistry, etc.) and in professional scientific researches (search of professional foreign language literature, writing scientific articles).

1 Materials and methods

1.1 Materials of investigations

This study investigates a model of CLIL-scaffolding and scholars’ perceptions of its effectiveness, implemented at the Herzen State University of Russia, during blended finding out of Applied Math and Computer Science in two courses on Computer simulation and Data-Mining. Manuals on Arena Simulation and Weka systems, with texts of actual scientific articles of the fields of computer simulation and data mining, became the didactic materials for CLIL-scaffolding and DDL.

1.2 Content and Language Integrated Learning

Content and Language Integrated Learning (CLIL) is an educational approach where some content learning (like a topic on Programming in R) is taught in an additional language (such as English language in Russia). This term was first coined in 1994 by David Marsh and Anne Maljers[Marsh, 2002]. In our days it has become one of the most significant and talked-about innovations in teaching English as a foreign language over the two decades and is now promoted by educationalists and institutions around the world.

C-Four dimensions (4 Cs) form a conceptual framework (Coyle 2008), which connects content, cognition, communication and culture[Coyle et. al, 2010]. Model of CLIL is shown on Figure 1.
In CLIL courses, for example, “Computer simulation” or “Data-mining,” should be taught to students in Russian, but programs, texts of manuals or articles they should read in English. Due to this students should not only learn about the subject in computer science, but they will also gain relevant vocabulary and language skills.

CLIL is closely connected with scaffolding methods, developing student’s learning strategies and modern corpus resources.

Scaffolding is crucial for decreasing learner’s demotivating, because it ensuring that learners feel more successful in subject and language tasks. It can take a wide range of forms – from choosing the language we use in class to breaking down tasks into smaller steps. Of particular use in CLIL classrooms are such language tools as word banks, glossaries and writing frames, on-line dictations which would not be necessary in a first or second language learning environment but here enable students to communicate in a meaningful way. Scaffolding ought to encourage students to use different learning strategies and find the ones that work for them. This means that most effective learners are aware of how they learn and think about which learning strategies they use for different tasks. They can use a wide range of learning strategies. But not effective students also need some time to reflect on new strategies after its using in order to enable to reformat its’ in the future for themselves. Therefore, scaffolding architecture should support various student learning strategies and have an expanded arsenal of innovative didactic tools.

Comparative analysis of results in [Yakaeva, 2017] displayed that subject teachers are absolutely responsible and deliberate in using language scaffolding in the process of teaching Mathematics through CLIL in University courses. In the process of study, methods of language scaffolding were revealed and analyzed on the foundation of assessment parameters of methods and techniques for teaching a foreign language. Authors conclude that, out of all methods and ways of language scaffolding, it is more popular for explanation of unknown mathematical terms.

### 1.3 Basis of Data Driven Learning

Data-Driven Learning is based on the idea that language is seen as data, and students as researchers performed to discover useful information in teacher-driven tasks. This pedagogical approach is based on the well-known paradigm “data - information - knowledge” according to the DKIW pyramid and the template approach to teaching grammar and vocabulary. Learning how to formulate language issues, how to use resources, how to obtain data and how to interpret them is just fundamental for the student’s autonomy and activity. While arriving at their own conclusions through these procedures, students use their higher-order thinking skills, according to Bloom’s taxonomy, for knowledge-creating.

In DDL, students can use the same corpus tools as professional linguists use, namely a corpus of texts that have been sampled and stored electronically, and a concordancers, that are search engines designed especially for linguistic analysis.


Some tools have been specially created for learning purpose, such as Micro-concord, WordSmith, WriteBetter, SkELL and so on. Although the practice of DDL more often concerns the

---

training of professional translators [Gorina, 2018] [Bernardini et al., 2008] or the further linguists [Gavrilova Kogan, 2016] [Cherniakova, 2011]. It can become even more valuable in the development of search and maintenance of foreign language skills in disciplines of the professional cycle in the science and engineering courses or computer science.

There are some examples of successful corpus-based scaffolding practices in physics CLIL-classes [Carloni, 2018] and an experience of DDL for language for special purpose in vocational schools, for example, a school for hairdressers [Corino & Ernesti, 2019].

2 General scheme for solving the problem

Having studied the DDL, CLIL, scaffolding theories and the most effective tools for organizing of DDL such as AntConcLab [Nymm et al., 2017], SketchEngine [Thomas, 2016 a], Skell [Thomas, 2016 b], LexSite-LexTutor [Berg et al., 2019] and LexTutor [Cobb, 2007], we developed an approximate algorithms of teacher’s actions for scaffolding in the educational work with the texts of manuals for computer programs (Fig. 2). In our opinion, the aforementioned corpus-based systems can be used by a teacher as a tool for compiling special training dictionaries and exercises that reflect the specifics of the vocabulary of the entire discipline or separate lessons.

![Figure 2: Corpus scaffolding for terminology training](image)

Despite the richness of observed tools, each system has drawbacks, therefore, as can be seen from its characteristics listed in Table 1, we should talk about creating an integrated scaffolding from the selected tools, taking into account its efficiency and teacher’s / student’s convenience.

Basing on the described tools, subject teacher can create online flashcards, online dictations, anagrams, crosswords and exercises based on the concordances of professional texts and word lists that he has built himself previously.

Further, we explain the examples that are obtained for the texts of the manual on the Weka data mining system for the course of Data-mining. LexTutor’s flashcard can be prepared by a teacher to
Figure 3: An example of LexTutor flashcards for basic terminology in Data-Mining course
Table 1: Corpus manager’s characteristics

<table>
<thead>
<tr>
<th>Corpus manager</th>
<th>User’s Corpora Link</th>
<th>User’s texts Link</th>
<th>Didactic support</th>
<th>Interface service</th>
</tr>
</thead>
<tbody>
<tr>
<td>AntConcLab</td>
<td>free</td>
<td>5</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>LexTutor</td>
<td>free</td>
<td>5</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>LexEngine</td>
<td>Commercial</td>
<td>5</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>SkELL</td>
<td>Commercial</td>
<td>5</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>LexSite-LexTutor</td>
<td>Commercial</td>
<td>5</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 4: An example of exercises on keywords drill formatted by LexTutor KeyWords instrument for Weka Manual.

work out the basic concepts of a data mining course in the second language (in English). Also, it is possible to use such cards only in native language and in a combined version, for example, the terms are in English, on the back of the card is the definition/translation in the native language. Figure 3 shows the preparation of LexTutor’s flashcards for the concepts of Chapter 5 of the Weka’s system guide on the theme “Explorer”. In the same system, it is possible to organize creative exercises to study the actualization of the lexical unit in a specialized text, based on a reserved or created by user corpus.

Besides, Dictator is a very useful and robust LexTutor’s tool, allowing us to operatively organize audio on-line dictations in training and test modes on different language constructions (words, sketches and sentences)(see Fig. 4). Due to Dictator we can organize scaffolding in listening, spelling and learning new terminology, as well as unfamiliar words. Moreover, the voice of the speaker can vary in speed and pronunciation (Special English / American English).

Now let consider the fragment of didactic work with SkELL using the word: cluster as an
Figure 5: Results obtained for keyword *cluster* by SkELL’s instruments: Examples and Word Sketches
example. Computer Science class student, usually gives a translation of the word *cluster* as a term: a group of similar things. To expand student’s horizons in both professional and linguistic fields, one’s can offer the following types of exercises for working with this word, as with a lexical unit. All sentences are taken from BNC with the help of SkELL.

1. Find some sentences where this word should be used in a direct meaning.
Examples of learner’s choice:
1) For arrays with clusters additional terms are needed.
2) Also content cluster performance indicators are available.

2. The word cluster despite meaning in Russian "кластер, скопление", can be used in meaning: “группироваться”, “гроздь”, “кисть”, “тесниться”, “связываться”. Find some examples with SkELL.
Examples of learner’s choice:
1) Hundred of rapidly moving galaxies often cluster tightly together.
2) The small town is clustered around the castle.
3. With the help of SkELL, search some clusters with this word. Try to classify this examples.
Examples of learner’s choice: Cluster sampling, robust clusters, cluster formation, cluster promotion.
4. Compose sentences where this word should be used in a direct meaning.
Example of learner’s choice: Cluster data describes data where many observations per unit are observed.

5. Compose sentences where this word should be used in additional meaning.
Example of learner’s choice: The vine is very productive and cluster thinning is required.

6. Make a sentences where the given word cluster will appear as a subject / addition / circumstance / predicate.
Examples of learner’s choice:
Modern clusters usually consist of nodes with multiple CPU’s or cores, sometimes “multi-threaded”...
2) Next I create a simpler data presentation, God bless Excel, by creating two big clusters next to each other.
3) Several different clustering methods are provided.
4) They proposed a clustering algorithm to measure dependencies.
5) The remaining six groups clustered approximately in chronological order.

7. Find and explain the connection between word cluster and words discover, instrument, analysis and so on.
Example of learner’s choice: In Search Engine technology, the mutual information between phrases and contexts is used as a feature for k-means clustering to discover semantic clusters (concepts).

8. Choose a number of synonyms for the word: cluster.
Example of learner’s choice: accumulation, congestion, collection, aggregation, gathering.
Thus, one can work on the vocabulary at the lesson, but also it can be included as independent homework of students. Corpus scaffolding by SkELL for exercises above are shown on Figure 5.

Finally, we dwell on the problems of organizing the search for relevant scientific literature and identifying terminology with the subsequent construction of exercises for working out professional terminology. We’ll demonstrate the solution of these research problems using AntLab’s programs.

For example, a student’s task is to select scientific texts for the formation of a literary review according to subject area keywords (known by himself or given by the teacher). Thus teacher’s task is to select the terminology and standard phrases for further exercises forming. On Figure 6 we show the scheme of AntLab programs sequential execution. By the AntCorGen program, we can search for texts in the PLOS ONE library across a wide range of areas of knowledge. Moreover, it is possible to organize the search separately for annotations and keywords, as well as for all text sections together. Using the AntFileConverter, EndCodeAnt and VariAnt service programs, it is necessary to carry out the pre-processing procedure with the selected texts.

Further, using the AntConc program we can solve the central part of our quantitative research.
Figure 6: Corpus scaffolding for terminology and grammar training with AntLab
Figure 7: Screenshots of AntConc and ProtAnt programs results for Arena Simulation manuals
Concordancer AntConc is very robust service for procedures of lemmatization, frequency dictionary organizing, keywords searching, building of collocations, N-Gramms and concordancers.

The updating of list of keywords phrases and terms by AntConc program, allows us to expand the set of studied texts. This set can be increased by using the ProtAnt program which selects texts that are closest in terms of lexical composition to the educational task. It takes a corpus of texts and compares them either individually or as a whole against a reference corpus or list of keywords to find characteristic features in the target files. After this step, it is necessary to return to quantitative processing and text profiling by AntConc and ProtAnt programs on expanded text corpus.

After this steps, it is necessary to return to work on quantitative research in the AntConc program, but with an expanded composition of educational texts, as well as to profiling the resulting dictionary and text markup. By the next step, it is necessary to repeat the quantitative processing and text profiling by AntConc and ProtAnt programs on expanded text corpus. In purpose to fragment texts into sentences AntMover program is used. It saves each example in a separate file. Next, the ProtAnt program again processes new sentences-files and identifies those that are lexically or grammatically suitable for generating exercises using a selected task designer program. Figure 7 show some results given by AntConc: Concordance and Collocates for for term “attribute”, and file prototyping by ProtAnt program for Arena Simulation manuals and list of keywords.

Because there is no didactic service in AntLab, to form exercises one can take such services as: The Teacher’s Coner, Quizlet or with on-line test makers: Hot Potatose or some others.

Conclusions

The first outcomes seem to be quite positive and it looks like DDL was very well received by bachelors from Math and Computer Science classes. The experiences presented in this paper suggest that DDL activities to CLIL-scaffolding have proven to have great benefits on students’ language skills development, along with motivation and involvement. When students were asked about the corpus-based activities, the majority of them declared that they found it useful and relatively easy to work with the concordancers and with authentic data. All of them gave appreciation for this new approach, they found it very useful to be able to work on their own or with the guidance of the teacher, but using the professional linguistic software to explore the language.

In our opinion, the proposed approach can lead to a synergistic increasing [Mayer, 2015] of students’ language skills development in the professional sphere. It can support foreign language competences formed at the previous stages of education, and will scaffold for applied math and computer science bachelors in software learning.

Acknowledgement

The research was supported by the Russian Science Foundation (RSF), Project “Digitalisation of the high school professional training in the context of education foresight 2035” (No 19-18-00108).
References


