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

Xenia Piotrowska[1]
krp62@mail.ru

Tamara Alekseeva[2]
ktv@gukit.ru

[1]Herzen State Pedagogical University of Russia,
[2] Saint-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

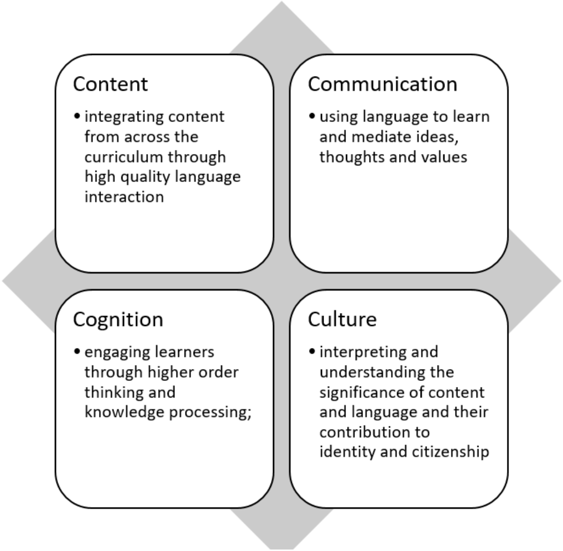

Content
• integrating content from across the curriculum through high quality language interaction
Communication
• using language to learn and mediate ideas, thoughts and values
Cognition
• engaging learners through higher order thinking and knowledge processing;
Culture
• interpreting and understanding the significance of content and language and their contribution to identity and citizenship

Figure 1: 4 Cs Conceptual Framework of CLIL

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.

Tim Jones coined the term data-driven learning in 1991 [Jons, 1991], but only since 2009 numerous books were followed: Anderson и Corbett (2009)[1], Reppen (2010)[2], Bennett (2010)[3], Flowerdew (2012)[4], Boulton и Tyne (2014)[5], Friginal (2018)[6].

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

[1] Anderson, W., Corbett, J. Exploring English with online corpora. London: Palgrave Macmillan, 2009.

[2] Reppen, R. Using corpora in the classroom. Cambridge: Cambridge University Press, 2010.

[3] Bennett, G. Using corpora in the language learning classroom: Corpus linguistics for teachers. Michigan: University of Michigan Press, 2010.

[4] Flowerdew, L. Corpora and language education. Basingstoke: Palgrave Macmillan, 2012.

[5] Boulton, A., H. Tyne. Des documents authentiques aux corpus: Démarches pour l’: To apprentissage des langues. Paris: Didier, 2014.

[6] Friginal, E. Corpus linguistics for English teachers: Tools, online resources, and classroom activities. Abingdon/New York: Routledge, 2018.

training of professional translators [Gorina, 2018] [Bernardini et al., 2008] or the futher 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.

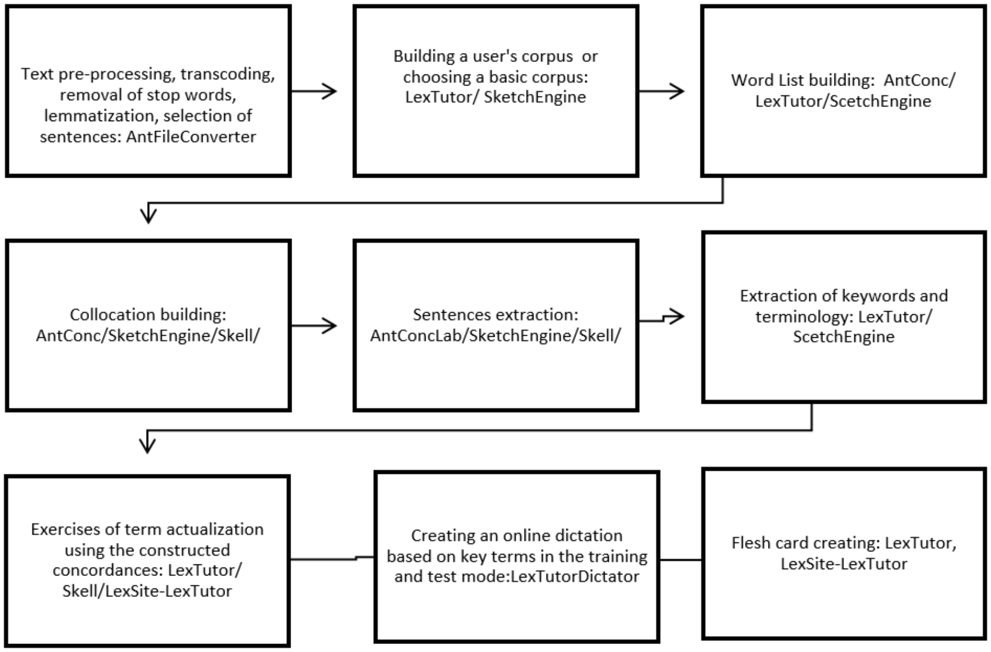
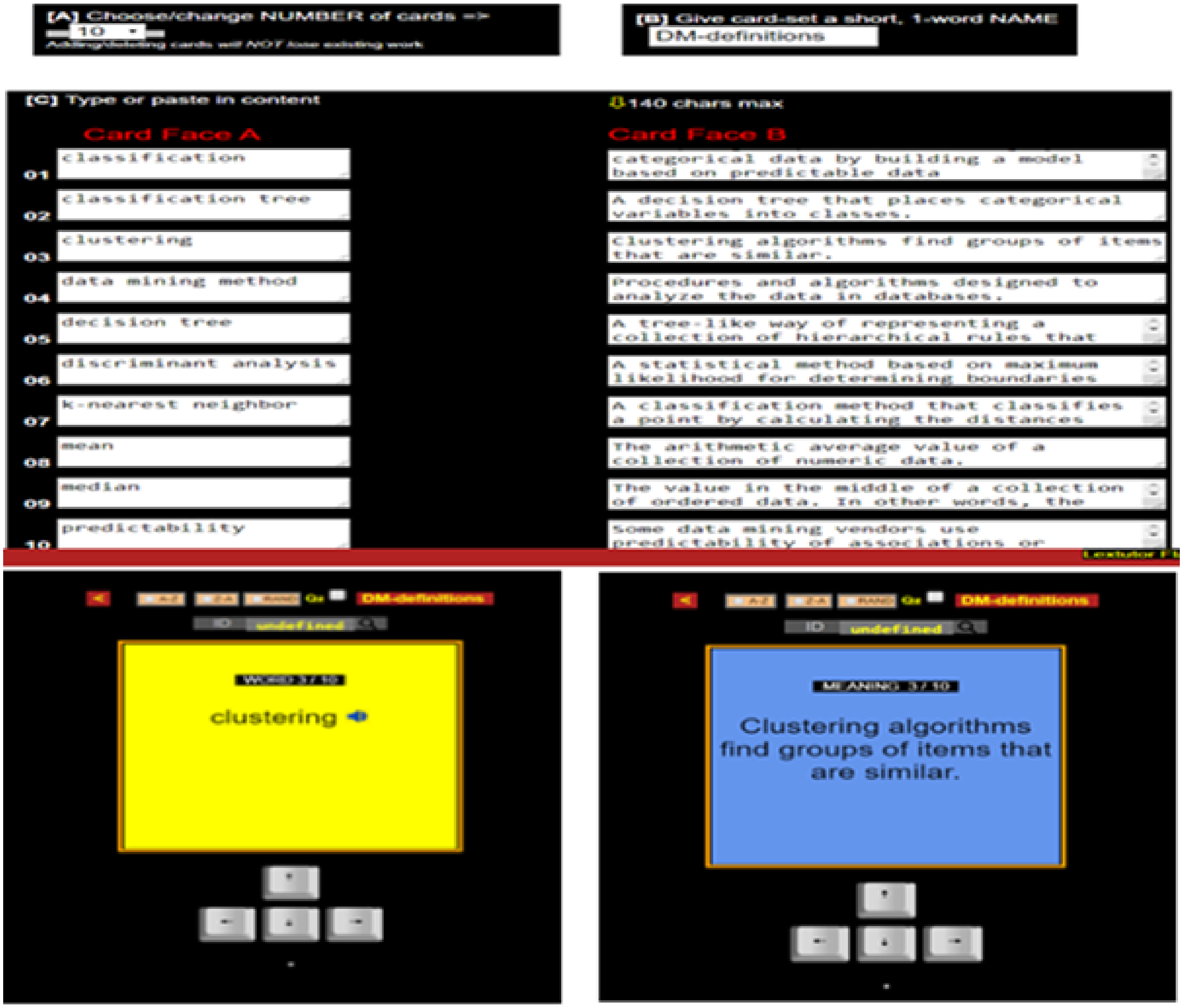

Text pre-processing, transcoding, removal of stop words, lemmatization, selection of sentences: AntFileConverter
Building a user's corpus or choosing a basic corpus: LexTutor/ SketchEngine
Word List building: AntConc/ LexTutor/ScetchEngine
Collocation building: AntConc/SketchEngine/Skell/
Sentences extraction: AntConcLab/SketchEngine/Skell/
Extraction of keywords and terminology: LexTutor/ ScetchEngine
Exercises of term actualization using the constructed concordances: LexTutor/ Skell/LexSite-LexTutor
Creating an online dictation based on key terms in the training and test mode:LexTutorDictator
Flesh card creating: LexTutor, LexSite-LexTutor

Figure 2: Corpus scaffolding for terminology training

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



[A] Choose/change NUMBER of cards =>

10

Adding/deleting cards will NOT lose existing work

[B] Give card-set a short, 1-word NAME

DM-definitions

[C] Type or paste in content

140 chars max

	Card Face A	Card Face B
01	classification	categorical data by building a model based on predictable data
02	classification tree	A decision tree that places categorical variables into classes.
03	clustering	Clustering algorithms find groups of items that are similiar.
04	data mining method	Procedures and algorithms designed to analyze the data in databases.
05	decision tree	A tree-like way of representing a collection of hierarchical rules that
06	discriminant analysis	A statistical method based on maximum likelihood for determining boundaries
07	k-nearest neighbor	A classification method that classifies a point by calculating the distances
08	mean	The arithmetic average value of a collection of numeric data.
09	median	The value in the middle of a collection of ordered data. In other words, the
10	predictability	Some data mining vendors use predictability of associations or

WORD 3/ 10

clustering

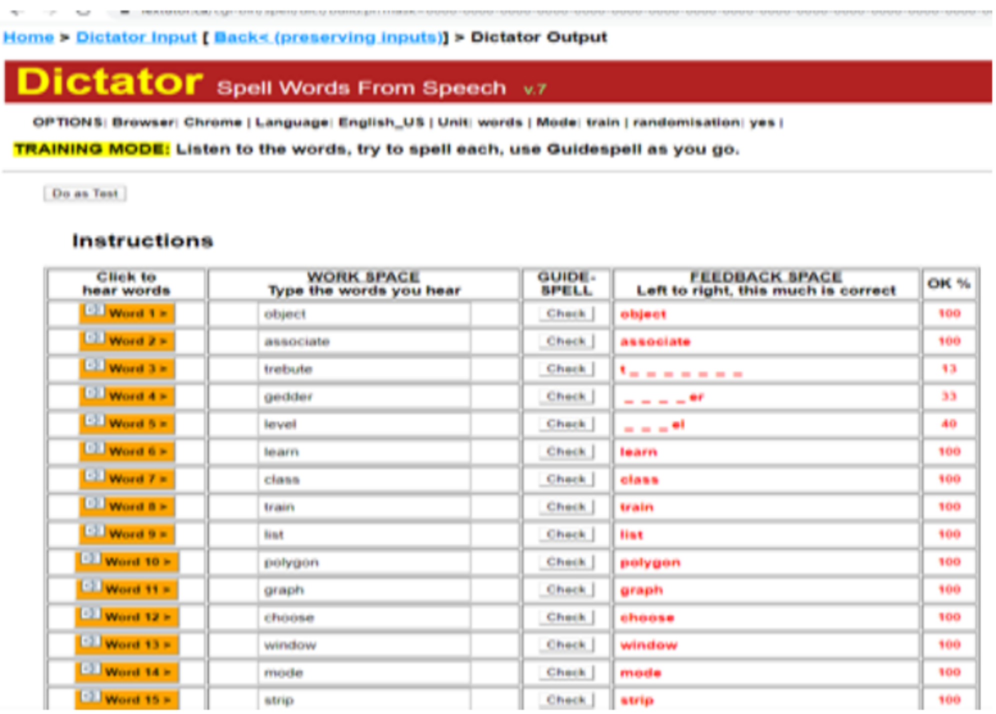
MEANING 3/ 10

Clustering algorithms find groups of items that are similar.

Figure 3: An example of LexTutor flashcards for basic terminology in Data-Mining course

Table 1: Corpus manager's characteristics

Corpus manager	User's Corpora Link	User's texts Link	Didactic suppurt	Interface service
AntConcLab	free	5	0	5
LexTutor	free	5	5	2
LexEngine	Commercial	5	0	5
SkELL	Commercial	5	0	5
LexSite-LexTutor	Commercial	5	4	5



Home > Dictator Input [Back< (preserving inputs)] > Dictator Output

Dictator Spell Words From Speech v.7

OPTIONS: Browser: Chrome | Language: English_US | Unit: words | Mode: train | randomisation: yes |

TRAINING MODE: Listen to the words, try to spell each, use Guidespell as you go.

Do as Test

Instructions

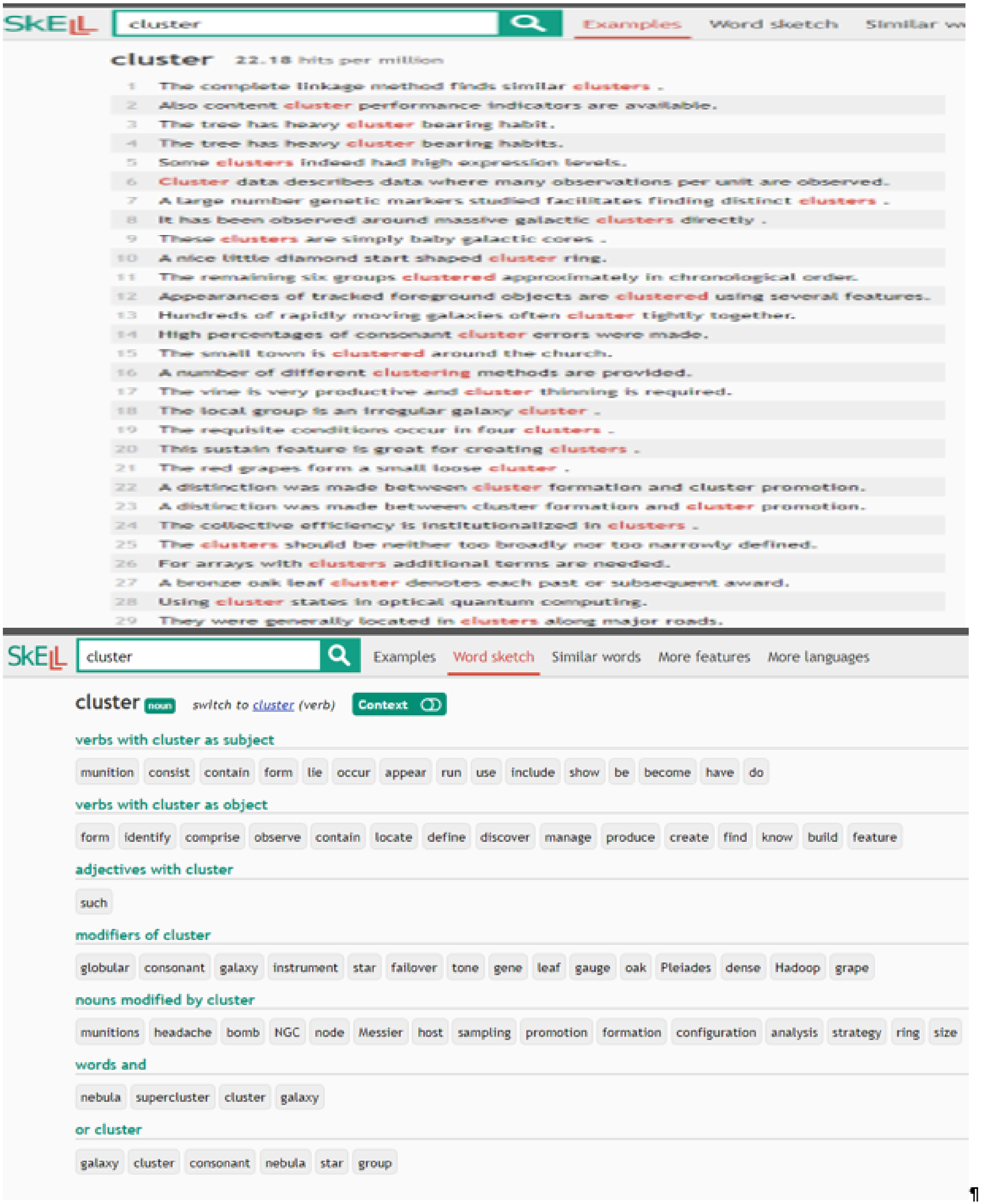
Click to hear words	WORK SPACE Type the words you hear	GUIDE-SPELL	FEEDBACK SPACE Left to right, this much is correct	OK %
Word 1 >	object	Check	object	100
Word 2 >	associate	Check	associate	100
Word 3 >	trebute	Check	t_ _ _ _ _ _ _	13
Word 4 >	gedder	Check	_ _ _ _er	33
Word 5 >	level	Check	_ _ _el	40
Word 6 >	learn	Check	learn	100
Word 7 >	class	Check	class	100
Word 8 >	train	Check	train	100
Word 9 >	list	Check	list	100
Word 10 >	polygon	Check	polygon	100
Word 11 >	graph	Check	graph	100
Word 12 >	choose	Check	choose	100
Word 13 >	window	Check	window	100
Word 14 >	mode	Check	mode	100
Word 15 >	strip	Check	strip	100

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



cluster 22.18 hits per million

1 The complete linkage method finds similar clusters .
2 Also content cluster performance indicators are available.
3 The tree has heavy cluster bearing habit.
4 The tree has heavy cluster bearing habits.
5 Some clusters indeed had high expression levels.
6 Cluster data describes data where many observations per unit are observed.
7 A large number genetic markers studied facilitates finding distinct clusters .
8 It has been observed around massive galactic clusters directly .
9 These clusters are simply baby galactic cores .
10 A nice little diamond start shaped cluster ring.
11 The remaining six groups clustered approximately in chronological order.
12 Appearances of tracked foreground objects are clustered using several features.
13 Hundreds of rapidly moving galaxies often cluster tightly together.
14 High percentages of consonant cluster errors were made.
15 The small town is clustered around the church.
16 A number of different clustering methods are provided.
17 The vine is very productive and cluster thinning is required.
18 The local group is an irregular galaxy cluster .
19 The requisite conditions occur in four clusters .
20 This sustain feature is great for creating clusters .
21 The red grapes form a small loose cluster .
22 A distinction was made between cluster formation and cluster promotion.
23 A distinction was made between cluster formation and cluster promotion.
24 The collective efficiency is institutionalized in clusters .
25 The clusters should be neither too broadly nor too narrowly defined.
26 For arrays with clusters additional terms are needed.
27 A bronze oak leaf cluster denotes each past or subsequent award.
28 Using cluster states in optical quantum computing.
29 They were generally located in clusters along major roads.

cluster noun switch to cluster (verb) Context

verbs with cluster as subject

munition consist contain form lie occur appear run use include show be become have do

verbs with cluster as object

form identify comprise observe contain locate define discover manage produce create find know build feature

adjectives with cluster

such

modifiers of cluster

globular consonant galaxy instrument star failover tone gene leaf gauge oak Pleiades dense Hadoop grape

nouns modified by cluster

munitions headache bomb NGC node Messier host sampling promotion formation configuration analysis strategy ring size

words and

nebula supercluster cluster galaxy

or cluster

galaxy cluster consonant nebula star group

¶

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: "*группироватьъся*", "*гроздъ*", "*кистъ*", "*теснитъся*", "*связыватъся*". 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.

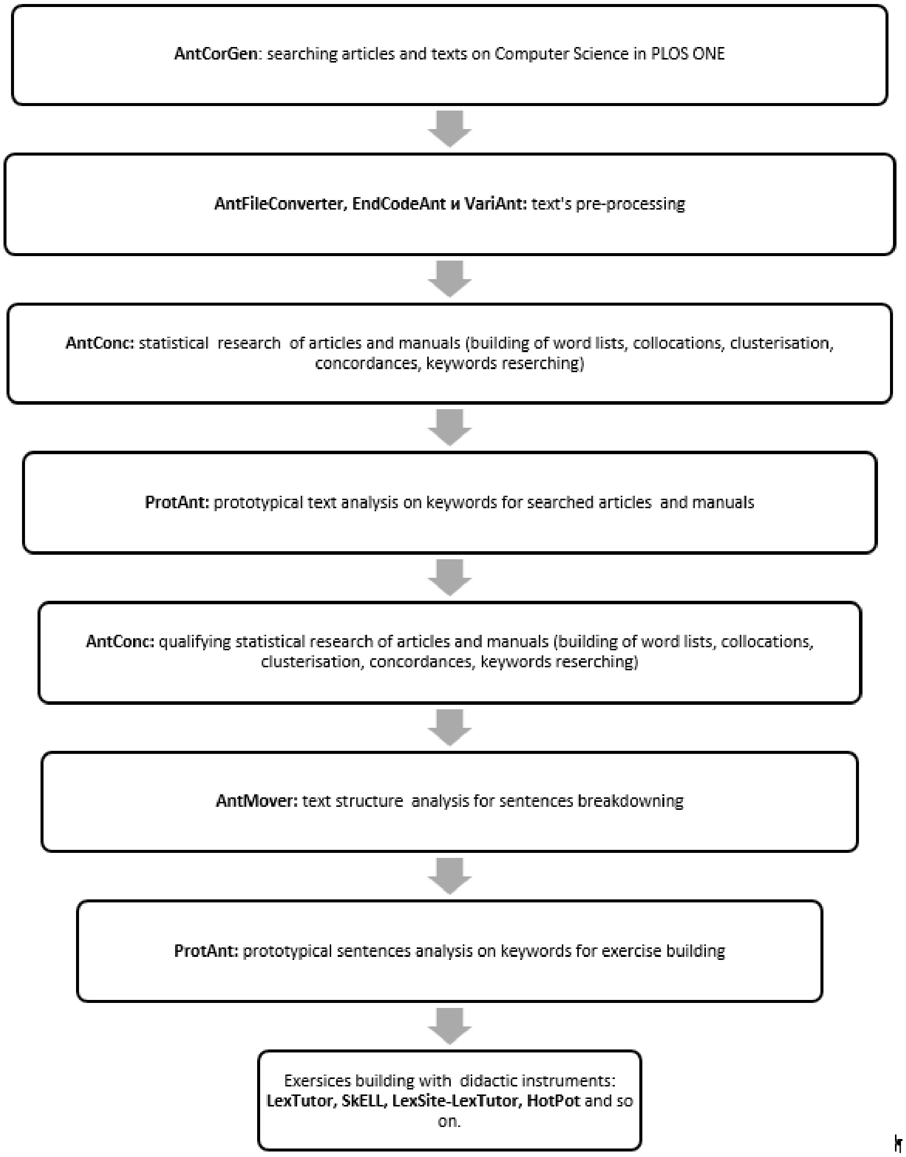
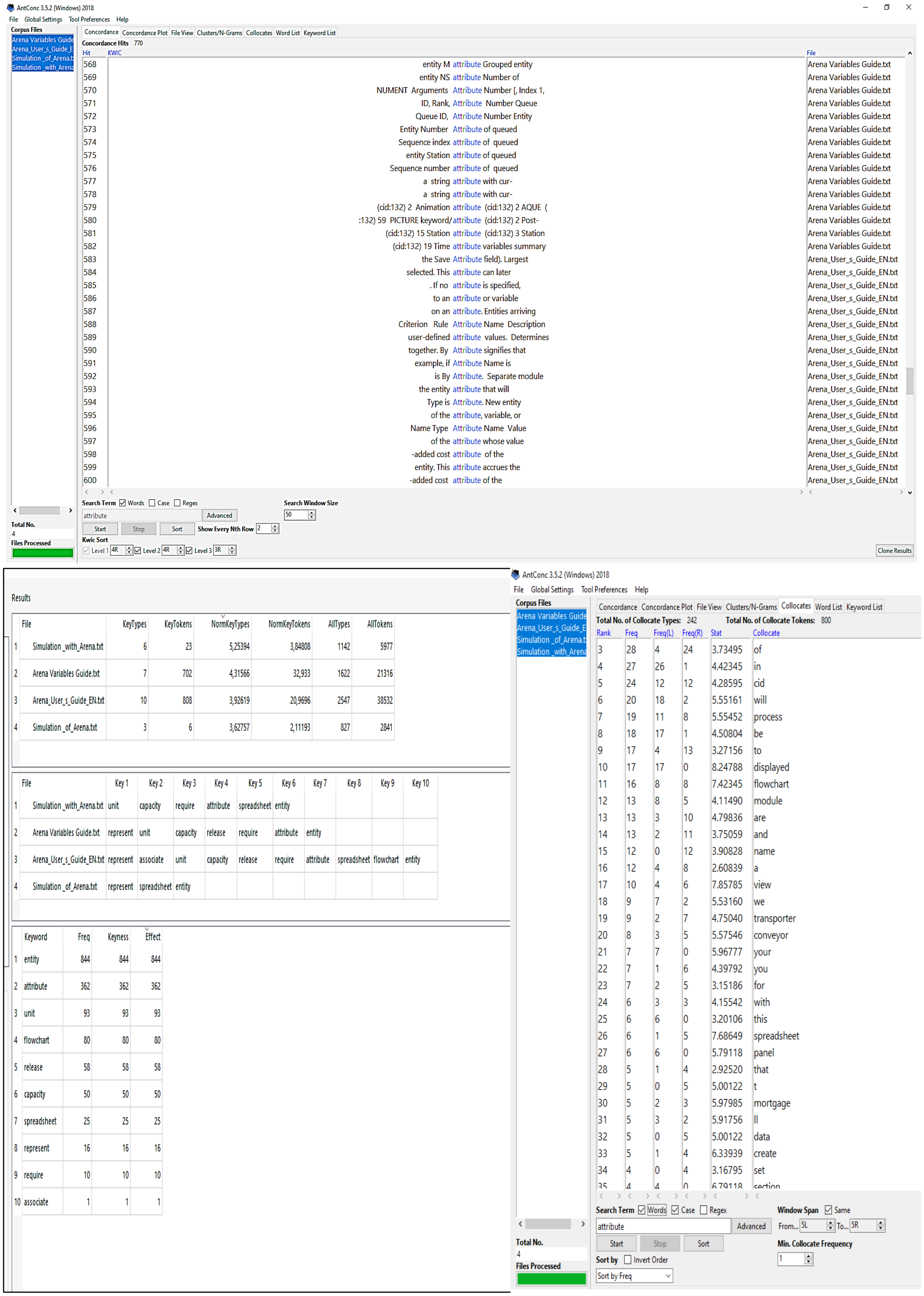

AntCorGen: searching articles and texts on Computer Science in PLOS ONE
AntFileConverter, EndCodeAnt и VariAnt: text's pre-processing
AntConc: statistical research of articles and manuals (building of word lists, collocations, clusterisation, concordances, keywords reserching)
ProtAnt: prototypical text analysis on keywords for searched articles and manuals
AntConc: qualifying statistical research of articles and manuals (building of word lists, collocations, clusterisation, concordances, keywords reserching)
AntMover: text structure analysis for sentences breakdowning
ProtAnt: prototypical sentences analysis on keywords for exercise building
Exersices building with didactic instruments: LexTutor, SkELL, LexSite-LexTutor, HotPot and so on.

Figure 6: Corpus scaffolding for terminology and grammar training with AntLab



AntConc 3.5.2 (Windows) 2018

Corpus Files: Arena Variables Guide, Arena_User_s_Guide_E, Simulation _of_Arena.t, Simulation _with_Arena

Concordance Hits 770

Hit	KWIC	File
568	entity M attribute Grouped entity	Arena Variables Guide.txt
569	entity NS attribute Number of	Arena Variables Guide.txt
570	NUMENT Arguments Attribute Number [, Index 1,	Arena Variables Guide.txt
571	ID, Rank, Attribute Number Queue	Arena Variables Guide.txt
572	Queue ID, Attribute Number Entity	Arena Variables Guide.txt
573	Entity Number Attribute of queued	Arena Variables Guide.txt
574	Sequence index attribute of queued	Arena Variables Guide.txt
575	entity Station attribute of queued	Arena Variables Guide.txt
576	Sequence number attribute of queued	Arena Variables Guide.txt
577	a string attribute with cur-	Arena Variables Guide.txt
578	a string attribute with cur-	Arena Variables Guide.txt
579	(cid:132) 2 Animation attribute (cid:132) 2 AQUE (	Arena Variables Guide.txt
580	:132) 59 PICTURE keyword/attribute (cid:132) 2 Post-	Arena Variables Guide.txt
581	(cid:132) 15 Station attribute (cid:132) 3 Station	Arena Variables Guide.txt
582	(cid:132) 19 Time attribute variables summary	Arena Variables Guide.txt
583	the Save Attribute field). Largest	Arena_User_s_Guide_EN.txt
584	selected. This attribute can later	Arena_User_s_Guide_EN.txt
585	. If no attribute is specified,	Arena_User_s_Guide_EN.txt
586	to an attribute or variable	Arena_User_s_Guide_EN.txt
587	on an attribute. Entities arriving	Arena_User_s_Guide_EN.txt
588	Criterion Rule Attribute Name Description	Arena_User_s_Guide_EN.txt
589	user-defined attribute values. Determines	Arena_User_s_Guide_EN.txt
590	together. By Attribute signifies that	Arena_User_s_Guide_EN.txt
591	example, if Attribute Name is	Arena_User_s_Guide_EN.txt
592	is By Attribute. Separate module	Arena_User_s_Guide_EN.txt
593	the entity attribute that will	Arena_User_s_Guide_EN.txt
594	Type is Attribute. New entity	Arena_User_s_Guide_EN.txt
595	of the attribute, variable, or	Arena_User_s_Guide_EN.txt
596	Name Type Attribute Name Value	Arena_User_s_Guide_EN.txt
597	of the attribute whose value	Arena_User_s_Guide_EN.txt
598	-added cost attribute of the	Arena_User_s_Guide_EN.txt
599	entity. This attribute accrues the	Arena_User_s_Guide_EN.txt
600	-added cost attribute of the	Arena_User_s_Guide_EN.txt

Search Term ☑ Words ☐ Case ☐ Regex: attribute — Advanced

Search Window Size: 50

Start | Stop | Sort | Show Every Nth Row: 2

Kwic Sort: ☑ Level 1 4R ☑ Level 2 4R ☑ Level 3 3R

Total No. 4

Files Processed

Results

	File	KeyTypes	KeyTokens	NormKeyTypes	NormKeyTokens	AllTypes	AllTokens
1	Simulation _with_Arena.txt	6	23	5,25394	3,84808	1142	5977
2	Arena Variables Guide.txt	7	702	4,31566	32,933	1622	21316
3	Arena_User_s_Guide_EN.txt	10	808	3,92619	20,9696	2547	38532
4	Simulation _of_Arena.txt	3	6	3,62757	2,11193	827	2841

	File	Key 1	Key 2	Key 3	Key 4	Key 5	Key 6	Key 7	Key 8	Key 9	Key 10
1	Simulation _with_Arena.txt	unit	capacity	require	attribute	spreadsheet	entity				
2	Arena Variables Guide.txt	represent	unit	capacity	release	require	attribute	entity			
3	Arena_User_s_Guide_EN.txt	represent	associate	unit	capacity	release	require	attribute	spreadsheet	flowchart	entity
4	Simulation _of_Arena.txt	represent	spreadsheet	entity							

	Keyword	Freq	Keyness	Effect
1	entity	844	844	844
2	attribute	362	362	362
3	unit	93	93	93
4	flowchart	80	80	80
5	release	58	58	58
6	capacity	50	50	50
7	spreadsheet	25	25	25
8	represent	16	16	16
9	require	10	10	10
10	associate	1	1	1

AntConc 3.5.2 (Windows) 2018

Corpus Files: Arena Variables Guide, Arena_User_s_Guide_E, Simulation _of_Arena.t, Simulation _with_Arena

Total No. of Collocate Types: 242 Total No. of Collocate Tokens: 800

Rank	Freq	Freq(L)	Freq(R)	Stat	Collocate
3	28	4	24	3.73495	of
4	27	26	1	4.42345	in
5	24	12	12	4.28595	cid
6	20	18	2	5.55161	will
7	19	11	8	5.55452	process
8	18	17	1	4.50804	be
9	17	4	13	3.27156	to
10	17	17	0	8.24788	displayed
11	16	8	8	7.42345	flowchart
12	13	8	5	4.11490	module
13	13	3	10	4.79836	are
14	13	2	11	3.75059	and
15	12	0	12	3.90828	name
16	12	4	8	2.60839	a
17	10	4	6	7.85785	view
18	9	7	2	5.53160	we
19	9	2	7	4.75040	transporter
20	8	3	5	5.57546	conveyor
21	7	7	0	5.96777	your
22	7	1	6	4.39792	you
23	7	2	5	3.15186	for
24	6	3	3	4.15542	with
25	6	6	0	3.20106	this
26	6	1	5	7.68649	spreadsheet
27	6	6	0	5.79118	panel
28	5	1	4	2.92520	that
29	5	0	5	5.00122	t
30	5	2	3	5.97985	mortgage
31	5	3	2	5.91756	ll
32	5	0	5	5.00122	data
33	5	1	4	6.33939	create
34	4	0	4	3.16795	set
35	4	4	0	6.79118	section

Search Term ☑ Words ☑ Case ☐ Regex: attribute — Advanced

Window Span ☑ Same: From... 5L To... 5R

Start | Stop | Sort

Min. Collocate Frequency: 1

Sort by ☐ Invert Order: Sort by Freq

Total No. 4

Files Processed

Figure 7: Screenshots of AntConc and ProtAnt programs results for Arena Simulation manuals

Concordancer AntConc is very robast 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

[Berg et al., 2019] Berg E., Kit, M. (2019) Reference and educational system Lexsite-Lextutor as information technology for foreign language learning //CEUR Workshop Proceedings, Vol.2562, http://ceur-ws.org/Vol-2562/paper-01.pdf

[Bernardini et al., 2008] Bernardini S., Castagnoli S.(2008) Corpora for translator education and translation practice // Topics in language resources for translation and localization (Edited by E.Y. Rodrigo). 2008. Amsterdam/ Philadelphia: John Benjamins Publishing Company. P. 38–55.

[Carloni, 2018] Carloni G. English-Taught Programs and Scaffolding in CLIL Setting:Case study // La didattica delle lingue nel nuovo millennio. Le sfide dell'internazionalizzazione. Atti del IV Congresso della società di Didattica delle Lingue e Linguistica Educativa DILLE (Università Ca' Foscari Venezia, 2-4 th february 2017), Ed. by C. M. Coonan, A. Bier and E. Ballarin, Venezia: Edizioni Ca' Foscari - Digital Publishing, Pp.483-498

[Gavrilova Kogan, 2016] Gavrilova A.V., Kogan M.S.(2016) Didakticheskiye aspekty ispol'zovaniya internet-resursov v organizatsii vneauditornoy raboty studentov //Voprosy metodiki prepodavaniya v vuze: yezhegodnyy sbornik. 2016. No 5 (19-1). S. 251-257.(In Rus) == Гаврилова А.В., Коган М.С. Дидактические аспекты использования интернет-ресурсов в организации внеаудиторной работы студентов //Вопросы методики преподавания в вузе: ежегодный сборник. 2016. No 5 (19-1). С. 251-257.

[Cherniakova, 2011] Chernyakova T.A. (2011) Ispol'zovaniye lingvisticheskogo korpusa v obuchenii inostrannomu yazyku // YAzyk i kul'tura. 2011, No 4. S. 119 – 125.(In Rus) == Чернякова Т.А. Использование лингвистического корпуса в обучении иностранному языку // Язык и культура. 2011, No 4. С. 119 – 125.

[Coyle et. al, 2010] Coyle, D., Hood, P., Marsh, D., 2010. CLIL: Content and Language Integrated Learning. Cambridge University Press, Cambridge

[Corino & Ernesti, 2019] Corino E., Onesti C.(2019) Data-driven learning: a scaffolding methodology for CLIL and LSP teaching and learning //Frontiers in Education, 2019, Vol., art. 7

[Cobb, 2007] Cobb T.(2007) Computing the vocabulary demands of L2 reading. //Language Learning Technology, 2007, Vol. 11. No 3. P. 38–64.

[Jons, 1991] Johns T. Should You Be Persuaded: Two Samples of Data-Driven Learning Materials // English Language Research Journal, 4, 1991. P. 1–16.

[Gorina, 2018] Gorina O.G. (2018) Instrumenty korpusnogo analiza v obuchenii inostrannomu yazyku //Vestnik Tomskogo gosudarstvennogo universiteta. No 435. S.187-194(In Rus) == Горина О.Г. Инструменты корпусного анализа в обучении иностранному языку //Вестник Томского государственного университета. 2018. No 435. С. 187-194.

[Marsh, 2002] Marsh, D. (Ed.). (2002). CLIL/EMILE European dimension: Action, trends and foresight potential. (European Commission Public Services Contract DG 3406/001-001). Retrieved from http://ec.europa.eu/languages/documents/doc491_en.pdf

[Mayer, 2015] Mayer R. V.(2015) Computer models for saltatory and continuous knowledge increase in teaching//International Journal of Open Information Technologies ISSN: 2307-8162 vol. 3, no. 9,(In Rus.) == Майер Р.В. Компьютерные модели скачкообразного и непрерывного увеличения знаний при обучении // International Journal of Open Information Technologies ISSN: 2307-8162 vol. 3, no. 9, 2015

[Nymm et al., 2017] Nymm V., Piotrowska X., Nõmm S. (2017) Using stochastic learning theory and corpus tools in CALL. In CEUR Workshop Proceedings, 2nd International Conference R. Piotrowski's Readings in Language Engineering and Applied Linguistics, Herzen State Pedagogical University, vol. 2233. Saint Petersburg; Russian Federation, November.

[Piotrowska, 2005] Piotrowska X. Computer-assisted language learning. The quantitative-linguistic basis of CALL methods(2005) (Computer-unterstützter Spracherwerb. Die quantitativ - linguistische Grundlage von CALLMethoden)// Quantitative Linguistik / Quantitative Linguistics, 2005: Pp. 897-908

[Piotrowska, 1991] Piotrovskaya K.R.(1991) Sovremennaya komp'yuternaya lingvodidaktika //Nauchno-tekhnicheskaya informatsiya. Seriya 2: Informatsionnyye protsessy i sistemy, 1991. No 4. S. 26-29.(In Rus) == Пиотровская К.Р. Современная компьютерная лингводидактика //Научно-техническая информация. Серия 2: Информационные процессы и системы, 1991. No 4. С. 26-29.

[Thomas, 2016 a] Thomas J. (2016) Discovering English with Sketch Engine: A Corpus-Based Approach to Language Exploration. 2nd ed. Versatile, 2016.

[Thomas, 2016 b] Thomas J.(2016) Discovering English with Skell: Workbook and Glossary. Versatile, 2016.

[Yakaeva, 2017] Yakaeva T., Salekhova L., Kuperman K., Grigorieva K. (2017) Content And Language Integrated Learning: Language Scaffolding And Speech Strategies //Modern Journal of Language Teaching Methods (MJLTM) Vol. 7, Issue. 9, September(2017) Published by EBSCO. Pp. 137–143