ISSN 1798-4769
Journal of Language Teaching and Research, Vol. 4, No. 1, pp. 105-112, January 2013

doi:10.4304/jltr.4.1.105-112

A Study on the Application of Data-driven Learning in Vocabulary Teaching and Leaning in China's EFL Class

Xiaowei Guan
School of Foreign Languages, Dalian University of Technology, Dalian, China

***Abstract*—Data-driven learning (DDL) developed from corpus linguistics plays a pioneering role in the evolution of EFL teaching, allowing the learners to indentify and induce language rules by observing numerous real corpora concordances. With the guidance of new teaching notion advocated by College English Curriculum Requirements in China, DDL model has become the tendency for college English learners' efficient and autonomous learning. It becomes urgent to take advantage of this method and apply it to college vocabulary teaching. Compared with traditional foreign language teaching and learning method, data-driven learning is characterized by "autonomic learning", "authentic language input", "self-discovery", and "bottom-up inductive learning", which will conducive to the formation of the students' personalized learning abilities and the development of their autonomic learning abilities.**

***Index Terms*—data-driven learning (DDL), vocabulary teaching, EFL, corpus**

From the 1990s, computer network as the core of modern information technology are developing rapidly, providing more favorable conditions and vast space for foreign language teaching. The international education circles attach great importance to the development of learners' basic knowledge and practical abilities to autonomically access, analyze, process and use the information through computer and network.

In 2004, the Ministry of Education of China promulgated the "College English Curriculum Requirements (For Trial Implementation)" (2004) which marks the beginning of a new round of college English teaching reform. One of the most important parts of this reform is to vigorously promote the application of information technology in college English reform, seeking to use the new teaching model nationwide. "College English Curriculum Requirements" (2007) holds "the new model should be based on modern information technology, particular network technology, so that English language teaching will be free from the constraints of time and place, taking into consideration students' individualized and autonomous learning". English Teachers should apply a large number of advanced information technology into college English curriculum design, and develop and construct a variety of computer and network-based courses.

These requirements not only clearly indicate the emphasis of our college English teaching in language teaching, but also put forward a new aim to improve the current teaching methods.

I. The Problems of Current College English Vocabulary Teaching and College English Teaching Reform in China

Vocabulary is a fundamental component of a language and of critical importance to the EFL learners. Vocabulary teaching is an important component of language teaching.

However there are still many problems existing in vocabulary teaching in China's EFL class. The teaching forms are relatively simple, and the teaching methods are lack of innovation. In traditional vocabulary teaching, the sample sentences are usually extracted from a certain dictionary or compiled by teachers, which contain a limited amount of information and are difficult to guarantee the authenticity of the sentences. The Sentences, which show no adequacy and vividness, fail to arouse the students' attention, and is unbeneficial to cultivating learner' initiative. In general, the present English vocabulary teaching still sticks to the teachers and textbooks centered pattern, and the top-down traditional foreign language teaching mode, in which teaching content and methods focus on abstract explaining and simple exercises.

II. Corpus

A corpus (plural 'corpora') is simply a collection of texts. A corpus is a large and principled collection of naturally occurring texts. The size of a corpus can range from tens of millions of words to a few thousand. The texts can be either transcripts of spoken language (increasingly with sound or visual files attached) or written language that has been scanned from books, newspapers etc. or downloaded electronically.

With the rapid development and wide application of computer technology, computer technology-based corpus technology becomes mature and become a powerful tool for the language study and teaching. Great concern given by the language researchers and language teachers, contemporary large-scaled corpora came into being.

For example, British National Corpus (BNC) comprises approximately 100 million words of written texts (90%) and transcripts of speech (10%), which aims to represent the universe of contemporary British English. Collins Birmingham University International Language Database (COBUILD) comprises approximately 500 million words. These large corpora can be used to help us study on the cross-cultural and cross-language English comparative analysis.

Corpus Linguistics in China began in the 1980s, which has been rapidly developed during the past 20 years. Now many corpora have been built. JDEST (Jiao Da English Corpus for Science and Technology) and Chinese Learner English Corpus (CLEC) was developed around the year 2000. There are other famous corpora in China such as the Corpora of English Education in China (CEEC), International Corpus of Learner English (ICLE), and Spoken and Written English Corpus of Chinese Learners (SWECCL). A corpus is always designed for a particular purpose and the type of corpus will depend on its purpose (He, 2004).

For the study on corpus in foreign language teaching, many of the relevant theories and applications of corpus have been introduced. Yang (2002) holds the view that the applications of corpus are reflected in the statistics of language frequency, dictionary compilation, the study on vocabulary collocation, language teaching and natural language processing. Zhen (2005) makes a systematic introduction and illustration of the idea, methods and techniques of corpus based data-driven foreign language learning. He believes that compared with traditional English teaching and learning, data-driven learning is characterized by “autonomic learning”, “authentic language input”, “self-discovery”, and “bottom-up inductive learning”.

III. Data-driven Learning

One of the significant features of corpus-based study is data-driven. The data-driven quantitative analysis makes us discover the problems that we could not find by intuition. DDL (data-driven learning) refers to the discovery and exploration-based learning model based on corpus by using the original data in the corpus or the retrieval results by the corpus se retrieval tools.

A. DDL, Constructivism, Lexical Grammar Theory

Constructivism as a paradigm or worldview posits that learning is an active, constructive process. The learner is an information constructor. People actively construct or create their own subjective representations of objective reality. New information is linked to prior knowledge, thus mental representations are subjective (Feng & Cai, 2009).

In the view of constructivist, learning is a constructive process in which the learner is building an internal illustration of knowledge, a personal interpretation of experience. This representation is continually open to modification, its structure and linkages forming the ground to which other knowledge structures are attached. Learning is an active process in which meaning is accomplished on the basis of experience. This view of knowledge does not necessarily reject the existence of the real world, and agrees that reality places constrains on the concepts that are, but contends that all we know of the world are human interpretations of our experience of the world. Conceptual growth comes from the sharing of various perspectives and the simultaneous changing of our internal representations in response to those perspectives as well as through cumulative experience.

The Theory of Constructivism is the theoretical basis of "the English teaching model based on computer and classroom" (Zhang, 2010, p.59). Data-driven learning embodies the Theory of Constructivism by making language learners research the language system on their own and become the masters of language learning.

Lexical Grammar Theory is an important theoretical contribution of linguist Sinclair for language study. Sinclair (2009) believes that a description of English cannot divide the language into two separate components, lexis and grammar, since grammatical features are decided by lexis and all lexical elements can have grammatical patterns. He observes that lexical meaning and grammatical meaning make up the word meaning, as neither can exist without the other, and many uses of words and phrases show a tendency to co-occur with certain grammatical choices. Lexical Grammar Theory maintains that language teaching should start from lexis and their core meanings and the most typical collocations in language teaching and learning. Special attention should be paid to that most frequently occurring high-frequency lexis. In this theory, the meaning of word depends on their surrounding lexical items and frequent semantic selective tendency, and the context of the words meaning is constructed by highlighting word’s collocation, colligation, semantic preference and semantic prosody semantic prosody by corpus.

B. Definition of DDL

DDL (data-driven learning) is defined by Johns (1991) in 1991 as “the use in the classroom of computer-generated concordances to get students to explore regularities of patterning in the target language, and the development of activities and exercises based on concordance output.”

It is an approach based on corpus and concordance-based materials to learn language. Learners who have a certain problem will use of retrieval software to discover rules and draw conclusions on the basis of observing and analyzing a large number of real corpus, and master a grammatical structure or word usage through real-time practice. So the

learning is also known as "research-then-theory" approach. Since the concept was proposed in 1991, an increasing emphasis on data-driven learning has been paid at home and abroad. Many scholars in China have proposed the application of DDL into the innovation and development of English teaching.

Hunston (2002, p.170) also points out that "DDL involves setting up situations in which students can answer questions about language themselves by studying corpus data in the form of concordance lines or sentences".

C. The Characteristics of DDL

There is a big difference data-driven language learning and traditional teaching mode. Its characteristics can be summarized in four aspects (Zhen, 2005).

First, DDL centers on learner autonomy. Data-driven learning emphasizes the students' autonomic learning, in which activities in the class are student-centered rather than led by the teacher to give full play to their personal characteristics. Learners are not passive recipients of the knowledge, but take on the active roles of discoverers and researchers, sorting through massive language data to discover rules and patterns embedded in the data, and can self-regulate learning strategies according to their own requirements. In other words, learner autonomy is cultivated by encouraging students to be responsible for their own learning.

Second, DDL uses authentic, rich massive corpora as the main language input. Corpus-based data-driven learning provides students with high quality, vast amounts of language data from real communicative activities. In a word, data-driven learning can create an authentic language environment for students to improve their language intuition to practice their ability to deal with language variation, in order to help them acquire authentic language.

Third, DDL emphasizes the exploration and discovery of learning process. Students learn through problem-solving activities rather than being instructed directly by the teacher. Data-driven learning provides students with a lot of real data based on corpus, to guide students to observe the learning process according to their own needs to experience, explore, and discover the knowledge of language. As a result, the language knowledge that students acquire will be more authentic and systemic, and the impression will be deeper.

Finally, DDL advocates bottom-up, inductive learning. Not in the data-driven learning, students first come into contact with a large amount of authentic language data, but not prescriptive grammatical rules. After their independent observations, they will generalize grammatical rules. With concordance software, students can easily obtain a list of contextualized examples of the investigated feature when dealing with tasks such as the acquisition of grammatical structures and lexical items.

D. Three Procedures of DDL

Tim Johns (1991) describes his procedures of Identify-Classify-Generalize for classroom based on concordances and data-driven learning.

The first step of the procedure is to identify the structure under examination. It is possible for the structures or words to be teacher- or class- generated. Class-generated questions would create an immediate interest in the lesson as it would be a response to a learner's question. After identifying the area of inquiry a concordance research is necessary to find the citations. The citations are then edited to produce a list of the structures or words in the chosen context.

Classification is the second step of Johns' procedure. It is necessary so that the learners will not be discouraged by encountering overwhelming files of data. The teacher may make the classification for the students if the corpus is too big or the citation is too difficult.

Generalizing is the act of inductively constructing rules describing the usage of the structures or words. The act of generalizing represents an essential part of the learning process with a DDL activity because students are actively engaged with the cognitive process of generalizing rules for the language. This process of generalizing is completely foreign to many Chinese students who are educated in an educational system that values memorization rather than the production of generalizations or theories.

IV. The Application of DDL in Vocabulary Teaching

In China, though corpus is still dominantly used in language research, there have been several researchers putting it to classroom teaching. Based on the theoretical directions, both home and abroad, latest researches have attempted to integrate corpus to language learning, vocabulary learning dominantly. In effect, DDL can provide such a space in which natural language is the dominant component, for in the term of "data-driven learning", "data" refers to authentic texts, the fundamental elements of the corpora.

Vocabulary is the "building blocks" of language, and is the basis of language understanding expression.

DDL is an advanced computer-aided teaching mode based on corpus index. DDL advocates students to take the initiative to explore vocabulary usage to accurately grasp the vocabulary by observing authentic linguistic phenomenon. Through DDL, the authentic data can assist learners in getting accustomed to the target language communication and help them acquire the language use successfully. The fact that the corpus can offer authentic materials conveniently makes data-driven learning valuable in foreign language pedagogy.

This new teaching model can really enrich the amount of information in the classroom. On the other hand and change the traditional vocabulary teaching method. More importantly, the large amount of language materials in corpus will

provide learners with a variety of inductive and deductive language learning opportunities, to deepen learners' acquisition of the target knowledge.

It not only brings about a challenge to the traditional foreign language teaching model which centers on teachers and textbooks, but also provides a new idea to solve the various problems in English vocabulary teaching.

A. The Innovative Application of DDL in Vocabulary Teaching

With the extensive application of information technology, the growing popularity of multimedia technology in language teaching no doubt provides the necessary technical prerequisites for foreign language learning. At present, classroom teaching equipments have realized networking, and students' abilities of using network technology have gradually increased. Multimedia teaching has been applied into various disciplines. The major corpora have become shared resources, and part of the corpus retrieval software can be downloaded directly on the computer. All these factors provide research basics for the implementation and carrying out of data-driven learning in foreign language teaching.

In vocabulary learning, learners can enter the word string that they want to retrieve, and will extract the examples related to this word or a particular language phenomenon with corpus retrieve function within a few seconds from a corpus of thousands of millions of words.

Learners can find the grammatical rules of the word by the help of these instances to achieve the effect of drawing inferences. The examples can also help them grasp the usage of words, their collocations, their language environment, and their co-occurrence with a specific grammatical structure. The observation and analysis of concordance can deepen their impression on some of the vocabulary and linguistic phenomena, and enhance their language awareness, which will make them truly master the word usage and collocation context in the target language in the process of the self-exploration.

One of the main techniques of corpus linguistics is a word concordance, which should also be a common means of data-driven learning. The target words are always presented in the KWIC (keywords in context) format, which lists all the contexts of the same word together, where the target word is always highlighted, which can save learners' energy and help to focus their attention. In this way, the learners could learn the target words intentionally. Concordance line refers to the co-occurrence of keywords and their context. Currently, much corpus software has index functions, such as Mconcord, Wordsmith. Inputting keywords, the software automatically retrieves the corpus, showing the context with a fixed number of words around the keywords, and the keywords are displayed on the screen. The number of words on the left or the right of the keywords makes up the "word span" of the keyword. The words in the word span constitute the context of the key words. The context is a continuous text around the keywords, which can be extendedly displayed in the line, paragraph, or even discourse that the keywords lie in. Retrieval can be widely used in the study of English vocabulary, grammar and discourse.

At present, data-driven learning is mostly applied in English vocabulary teaching, including collocation, colligation, and semantic prosody.

a. Collocation

Collocation is undoubtedly one of the most important concepts in the field of linguistics and applied linguistics. Sinclair (1991) defined collocation as the co-occurrence of two or more words within a short distance in the text. For general language teaching and research, collocation is a sequence of words with certain non-idiom meaning in the text and used following certain grammatical forms, the words making up the sequence co-occur with a greater probability than accident. (Wei, 2001)

The analysis of collocation is of great significance in the study of word behavior. Words and words' collocation not only play a restrictive role to the establishment of the syntactic structure relations, but also are the fundamental basis for the realization of meaning and the elimination of ambiguity.

Now, we will take the word "avoid" for an example. Teachers need to guide students how to observe the usage and the overall features of the collocation of "avoid" through the concordance lines. We carry out KWIC retrieval in BNC (British National Corpus, 100 million) online. Part of the concordance lines are shown in Fig. 1.

A	B	C	. We have always tried at every stage to to erm	avoid	a commitment of any kind to an outer northern route . We
A	B	C	faith would be given their proper place . This concern to	avoid	a false polarisation was certainly justified ; but it is a real
A	B	C	And with three eclipses , you may not be able to	avoid	a fateful attraction . Letters section follows Tax on food IF
A	B	C	For example , it is important for an inexperienced solo pilot to	avoid	a field landing until he has had some training in how to
A	B	C	EC , in the sense that the present German wish to	avoid	active external commitments and to concentrate on internal
A	B	C	people to believe they could manage independent living and	avoid	admission to residential care . Social workers above all have
A	B	C	for the two insulin species on routine monitoring . Attempts to	avoid	altered glycaemic control , which is a confounding variable ,
A	B	C	network diagrams is the dummy job . This is necessary to	avoid	ambiguity , unnecessary constraints in the plan and to avoid
A	B	C	the military 's desire , on the one hand , to	avoid	attacks on its competence or political attitudes and , on the
A	B	C	occasions he has tried acid , it was used properly to	avoid	bad trips ; thus , he had " come to terms with
A	B	C	see the Shah , they walked around the palace garden to	avoid	being overheard by Mosadeq " a spies or microphones . The Shah
A	B	C	(c) Complete the sequence with reverse punch Faults to	avoid	Beware of over-committing yourself to the foot sweep . The
A	B	C	camera film . But tourists who want value for money should	avoid	buying a postcard in Bahrain or hiring a deckchair in Japan .
A	B	C	then the other issue I think is criterion twelve , erm	avoid	conflict with mineral and non-mineral development . Again I
A	B	C	and people went up one aisle and down the next to	avoid	confusion . This worked very well , but in 1988 people were
A	B	C	of the electricity industry , is under pressure to	avoid	creating a private monopoly with BR services . He also appeared
A	B	C	practices " where managers may choose to play safe inorder to	avoid	criticism " if anything goes wrong " , whereas their staff ,

Figure 1. Concordance lines of "avoid".

Before class teachers need to edit the original concordance lines and show them to students in the classroom. Under the guidance of teachers, students may find two main grammatical structures of "avoid" through observations, that is, "avoid + noun" and "avoid + doing". However, students cannot directly observe typicality of the collocated words, as well as the lexeme distribution (lexeme refers to the location of collocation words in the span), which is determined by statistical tools using Z-score or T-score. The higher the Z -score or T-score, the more typical the collocation is, and vice versa. We can use such as TACT and CAST software. Fig. 2 is the collocation words of "avoid" counted from the BROWN Corpus by Zhen (2005) (in descending order of the Z-score). Teachers can try to provide the collocation words and concordance lines of "avoid" for students to observe and generalize. They may also contrast the collocations of "avoid" of learners to point out their errors.

搭配词	Z值	搭配词	Z值	搭配词	Z值
unaggressive	42.08	rebuke	29.74	disgrace	24.27
sullying	42.08	nudity	29.74	pungent	21.00
puncturing	42.08	fugitive	29.74	nemesis	21.00
musing	42.08	bigotry	29.74	instability	18.78
nicked	42.08	stasis	24.27	repel	17.13
eye-strain	42.08	insomnia	24.27	convict	17.13
debacle	42.08	infliction	24.27	congestion	17.13
stuffy	29.74	disruption	24.27	scandal	14.82

Figure 2. Z-score of collocates of "avoid".

b. Colligation

Colligation is an important concept in the study of word collocation, which refers to the combination of abstract grammatical categories in the text, that is, the grammatical structure and framework of the collocation. For example, "V + N" represents a colligation. Through the establishment of colligation, we can find the grammatical pattern of the vocabulary. Words with different meanings have different grammatical patterns, and different words in the same grammatical pattern have certain connection in the meaning. For example, grammatical patterns of each word are given in Collins COBUILD English Dictionary. One of the grammatical patterns of "blaze" is "VP with n" (Her Eyes blazed with fury.) Mastering the characteristics of grammatical patterns is of great important for students to use foreign languages fluently and accurately.

Zhen (2005) took the noun "matter" as an example, adopting KWIC retrieval method by using Wordsmith software to search it from the BROWN corpus. Part of the concordance lines are shown in Fig. 3. Students can observe the concordance lines to find several major grammatical patterns of "matter": "a N of n /-doing", "a N-of pl-n", and a fixed phrase.

1	ng a forward roll is simply a	matter	of copying another child who can
2	only to a very limited extent a	matter	of devising new machinery of cons
3	this at all It is largely a	matter	of finding passages that suit one
4	As he grows older, it may be a	matter	of providing some accustomed object
5	two problems to consider, one is a	matter	of adjusting the fiscal calendar, the
6	South In this case it is primarily a	matter	of conflict of racial
7	man's face. It was simply a	matter	of curiosity, a natural right to
8	distance from earth, but is a	matter	of the development of the human
9	seek to be systematic; is not a	matter	of intuition," history has this in common
10	man's face. It was simply a	matter	of curiosity, a natural right to examine
11	principle. One problem is a	matter	of shifting dates; the other is
12	assented, vowing that in a	matter	of dollars and cents, his brother
13	be completely enclosed in a	matter	of three or four days. Then you can
14	were wholly inadequate. In a	matter	of months the war department
15	The result dramatically visible in a	matter	of days in the family's disrupted daily
16	to this experiment, nor, as a	matter	of fact, in those who were continuous
17	echoed mockingly. "What's the	matter,	Joe, you scared of me? Think I'm
18	E mud, at the rain. "Oh, what's the	matter	with me"? He demands
19	was about to enter college. As a	matter	of fact, Albert Flint expressed
20	on the spur of the moment. As a	matter	of fact, he wouldn't have cared

Figure 3. Concordance lines of "matter'.

c. Semantic Prosody

Semantic prosody refers to the semantic atmosphere created by the typical collocates of the keyword in its context (Wei, 2002a), which can be roughly divided into positive prosody, neutral prosody or intricate mixed prosody and negative prosody (Stubbs, 1996). Semantic prosody opens up a new research orientation for corpus linguistics, and provides a new perspective to the observation and description of word behavior (Wei, 2002b). According to Wei (2002a), the study on semantic prosody can be carried out by the following methods. (1) Build up a colligation, summarize and describe the semantic prosody of keywords based on data; (2) Calculate collocates, study semantic prosody by data-driven approach; (3) Build up the structure of semantic prosody based on a combination of data and data-driven approach. Regardless of which method to be adopted, the study must be data driven. Only with corpus, the semantic prosody that the words may have can be described scientifically by the statistics, analysis and generalization of data.

For example: Wei (2002b) studied the semantic prosody of "cause" by using Wordsmith to retrieve in JDEST corpus. "Cause" appears a total of 949 times, whose significant collocations can be determined by Z-score test, as shown in Fig. 4. Significant collocations reveal the typical behavior of the node word, which is the basis of the research of semantic prosody.

Collocates	Co-occur with node	Corp. Freq.	Z-score	Collocates	Co-occur with node	Corp. Freq.	Z-score
failure	34	848	21.36	diseases	4	131	6.26
bleeding	9	97	17.38	distortion	4	150	5.77
damage	21	511	17.03	erosion	4	153	5.70
death	15	334	15.13	faults	4	159	5.57
unemployment	7	85	14.39	collisions	3	97	5.46
injury	10	197	13.23	fission	5	250	5.38
trouble	6	107	10.83	fire	5	254	5.33
problems	27	1897	9.91	illness	3	110	5.06
disruption	4	58	9.89	cracking	4	203	4.77
harm	3	35	9.62	defects	5	305	4.71
worry	3	39	9.08	fault	5	316	4.60
deterioration	4	69	9.00	concern	5	338	4.38
wear	8	259	8.91	strain	7	621	4.23
disease	13	672	8.49	inflation	3	147	4.22
pain	6	166	8.46	vibration	4	257	4.07
injuries	4	80	8.30	loss	8	822	4.00
symptoms	7	236	8.13	dust	3	170	3.84
odor	3	54	7.62	cracks	3	173	3.79
breakdown	4	101	7.28	fall	4	285	3.78
collision	4	110	6.93	difficulties	4	298	3.65
diabetes	4	110	6.93	delay	3	252	2.89
disorder	3	65	6.87	reduction	6	758	2.86
errors	7	341	6.48	separation	4	405	2.86
stresses	9	540	6.39	shock	3	305	2.47
corrosion	8	444	6.36	fatigue	4	483	2.44
instability	4	128	6.34	crack	3	346	2.21

Figure 4. Features of Collocates of "cause".

B. Construct a Corpus

Obtaining the corpus data becomes very easy by the support of computer technology and network. In the process of constructing a corpus, you need to consider the following principles. (1) Principle of authenticity. The data collected in the corpus must be authentic and natural. (2) Principle of representativeness. The selection of the corpus should cover a wide range of content and fields, and the constitution and selection of corpus should not be limited to one area, but take into account the representativeness and balance of the selected corpus in different domains of language. (3) Principle of dynamic. The large amount of data in the corpus needs to be updated constantly. Although a corpus collects a massive data, it cannot cover all the language content. Languages develop with the development of society and are dynamic. (4) Principle of openness. Corpus itself is an open system, into which a variety of text resources and audio materials are filtered, and can interconnect with other corpora.

According to the needs of different courses, we can first construct small corpus for our teaching. For example, in order to explain the abstract translation in science and technology paper, teachers can in advance self-construct Science and Technology Abstract Corpus, which includes a certain number of Chinese and English abstracts from online China Knowledge Resource Integrated Database. The selected abstracts may be classified in accordance with the disciplines, and the right amount of original data for each discipline should be included to construct the Chinese Students' Science and Technology Abstracts Corpus. In addition, we can also self-construct Foreign Scholars' Science and Technology Abstracts corpus.

In Class, teachers could allow students to contrast the Chinese students' English abstracts to their Chinese abstracts, let them make genre analysis, and summarize language differences between English and Chinese abstracts, such as tense, voice and so on. In addition, by comparing the abstracts of Chinese students with foreign scholars, we can guide students to discover the errors usually made by the Chinese students in abstract writing, and promote teaching and academic development.

C. The Vocabulary Teaching Mode Based on DDL

According to the teaching objectives and content, teachers will select appropriate content from the corpus to produce the illustrated teaching classes, and design individualized teaching task in view of the characteristics and language skills of students. On the first stage, determine the problems to be solved and the solutions to the problem by using corpus. On the second stage, carry out classroom activities based on corpus to guide students to find answers to the questions independently and participate in classroom exchanges. On the third stage, assign some tasks to enable students to use corpus resources with target.

Teachers can fist give students the available corpus, and teach them how to use the retrieval tools for autonomic learning. Here are several teaching modes.

Mode 1: Before class, teachers need to extract the concordance lines of a word in the corpus by retrieval software, and in the class, show the data to the students by multimedia. Next, the students are asked to discuss the law behind the authentic language phenomena in groups in the classroom. If permitted, we can also enable students to independently use the retrieval software to extract and analyze more real data from the corpus. Finally, under the guidance of the teacher, students need to summarize the characteristics of collocation, grammar patterns, context and co-occurrence.

Mode 2: Before starting a new lesson, first teachers need to determine word to learn in this unit, and require students to produce "co-occurrence in context" of the word in groups with corpus to observe and induce language rules. In the class, each group exchanges their retrieval results, and then summarizes the rules of usage again. At the same time, teachers can lead students to contrast the results of discussion to that in the dictionary, and further find examples from the corpus for those usages which are not clear enough to strengthen the students' language awareness.

Mode 3: teachers can produce co-occurrence of a word from a corpus according to the difficult levels of the unit, and replace the word in the concordance lines with other unrelated alternative symbols. Students are asked to determine its part of speech with the context before and after the word, guess its meaning and summarize its usage.

D. The Orientation of Teachers' and Students' Role

There is great difference between the traditional teaching model and corpus-based data-driven teaching mode, the latter reflecting the principle of combining practicability, infotainment and interesting in English teaching in favor of mobilizing the enthusiasm of both teachers and students. This new teaching mode, in particular, reflects the dominant position of and the leading role of teachers in the teaching process.

Teaching space becomes open, and teachers' role changes from the traditional knowledge initiator to the organizers and instructors of the teaching process, the assistors and promoters of the construction of meaning. The status of students changes from passive recipients into active constructors of knowledge.

The most critical step in DDL is learners summarize the concordance lines to arrive at language rules under the guidance of teachers. DDL emphasizes the student-centered exploratory learning mode, which fundamentally alters the traditional top-down teaching mode. Students, who change to researchers, constantly summarize from a large number of language input. The learning process makes them get the satisfaction of success, enhance their self-confidence, and further stimulates their interest in learning. The questions brought up in DDL could be set by the teacher according to the syllabus, or could be created by learners in their summarization. Learners collect large amounts of authentic information from the "context" in corpus, breaking the traditional teaching mode in which learners only rely on reference books and dictionaries to learn a word, so that students can get a lot of authentic language input.

V. Conclusion

Foreign language teaching has always insisted on the innovation of teaching method. Corpus-based data-driven teaching and learning provide a large number of instances of real context, and create a learning environment to attract learners' attention, be conducive to enhance their memory and help them to use context to obtain the word semantics and summarize the grammatical rules. In DDL, the student-centered classroom design puts more emphasis on classroom interaction, in which students can communicate through their own understanding of the language knowledge inducted from the corpus, to achieve the purpose of the acquisition of language rules. This learning mode emphasizes the learner's autonomic learning ability to explore and discover language knowledge based on corpus according to their own needs so as to continuously introspect and induce language rules. DDL changes the dominant position of teachers in traditional teaching. Teachers change from initiators into the organizers and counselors of teaching. The relation between teachers and students has thus become more mutually cooperative.

References

[1] Higher Education Department of Education Ministry. (2004). College English curriculum requirements (For Trial Implementation). Shanghai: Shanghai Foreign Language Education Press.
[2] Higher Education Department of Education Ministry. (2007). College English curriculum Requirements. Beijing: Foreign Language Teaching and Research Press.
[3] He, A.P. (2004). Corpus linguistics and English language Teaching. Beijing: Foreign Language Teaching and Research Press.
[4] Yang, H. Z. (2002). An introduction to corpus linguistics. Shanghai: Shanghai Foreign Language Education Press.

[5] Zhen, F.C. (2005). Corpus-based data-driven foreign language learning: ideas, methods and techniques. *Foreign Language World*, 4, 19-27.
[6] Feng, Y. F. & Cai, L. (2009). A research-oriented approach to college English integrated reading course under multimedia environment. *Foreign Languages and Their Teaching*, 11, 28-31.
[7] Zhang, T. W. (2010). New perspectives and reflections of the reform in college English teaching—review on the integration of computer and networks into foreign language curriculum. *Computer-Assisted Foreign Language Education in China*, 135, 59-63.
[8] He, A.P. (2009). Sinclair's lexical grammar: interpretation and application. *Foreign Language Research*, 5, 52-57.
[9] Johns, T. & P. King. (1991). Classroom Concordancing. Birmingham: University of Birmingham.
[10] Hunston, S. (2002).Corpora in applied linguistics. Cambridge: Cambridge University Press.
[11] Johns, T. (1991). Should you be persuaded: two examples of data driven learning. *ELR Journal* (New Series), 4. 1-16.
[12] Sinclair J. (1991). Corpus, concordance, collocation. London: Oxford University Press.
[13] Wei, N.X. (2001). The definition and study of word collocation. Shanghai: Shanghai Jiao Tong University Press.
[14] Wei, N.X. (2002a). The approach of the study on Semantic prosody. *Foreign Language Teaching and Research*, 4, 300 - 307.
[15] Stubbs, M. (1996). Text and corpus analysis. Oxford: Blackwell Publisher.
[16] Wei, N.X. (2002b). A Corpus-driven study of semantic prosodies in specialized text. *Modern Foreign Languages*, 2, 165-175.

Xiaowei Guan was born in Shenyang, China in 1979. She received her PH.D. degree in computer application from Dalian University of Technology, China in 2009.

She is currently a lecturer in the School of Foreign Languages, Dalian University of Technology, Dalian, China. Her research interests include machine translation and natural language processing, E-C and C-E translation and contrast.