ISSN 1870-4069

Spoken English Learner Corpora

Olga Kolesnikova, Oscar-Arturo González-González

Instituto Politécnico Nacional, Escuela Superior de Cómputo,
Ciudad de México, Mexico

kolesolga@gmail.com, oscar.ar-56@hotmail.com

Abstract. In this paper we present a survey of some most significant spoken English learner corpora created up to date. Spoken learner corpora which include speech generated by learners are important in many areas of research and practice, in particular, for identifying typical pronunciation errors of learners of English as a second language (ESL), English as a foreign language (EFL), or English as a lingua franca (ELF). The data on common errors is helpful in designing more effective methods of pronunciation teaching as an aspect of language training. Also, error patterns can be implemented in intelligent tutor systems for English learning in order to design explanations and exercises in the error-preventive way and to generate a relevant feedback to the learner. The corpora we survey in this article include various types of English speech generated by learners with Arabic, Chinese, French, German, Greek, Japanese, Korean, Norwegian, Polish, Spanish, among others, as their first language (L1). Some English learner corpora described here are created for a single L1, other corpora are compiled for various first languages. Also, learner corpora vary depending on what type of English they exhibit: ESL, EFL, ELF or their combinations.

Keywords: Spoken English learner corpus, accented English speech, English as a second/foreign language, English as lingua franca, pronunciation errors.

1 Introduction

An English learner corpus is a collection of written and/or spoken texts produced by learners of English as a second language (ESL or English as L2), or English as a foreign language (EFL), or English as a lingua franca (ELF). Learner corpora are used by researchers, teachers of English, and learners for various purposes, one of them is error recognition and analysis. The results of error analysis can be applied in English language teaching in a conventional classroom environment as well as in computer assisted training or intelligent tutor systems.

Pronunciation is one of the major aspects of English training. For an efficient acquisition of English pronunciation, the knowledge of most common errors can aid in the development of adequate methods which will help the learner to understand how the English sounds are generated and recognized in speech. Another element of English teaching is the use of pronunciation exercises for speaking practice; such exercises can be designed in a more effective manner taking into account common error patterns in

order to prevent mispronunciations on the one hand, and on the other hand, correct them with drills targeted at the specific errors known in advance.

In order to successfully implement the data on learner errors, the latter must be identified, studied, classified, and formalized. Formalization of error patterns or rules is necessary for their application in intelligent tutor systems and various forms of online learning. Learner corpora include the systematized and usually annotated material necessary for researchers and education professionals to work on errors, therefore, the importance of such corpora can hardly be overestimated.

In this paper we present a survey of some most significant spoken English learner corpora collected for various first languages (L1), i.e., the speech recorded in such corpora is accented by different mother tongues of English learners. For instance, a Spanish L1 – English L2 spoken corpus contains texts spoken by English learners whose first language is Spanish. In our survey, we give description of spoken English corpora with various first languages: Arabic, Chinese, French, German, Greek, Japanese, Korean, Norwegian, Polish, Spanish, among others. Some English learner corpora described here are created for a single L1, other corpora are compiled for several first languages. Also, learner corpora vary depending on what type of English they exhibit: ESL, EFL, ELF, or their combinations.

Our descriptions of corpora are structured as follows. First, we give the title of the corpus, then we indicate the name of the coordinator/s and/or head/s responsible for the corpus compilation, in what institution or organization the corpus was created, the aim of the corpus, the subjects, themes, topics covered in texts, some details on speech recording, then how the recorded material was processed, transcribed, and annotated. The extension and completeness of the descriptions depend on information found in public domain to the best of our knowledge.

In some cases, a collection of spoken texts is called a database instead of corpus. The difference between a corpus and a database is that the latter is aimed to include a wide range of data types: from spontaneously generated texts (for example, speech in informal interviews or chats) to texts read by learners in a formal environment of a classroom or during a test or exam. Sometimes, such databases include not only speech but also written texts: essays, summaries, description, reports, etc.

There are English learner corpora which form a part of larger corpora: for example, a multi-layered learners' corpus called AixOx (Herment, Tortel, Bigi, Hirst, Loukina 2012) includes French native speakers reading French passages, English native speakers reading English passages, as well as French speakers reading English and English speakers reading French.

The rest of the paper is organized as follows. Sections from 2 to 12, arranged in the alphabetic order of the first languages of English learners, describe English learner corpora created for a particular single F1. Section 13 presents corpora compiled for multiple first languages, each corpus is described in a separate subsection arranged in the alphabetic order of the corpora titles.

2 Arabic

2.1 Qatar Learner Corpus

The Qatar learner corpus is an Arabic L1 – English L2 spoken corpus. It was created by Yun (Helen) Zhao, at Modern Language Department of the Carnegie Mellon University, USA. The recorded participants are ESL learners with Arabic, mostly Qatar, as their first language. The corpus includes spoken interviews of 19 participants. For each participant, the following metadata is available: first name, grade, nationality, gender, the rates of the participant's reading skills and language usage, as well as the average English rate, see Table 1. The corpus is freely available online[1]. The link to the Qatar learner corpus is located at the website of Université catholique de Louvain[2]. At this page, an extensive list of learner corpora for various first and second languages around the world can be found.

Table 1. Data on some participants recorded in the Qatar learner corpus.

Name	Grade	Nationality	Gender	Reading Skills	Language Usage	Average English
Sam	12	Qatari	Male	39.61	65.76	52.685
Abe	12	Qatari	Male	62.57	73.67	68.12
Charles	11	Qatari	Male	61.12	80.31	70.715
Tom	12	Qatari	Male	44.35	39.31	41.83
Larry	12	Qatari	Male	93.66	89.91	91.785
Bill	12	Qatari	Male	75.32	99	87.16
Jenny	12	Qatari	Female	77.63	97	87.315
Nancy	12	Qatari	Female	96.81	99	97.905
Lucy	12	Qatari	Female	99	97.2	98.1
Anne	12	Qatari	Female	94.54	87.75	91.145
Alice	11	Qatari	Female	78.46	98.8	88.63
Paula	11	Qatari	Female	53.06	57.4	55.23
Pat	12	Qatari	Female	53.37	84.62	68.995
Tina	12	Qatari	Female	79.65	89.32	84.485
Linda	11	Qatari	Female	66.95	90.51	78.73
Donna	11	Kuwaiti	Female	71.32	91.1	81.21

3 Chinese

The interest of Chinese researches and educationalists for the English language acquisition is quite big as evidenced by many spoken English learner corpora created for this first language. In this section we give a brief description of most significant of them.

[1] http://talkbank.org/ data/SLABank/English/
[2] https://www. uclouvain.be/en-cecl-lcworld.html/

3.1 BICCEL

BICCEL stands for the Bilingual Corpus of Chinese English Learners. It is a spoken corpus which contains the recordings of speech produced at the National Oral English Test by forth year students majoring in English. The collected material spans the years from 2001 to 2005 and includes 1,100 test participants. This corpus also contains written texts which are in-class assignments. The work on the BICCEL was supervised by Prof. Wen Qiufang and done by Dr. Wang Jinquan at the National Research Center for Foreign Language Education, Beijing Foreign Studies University, China.

3.2 CUCASE

CUCASE is the City University Corpus of Academic Spoken English compiled by David Yong Wey Lee at the City University of Hong Kong. It is a multimedia collection of learner speech, and it also includes data produced by native English speakers.

3.3 COLSEC

COLSEC is the acronym for the College Learners' Spoken English Corpus. This corpus consists of the transcribed speech of non-English university majors produced at the National Spoken English Test. This corpus was created by Yang and Wei (2005). The COLSEC was used in the work of Luo, Yang, and Wang (2011) to define mispronunciation rules. These rules and the statistics of mispronunciations observed in the COLSEC allowed the authors to construct pronunciation lexicons in which prior probabilities was indicated. Prior probabilities reflect how likely each type of error might occur as language models for automatic speech recognition (ASR) systems and applications.

3.4 ESCCL

ESCCL is the English Speech Corpus of Chinese Learners. It contains dialogues read aloud, and it was created by Chen Hua from Nantong University, Wen Qiufang from the Beijing Foreign Studies University, and Li Aijun from the Chinese Academy of Social Sciences (Hua, Qiufang, Aijun 2008). The recorded speech was produced by participants at four different educational backgrounds. They were asked to complete two tasks: to read a dialogue aloud and to produce a spontaneous dialogue on a given topic. The recordings were collected in different parts of China and in its ten major dialectal areas. The authors claim that the quality of their recording is higher than the recordings stored in other corpora: the recording for the ESCCL was done in language laboratories by MP3- H06 at the sample rate of 16,000 (16 kHz, 16 bit mono PCM).

The recorded participants cover almost all learners under formal classroom instruction, with an interval of three years between adjacent groups. In each part of China (also in each dialectal district), at least 30 junior middle school students, 30 senior middle school students, 30 college English majors, and 30 English majors for Master degree were willingly recorded. In each group, the number of male and female students was well balanced.

The corpus also includes prosody annotations using both British and American annotation systems. The annotations were made on the computer with Praat software[3] (Boersma 2002) by 15 college English researchers. All the data were cross-checked by three phoneticians in China.

3.5 SWECCL

SWECCL stands for the Spoken and Written English Corpus of Chinese Learners. As its title shows, the corpus includes a spoken part, entitled as the Spoken English Corpus of Chinese learners (SECCL), and a written part, called the Written English Corpus of Chinese learners (WECCL). The SWECCL was compiled by Wei Qiufang, Wang Lifei, and Liang Maocheng (Wen, Wang, Liang 2005). Two versions of this corpus have been created up to date: the first version was completed in 2005, and the work on the second version was finished in 2007. The spoken section (SECCL) in its second version includes speech produced at the National Spoken English Test: in the years 2003-2006 by second-year English majors, and in the years 2000-2006 by forth-year English majors. The corpus also comprises longitudinal data of 40 hours of speech within the years 2000-2004.

3.6 TSLC

TSLC is the TELEC Secondary Learner Corpus, which includes written and spoken English Chinese (Allan 2002). TELEC stands for the Teachers of English Language Education Center. This institution maintains a computer network called TeleNex designed to provide support for language teachers in Hong Kong. The work on the TSLC corpus was done under TELEC auspices. The corpus was developed within a number of years beginning from 1994 and comprises now ten million words of running text, mostly written compositions; however, it also includes a small spoken part. The work on the corpus was led by Quentin Allan, the University of Hong Kong.

The TSLC has been used primarily for pedagogical purposes (Allan 1999). The TELEC staff members use the TSLC for developing teaching materials, designing lessons that address common problem areas, and answering questions asked by teachers on TeleNex webpage[4]. Through TeleNex, teachers who do not have time or expertise to carry out their own corpus investigations can still enjoy the benefits of learner corpora research.

4 French

4.1 ANGLISH

The ANGLISH database (Tortel 2008) was created for British English as L2. It includes L1 as well as L2 French speakers. The recording was done in an anechoic room, and 63 participants were recorded while reading and repeating texts. The reading part

[3] http://www.praat.org/
[4] http://www.telenex.hku.hk/telec/

includes 1,260 utterances. Continuous unprepared speech was also recorded. The participants included native speakers of British English (23 speakers: 13 female and 10 male), non-specialist working adult French speakers of English (20 learners: 10 female, 10 male), and second-year university French students of English (20 learners: 10 female, 10 male).

The recordings of the reading part of the corpus were manually segmented into phonemes and labeled with CVC codes using the Praat software[5] (Boersma 2002). The corpus is freely available on SLDR (Speech and Language Data Repository[6],) and its description can also be found at the website of Université catholique de Louvain[7].

4.2 AixOx

AixOx, a multi-layered learners' corpus, (Herment, Tortel, Bigi, Hirst, Loukina 2012) includes French native speakers reading French passages, English native speakers reading English passages, as well as French speakers reading English, and English speakers reading French.

4.3 Learners' Corpus of Reading Texts

This learner corpus includes unprepared reading of English texts. The texts are short abstracts of fiction or made-up dialogues. The corpus was compiled by Sophie Herment, Valérie Kerfelec, Laetitia Leonarduzzi, and Gabor Turcsan from Laboratoire parole et langage, Aix Marseille Université in Aix-en-Provence, France. The 54 recorded participants were first-year French students of the English Department at the above mentioned university. The corpus is accessible online[8] and can be freely downloaded.

4.4 CoNNECT

CoNNECT is the Corpus of Native and Non-native EFL Classroom Teacher Talk. It contains transcripts of native and non-native English lesson audio recordings performed in a secondary classroom of students ranging from A1 to B2 levels. The data was collected within the period from January 2009 till March 2011. The recordings were made in French-speaking Belgium and in Britain. The CoNNECT includes two sub-corpora: the native English sub-corpus with 108,988 words and non-native English sub-corpus with 56,526 words.

The native English recordings include 24 lessons, and the non-native English recordings include 14 lessons. The corpus has been used to analyze the linguistic features of native-speaker teachers' classroom language that could be useful to non-native foreign language teachers. Its native English part can also serve as a baseline for comparison with the non-native sub-corpus.

[5] http://www.praat.org/
[6] http://sldr.org/, http://sldr.org/sldr000731/
[7] https://www.uclouvain.be/en-cecl-lcworld.html/
[8] http://sldr.ortolang.fr/voir_depot.php?lang=en&id=15&allpreview=1/

The CoNNECT has been transcribed according to the guidelines used for the Louvain International Database of Spoken English[9] (LINDSEI).

The main objective of the CoNNECT (Meunier 2016) was to study native versus non-native teachers' speech, in particular, lexical choices and prosody. It was found in the analysis of the corpus that native teachers use prosody as a strategic pedagogical tool: they employ rising intonation to draw the learners' attention and longer pauses to prompt learners' feedback and reactions. Another result of the research is findings with respect to rephrasing strategies of teachers. Native teachers rephrase guidelines, feedback, task descriptions more often that non-native teachers do, and the former also use several types of rephrasing for the same turn, sometimes up to three variants. Non-native teachers tend to use full forms of terms in spoken interactions, for instance, *is not, could not, is going to*, while native teachers use contracted forms instead.

5 German

5.1 LeaP

The LeaP corpus (LeaP stands for Learning Prosody in a Foreign Language) by Milde and Gut (2002), see also (Gut 2004, 2012), consists of two sub-corpora: the first sub-corpus includes recordings of ESL learners, and the second one contains the speech of learners of German as a second language. The corpus is available to the scientific community[10] and can be downloaded upon the request to the authors.

The LeaP corpus was collected within the frame of the LeaP project, which was led by Ulrike Gut at the University of Bielefeld, Germany, within the time period from 2001 to 2003. The aim of the LeaP project was related to the acquisition of prosody by non-native speakers of German and English; therefore, the researchers were concerned with phonetic and phonological description of non-native prosody and exploration of learner variables that influence the language acquisition process.

The corpus data covers a wide range of speakers in terms of age, sex, native languages, level of competence, length of exposure to the target language, age at first exposure to the target language, and non-linguistic factors such as motivation to learn the language, musicality, and so forth. The age of the non-native speakers at the time of the recording ranged from 21 to 60. The data was collected from different groups of speakers: learners before and after a period abroad, before and after a four-month prosody training course, learners with different levels of competence; special attention was given to advanced learners who are hardly distinguishable from native speakers.

Four types of speech styles were recorded: nonsense word lists, readings of a short story (about 2 minutes), retellings of the story (between 2 and 10 minutes), free speech in an interview situation (between 10 and 30 minutes) The recordings were annotated manually and automatically on eight different tiers including pitch, tones, segments, syllables, words, phrasing, parts-of speech, and lemmas. The entire corpus consists of

[9] http://www.uclouvain.be/en-307849.html/

[10] http://www.philhist.uni-augsburg.de/de/lehrstuehle/anglistik/applied-inguistics/Forschung/leap/

359 annotated files, includes a total of 131 speakers, and the total amount of recording time is more than 12 hours.

Many research works have been performed on this corpus. For example, one the works (Carson-Berndsen, Gut, Kelly 2006) discovered regularities in non-native speech which can be used in a variety of pedagogical activities as well as in computer assisted training and automatic speech recognition.

5.2 GLBCC

GLBCC stands for the Giessen-Long Beach Chaplin Corpus, it includes transcribed interactions between native English speakers, ESL and EFL speakers. The corpus was compiled by Andreas Jucker and Sara Smith at the University of Giessen, Germany. The corpus can be accessed online[11] and downloaded upon a request to the authors.

In the process of corpus creation, pairs of students, in California (for English as native and second language) and in Giessen (for English as foreign language), participated in the experiment. They were asked to watch the first part of a silent Charlie Chaplin movie. One participant, called speaker A, was then asked to retell in a monologue what he/she had seen so far, while the other participant, called speaker B, watched the rest of the movie and told his/her partner the second part of the movie. In the end of the conversation, the two participants discussed several aspects of the movie on the basis of a few written prompts.

In the process of corpus compilation, 108 sessions were recorded involving 191 speakers. There were 83 A-speakers, 90 B-speakers, and altogether, the corpus comprises 35 American, 4 British, and 2 Australian native speakers. 77 non-native speakers are Germans, the others have a variety of linguistic backgrounds, including Hispanic, Japanese, and Korean.

6 Greek

6.1 YoLeCorE

YoLeCorE stands for the Young Learner Corpus of English, it is an English Greek spoken pedagogic corpus of video-recorded EFL classes. The corpus was created by Marina Mattheoudakis and Thomas Zapounidis at the Aristotle University of Thessaloniki, Greece (Mattheoudakis 2014). This audiovisual written and spoken corpus was compiled at the Third Model Experimental primary school in Evosmos, and it is an innovative pedagogic corpus which includes all language instances produced in a class of 8-9-year-old learners during one school year.

[11] http://ota.oucs.ox.ac.uk/headers/2506.xml/

7 Japanese

7.1 NICT JLE

NICT JLE corpus is the Japanese Learner English (JLE) corpus compiled at the National Institute of Information and Communications Technology (NICT) in 2004 by Emi Izumi, Kiyotaka Uchimoto, and Hitoshi Isahara, Kyoto, Japan. The corpus data includes the transcripts of audio-recorded speech samples of English oral proficiency interview test which is called ACTFL-ALC SST (Standard Speaking Test). The corpus contains 1,281 samples, 1.2 million words, 300 hours in total. This corpus is available online[12] and can be freely downloaded.

The metadata includes the proficiency level of the participants (9 levels) based on the SST scoring method; this makes it possible to easily analyze and compare the characteristics of interlanguage of each developmental stage. This is one of the advantages of the NICT JLE corpus.

The corpus is annotated with more than 30 basic tags for all files and with error tags for 167 files. The basic tags include tags for representing the structure of the interview, tags for the interviewee's profile, tags for speaker turns, and tags for representing utterance phenomena such as fillers, repetitions, self-corrections, overlapping, etc. The error tag set includes 47 tags of lexical and grammatical errors of learners.

In order to compare native and non-native English speech, the NICT JLE corpus includes a native English speakers' sub-corpus.

8 Korean

8.1 ETC

ETC stands for English (as a foreign language) Teacher Corpus, it contains teacher talks in language classrooms and includes a total of 247,398 words compiled through 62 hours of recording of EFL classes. The corpus was created by Ye-Eun Kwon and Eun-Joo Lee (Kwon, Lee 2014).

The EFL corpus consists of two sub-corpora: first, the teacher talk was collected from four Korean EFL teachers, and second, the speech recordings of five native university EFL instructors were made. All the non-native English teachers were teaching general English classes at three different universities in Seoul at the time of data collection. For comparative purposes, five native teachers who were teaching at two different universities in Seoul participated in the study. The ETC two sub-corpora are called non-native EFL teacher (NNET) corpus and native English teacher (NET) corpus. The NNET corpus includes 123,122 words, and the NET corpus contains 124,275 words.

The EFL teachers' age ranged from the late twenties to mid-forties. The Korean EFL teachers held graduate level degrees in English education or general education while the native EFL teachers had graduate level degrees in fields other than English or education. The participants' teaching experience ranged from three to nine years, and

[12] http://alaginrc.nict.go.jp/nict_jle/index_E.html/

the Korean EFL teachers had on average slightly less experience teaching English than native EFL teachers.

8.2 NICKLE

NICKLE stands for the Neungyule Interlanguage Corpus of Korean Learners of English. It includes the written part and the spoken part. The spoken part consists of student interviews and transcriptions of oral speech tests. The corpus was created by Ji-Myung Choi at the Yonsei University, Seoul, Korea.

9 Norwegian

9.1 EVA

The EVA Corpus of Norwegian school English was compiled as a part of the government-sponsored EVA Project (Evaluation of English in Norwegian schools), with Angela Hasselgren as the project leader and Anna-Brita Stenström as its advisor, both from the English Department at the University of Bergen, Norway. The recorded speakers were Norwegian pupils of 14–15 years.

The corpus consists of the transcripts of 62 pupils taking the EVA 8th grade oral test. This test includes three picture-based tasks: the first task involves describing, narrating, and discussing, the second task involves giving instructions and checking for understanding, and the third task consists in a role play, with one role fixed, i.e. read by the pupil.

The main part of the corpus includes about 35,000 words. Together with it, a smaller control corpus was compiled with 16 native British teenager speakers carrying out the same tasks.

10 Polish

10.1 PLEC

PLEC is the PELCRA Learner English Corpus created by Piotr Pęzik, Barbara Lewandowska-Tomaszczyk, University of Łódź, Poland[13]. PELCRA stands for Polish and English Language Corpora for Research and Applications, and it is also the name of a research group at the Department of English Language at the University of Łódź.

PLEC includes the written part and the spoken part. The corpus makes it possible to analyze many aspects of the phonetic, lexical, grammatical, and phraseological competence of Polish learners of English using quantitative and qualitative methods.

The corpus contains time-aligned interviews and other spoken interactions of Polish learners of English. The transcriptions of the corpus include manual annotations of mispronounced words, this permits researchers to study the relative frequency of word mispronunciations as well as possible patterns among them. The results of research on

[13] http://pelcra.pl/plec/ research/

pronunciation errors can be used to prioritize certain lexical items in pronunciation courses and thus they will help to develop syllabuses and materials in teaching English to Polish-speaking students.

11 Spanish and Catalan

11.1 BELC

BELC stands for the Barcelona English Language Corpus (Muñoz 2006); it is a written and spoken corpus created with the objective to do research on how age affects the acquisition of English as a foreign language. BELC was compiled by the research group Grup de Recerca en Adquisició de Llengües (GRAL) in the Department of English at the University of Barcelona. The coordinator of the research group is Dr. Carmen Muñoz Lahoz, other details and information can be found at the website of the University of Barcelona[14].

The recorded participants were 2,063 students from state schools in Catalonia, Spain. As they are residents of Catalonia, they are bilingual: their native languages are Spanish and Catalan.

According to the objective of the corpus, students of various age groups were recorded. Alongside with the age factor, another parameter was taken into account, namely, the number of hours a student passed learning English as a second language. So the students were recorded after having 200, 416, 726, and 826 hours of English language instruction in schools.

The written part of the corpus includes compositions dealt with a familiar topic: *Me: my past, present and future*, it was the first task completed by the subjects.

The spoken part of the corpus includes three other tasks performed by the subjects: first, oral narrative prompted by six pictures at which the subjects were to freely look before and during their spontaneous talk; second, oral semi-guided interview which began with a warm-up in the form of questions about the subject's family, daily life, and hobbies and then continued as a spontaneous talk on any other topics initiated by the interviewer as well as by the subject; third, role-play performed in randomly chosen pairs: one of the students played the role of the mother/father, and the second student, the role of the son/daughter. The latter had to ask permission to have a party at home and both role play partners had to negotiate the setting, time, music, eating, drinking, and any other activities and details.

The BELC data consists of the recordings of those students who could be followed longitudinally and for whom there are two, three, or four collection times over a period of seven years. However, not all subjects performed all the four tasks.

BELC was updated in 2014: its spoken part was expanded by adding more recordings of oral narratives which were also transcribed, and to its written part, more compositions were aggregated. Other details on BELC can be found online[15].

[14] http://www.ub.edu/web/ub/en/recerca_innovacio/recerca_a_la_UB/grups/fitxa/G/ADQULENG/equipInvestigador/index.html/

[15] https://www.uclouvain.be/en-cecl-lcworld.html/

11.2 SULEC

SULEC is the Santiago University Learner of English Corpus which includes written and spoken data. The written part contains compositions and argumentative essays, while the spoken section includes semi-structured interviews, short oral presentations, and brief story descriptions. The corpus was created starting in 2002 by a research group at the University of Santiago de Compostela led by Ignacio M. Palacios Martínez, see (Palacios Martínez 2005).

12 Taiwanese

12.1 LTTC English Learner Corpus

The LTTC English Learner Corpus consists of language samples produced by Taiwanese learners of English. LTTC stands for the Language Teaching and Testing Center. This center in cooperation with another institution, Graduate Institute of Linguistics (GIL), started the work on the corpus in 2007. The project director is Prof. Hintat Cheung from GIL and the co-directors are Dr. Zhao-Ming Gao from the Department of Foreign Languages at the National Taiwan University and Dr. Siaw-Fong Chung from the Department of English at National Chengchi University.

The participants were English learners who took the General English Proficiency Test (GEPT), a language proficiency examination developed and applied by the LTTC. The corpus includes 2,000 written samples and also includes a spoken section with 400 speech samples; both written and oral samples were collected at the Intermediate GEPT examination. For each participant the following metadata is given: the region of Taiwan (North, East, etc.) where the test was taken, the age, gender, education level of the test-taker, his/her major if the test-taker was a college graduate, whether the test-taker was a student or not, and whether he or she had lived in an English-speaking country, and if so, for how long.

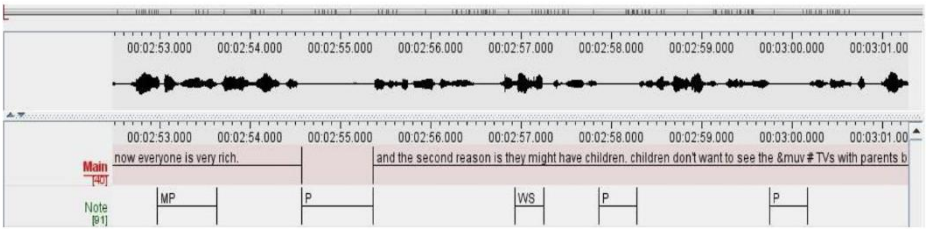

Figure 1. Soundwave and transcription of spoken data segment from the LTTC English Learner corpus.

The spoken data was first recorded on cassette tapes, then digitized and transcribed using the software ELAN (EUDICO Linguistic Annotator) (Hellwig, van Uytvanck, Hulsbosch 2008) as well as tagged using the CHAT (CHILDES) format (MacWhinney 2008). Tags were added inside the body of the transcriptions for repetitions, self-corrections, incomprehensible sounds, lengthened vowels, and other characteristics. Tags for filled and unfilled pauses, mispronunciations, and word stress errors were

added in the tier below that of the main transcription. Figure 1 presents an example of a soundwave and the corresponding transcription of the data after its processing with the ELAN software.

13 Various First Languages

13.1 ITAcorp

ITAcorp is the International Teaching Assistants corpus. International teaching assistants (ITA) are international graduate students employed usually by North American universities especially in the areas of engineering, mathematics, and sciences. ITAs participate in such activities as grading tests for large lectures, teaching break-out discussion sessions, and conducting office hours. The latter activity involve ITAs as tutors of undergraduates on homework problems, preparing them for tests, and answering questions on behalf of a supervising professor.

In practice, it turned out that many ITAs are not prepared for their roles in these sessions. Therefore, they are suggested to take advanced ESL for academic purposes and ITA preparation courses to improve their posterior performance as teaching assistants. The goal of these preparation courses is two-fold: to give instruction in English language usage in the ITAs' activities context and to teach pedagogy.

The ITAcorp consists of transcribed recordings of such preparation courses at a large Northeastern American university, which included different classroom activities and computer-mediated activities (chats): classroom discussions, lecture preparation, question answering, concept presentations, and office hours role plays. The corpus was started in 2005, a detailed description of its creation is given in (Reinhardt 2007). The work on corpus creation was done by Steven L. Thorne, Paula Golombek, and Jonathon Reinhardt at the Pennsylvania State University, USA (Thorne, Reinhardt, Golombek 2008; Reinhardt 2010).

The purpose of the corpus was to inform instruction in advanced spoken English for academic purposes and to do research on intercultural pragmatics and sociolinguistic issues.

Chinese, Thai, Korean, and other L1 English learners were recorded. The sub-corpus of office hours role plays includes approximately 103,000 tokens. In these role plays, the students played the ITA and student roles, and also, the ITA and an evaluator roles in a post-semester evaluation. Each role play is about 4 minutes long, the context of each role play is a student approaching an ITA with a typical problem that would need to be negotiated.

13.2 LONGDALE

LONGDALE is the Longitudinal Database of Learner English at the University of Louvain, Belgium. The project director is Fanny Meunier, and the team also includes Sylviane Granger, Damien Littré, and Magali Paquot.

The objective of this project is to construct a Learner English longitudinal database from learners of various L1s. These learners are followed over a period of at least three

years of their studies. Up to date the data was collected within 2008-2009-2010 and 2010-2011-2012 periods including only written data from German, French, Italian, Dutch, Turkish, and Brazilian learners[16]. The project team also plans to record interviews but this work is still in progress.

The LONGDALE also includes recordings of English learners of French (EN_FR), but only of young children. In this respect, this corpus is similar to FLLOC (French Learner Language Oral Corpus, see Myles and Mitchell 2007), in which the children are aged 7 to 11, or to CYLIL (The Corpus of Young Learner Interlanguage, see Housen 2002), which contains English L2 recordings of school pupils of different European nationalities, French being one of them.

13.3 LINDSEI

LINDSEI is the Louvain International Database of Spoken English Interlanguage (Gilquin, De Cook, Granger 2010). It is a collaborative project between several universities internationally, coordinated at the University of Louvain, Belgium. Started in 1995, this database now includes 21 sub-corpora, of which 14 are complete (Bulgarian, Chinese, Dutch, French, German, Greek, Italian, Japanese, Polish, 2 Spanish sub-corpora, Swedish, Taiwanese, Turkish), and seven are in progress (Arabic of Saudi Arabia, Basque, Brazilian Portuguese, Czech, Finnish, Lithuanian, Norwegian)[17].

The LINDSEI corpus is offered online[18] on CD-ROM containing over 1 million words, of which almost 800,000 were produced by learners, representing 11 different mother tongue backgrounds: Bulgarian, Chinese, Dutch, French, German, Greek, Italian, Japanese, Polish, Spanish, and Swedish. The corpora include 525 interviews and each interview consists of three tasks: first, a warm-up, in which learners were given a few minutes to talk about one of three set topics, second, a free informal discussion as the main part of the interview, and third, a picture description. The interviews are transcribed according to the transcription guidelines which can be found at the website of the University of Louvain[19].

13.4 MAELC

MAELC is the Multimedia Adult ESL Learner Corpus started in 2001 by Stephen Reder, Kathryn Harris, and Kristen Setzler of the Department of Applied Linguistics in Portland State University, USA, together with Portland Community College which provides adult ESL courses (Reder, Harris, Setzler 2003).

The corpus is a database of videos of classroom activities from four years of adult ESL classes from beginning to upper-intermediate proficiency. More than 3,600 hours of classroom interaction were recorded by six cameras and multiple wireless microphones.

[16] According to http://www.uclouvain.be/en-314347.html/ as of July 7, 2016.
[17] This data on sub-corpora was retrieved on July 7, 2016 from http://www.uclouvain.be/en-307845.html/.
[18] http://www.i6doc.com/en/collections/ cdlindsei/
[19] http://www.uclouvain.be/en-307849.html/

By now, about 150 hours of student language have been transcribed, this includes language of 250 low-level students with known background characteristics. The corpus also includes scanned copies of classroom written materials, student work, teacher logs, and teacher reflections. Many students were recorded several times per term and often in consecutive terms in different levels, which allows for longitudinal study of ESL acquisition.

The corpus was created with the purpose to do research on diverse ESL acquisition issues including longitudinal studies as mentioned above, in-depth case studies of adult learners of English, close examinations of dyadic and small-group interactions, which can focus on interactions between students from different L1 backgrounds, developmental studies of individual students who were recorded throughout several terms of study, among other themes.

13.5 T2K-SWAL Corpus

The T2K-SWAL Corpus is the TOEFL 2000 Spoken and Written Academic Language Corpus (Biber, Conrad, Reppen, Byrd, Helt, 2002; Biber, Conrad, Reppen, Byrd, Helt, Clark, Cortes, Csomay, Urzua 2004). It was created with the purpose to do an empiric study of texts used on listening and reading exams and determine if they accurately represent the linguistic characteristics of spoken and written academic registers, i.e., to diagnose the representativeness of English as a Second Language/English as a Foreign Language (ESL/EFL) materials and assessment instruments and see if it corresponds to real-life language usage.

Table 2. Data on T2K-SWAL corpus.

Register	Number of texts	Number of words
Spoken:		
Class sessions	176	1,248,800
Classroom management	40	39,300
Labs/in-class groups	17	88,200
Office hours	11	50,400
Study groups	25	141,100
Service encounters	22	97,700
Total speech:	**291**	**1,665,500**
Written:		
Textbooks	87	760,600
Course packs	27	107,200
Course management	21	52,400
Institutional writing	37	151,500
Total writing:	**172**	**1,071,700**
Total corpus:	**423**	**2,737,200**

The corpus includes 2.7 million words and is representative of the range of spoken and written registers that students encounter in U.S. universities. The data in the corpus includes written and spoken texts associated with academic life, including classroom

teaching, office hours, study groups, on-campus service encounters, textbooks, course packs, and institutional written materials (e.g., university catalogs, brochures), see Table 2.

Spoken texts were transcribed using a consistent transcription convention (see Edwards, Lampert 1993), then speakers were distinguished to a possible extent and some demographic information was added for each speaker (their status as instructor or student, etc.). The texts were annotated using various grammar tags, for details of annotation see (Biber et al. 2004).

13.6 ICNALE

ICNALE stands for the International Corpus Network of Asian Learners of English. It is a corpus of controlled speech and essays produced by learners of English in ten Asian countries and areas. For performing comparative studies, the corpus also includes speech of English native speakers. The project director is Shin'ichiro Ishikawa, Kobe University, Japan (Ishikawa 2011, 2012, 2013, 2014). ICNALE, as many other learner corpora, includes both written and spoken sections. The corpus can be accessed freely at its web page[20].

The spoken part of ICNALE contain the recordings of participants performing the following tasks: first, they respond to the personal attribute questions; second, they respond to the learning history questions; third, they answer the vocabulary size test, and forth, they take a telephone interview and respond to the questions concerning student's name, country, college, self-introduction (60 sec. speech), then each participant was asked to speak on some topics defined in advance, and the speech was recorded according in the following modes: Topic 1, Trial 1 (60 sec. speech after 20 sec. preparation), Topic 1, Trial 2 (60 sec. speech after 10 sec. preparation), Topic 2, Trial 1 (60 sec. speech after 20 sec. preparation), Topic 2, Trial 2 (60 sec. speech after 10 sec. preparation), then self-evaluation (0 to 5 points).

The questions and tests for the first three tasks were taken from the data collection sheet which can be obtained upon the request to the author of the corpus, as well as the topics used for the forth task.

The recorded participants were learners of English from the following countries: Hong Kong, Pakistan, Philippines, Singapore, Indonesia, and Thailand. For each participant the following data is included in the corpus: country or area of origin, sex, age, school grade (1, 2, 3, 4 ...), major or occupation; in case of students, their major at colleges; in case of employed people, their job.

The following academic genres were employed only for students: humanities, social sciences, science and technology, and life science. Information on the participants' proficiency test also included test name (TOEIC, TOEFL, etc.), the score in the above test; information on participants' motivation (using the scale from 1 to 6 points): integrative or instrumental motivation, strength of motivation, and the integrative motivation orientation score.

Also, information on the participants' English learning experiences (using the scale from 1 to 6 points): how much a participant studied English in their primary school days, in their secondary school days, in their college days, how much a participant

[20] http://language.sakura.ne.jp/icnale/index.html/

studied English in class, outside class, namely, at home, in the community, etc., how much a participant studied listening, reading, speaking, writing, how much a participant has been taught by English native participant, how much a participant has been taught pronunciation, presentation, essay writing.

13.7 CYLIL

The Corpus of Young Learner Interlanguage (CYLIL) contains English L2 data elicited from European School pupils and recorded at different levels of the participants' development including a longitudinal dataset of 6 learners followed during the period of three years as well as speech produced by other 40 learners. The corpus also includes the speech of eight native English children produced on the same tasks performed by other English learners recorded for the corpus; this serves as a baseline for comparative studies. The participants' L1 background was one of the following: Dutch, French, Greek, or Italian. In total, the corpus currently amounts to 500,000 words.

The creation of the corpus started in 1990 at Vrije Universiteit Brussel, Belgium, by Alex Housen (Housen 2002). The oral interviews consisted of informal free conversation and semi-guided speech tasks. Informants were asked to talk about events in their past, to describe pictures, to share opinions about movies they had seen, and to retell three picture stories with a variety of characters and actions.

The recorded speech was transcribed, segmented, coded, and annotated in CHAT format. The latter permits researchers to perform computer-aided analysis using the CLAN software. CHAT stands for Codes for the Human Analysis of Transcripts and CLAN is the acronym of Computerized Language Analysis. These too toolkits were designed for studying language learner speech by the CHILDES organization.

The purpose of the CHILDES system (Child Language Data Exchange System) is the study of child language and first language acquisition. However, it has been used also to do research on second language acquisition (SLA), speech pathologies, and discourse.

The CHILDES system[21] also contains electronically available corpora on child language, interlanguage, bilingual speech, and speech disorders (MacWhinney 2000). Access to the CHILDES database, the CHAT conventions, and the CLAN software is free. In its turn, CHILDES is a component of the TalkBank database[22] where many freely accessible corpora can be found.

13.8 ISLE

The ISLE speech corpus (Menzel, Atwell, Bonaventura, Herron, Howarth, Morton, Souter 2000) was created with the objective to implement a speech recognition method based on Hidden Markov Model in a computer assisted environment for teaching English at the intermediate level. The acronym ISLE stands for Interactive Spoken Language Education.

[21] http://childes.psy.cmu.edu
[22] http://talkbank.org/

Table 3. Data on the ISLE corpus.

Corpus section	Number of sentences	Linguistic issue	Exercise type	Examples
A B C	27 33 22	Wide vocabulary coverage (410)	Adaptation/ reading	"In 1952 a Swiss expedition was sent and two of the men reached a point only three hundred metres from the top before they had to turn back."
D	81	Problem phones, weak forms	Minimal pair item selection/ combination	"I said bad not bed." "She's wearing a brown wooly hat and the red scarf."
E	63	Stress, weak forms, problem phones, consonant clusters	Reading	"The convict expressed anger at the sentence." "The jury took two days to convict him."
F	10	Weak forms, problem phones	Description/Item selection/combination	"I would like chicken with fried potatoes, broccoli, peas and a glass of water."
G	11	Weak forms, problem phones	Description/Item selection/combination	"This year I'd like to visit Rome for a few days."

The corpus includes recorded speech of different types: reading simple sentences, pronouncing minimal pairs, giving answers to multiple choice questions by selecting an item from a list of options or/and combining items from different selections. English learners recorded for the corpus had German (23 learners) or Italian (23 learners) as their first language. English was learnt in its British variant.

The produced speech was recorded directly into WAV format, using a sampling rate of 16 kHz at a resolution of 16 bits. Some examples of the data included in the ISLE corpus are presented in Table 3.

Since the aim of this corpus is more specific in comparison with other more general purpose learner spoken corpora, the ISLE corpus was compiled to be applied as a tool in order to train the parameters and rules used in the recognition and diagnosis systems, to test the performance of the system on a known dataset, and to evaluate the contribution of speaker adaptation for improving the reliability of the native British English recognizer.

Taking the above mentioned tasks into account, the ISLE corpus had to be annotated at multiple levels: the word level, the phone level, and the stress level. This was necessary for determining the pronunciation errors (for instance, phone realization problems and misplaced word stress assignments, see examples in Table 4 and Table 5). Also, the corpus had to include various types of speech because the actual system for teaching EFL/ESL usually includes exercises of various grades of complexity (elementary, intermediate, advanced exercises).

Table 4. Examples of phone level errors from the ISLE corpus.

German			Italian		
from	**to**	**Example**	**from**	**to**	**Example**
oh	ow	pr**o**duce	eh	ey	s**ai**d
ax	ao	c**u**pboard	eh	ae	b**e**d
uw	ao	pn**eu**matic	ae	ey	pl**a**nning
aw	ow	**ou**tside	ih	iy	t**i**cket
aa	ae	st**a**ff	ay	iy	
ih	iy	d**e**ssert	oh	ow	b**io**log**i**c**a**l
-	p	**p**neumatic	ih	iy	
s	z	**s**aid	ax	ae	
v	w	**v**isa	-	ax	sheep
w	v	**w**eekend	-	hh	**h**onest
dh	d	**th**e	th	t	**th**in
-	w	bisc**u**it	s	z	**s**leep
-	b	thum**b**	jh	g	**g**in**g**er
g	-	fin**g**er	t	-	bai**t**
t	-	desser**t**			

Table 5. Examples of stress level errors from the ISLE corpus.

German	Italian
´report	´photographic
´television	´convict/con´vict
´contrast/contr´ast	´components

The ISLE corpus includes almost 18 hours of annotated speech and is based on 250 utterances selected from typical second language learning exercises. The ultimate purpose of the corpus usage is production of a relevant detailed feedback to English learners based on detection of their errors as well as selection targeted pronunciation exercises for error correction and further practice. The researchers who created and used the ISLE corpus on the tasks described above claim that the results of their work can be applied to any L1 (Menzel et al. 2000).

13.9 TCEEE

The Tübingen Corpus of Eastern European English (TCEEE) includes spontaneous spoken production data obtained by means of a semi-structured interview (Salakhyan 2012). The TCEEE was compiled by Elena Salakhyan at the Eberhard Karls University of Tübingen, Germany. The native languages of the participants were Russian, Ukrainian, Polish, and Slovak. The corpus includes a total of 60,000 words.

The corpus was created with the objective to study the Eastern European variety of English as one of the World Englishes (Berns 2005; House 2002; Jenkins 2007). However, since English speech was produced by non-native speakers of English, it can be viewed as an English learner corpus. The proficiency of the participants was in the range from the B1 to the C1 level according to Common European Framework of

Reference for Languages classification. The questions used in interviews elicited information about each speaker's English learning history, his/her profession and career, experience of participation in international projects as well as speech produced in a spontaneous conversation.

The TCEEE was applied in studies of tense and aspect usage and it was found (Salakhyan 2012) that Slavic speakers of English do not use the tense forms to a full extent which results in the phenomenon of simplifying and reducing the English system of tenses and aspects. These observations yield a conclusion of a possibility that Eastern European Englishes are emerging. However, in order that such varieties of English to be recognized, further studies concerning morphology, syntax, vocabulary, and semantics are required.

Acknowledgements. The author appreciates the support of Mexican Government which made it possible to complete this work: SNI-CONACYT, BEIFI-IPN, SIP-IPN grants 20162064 and 20161958, and the EDI Program.

References

1. Allan, Q. G.: Enhancing the language awareness of Hong Kong teachers through corpus data: The Telenex experience. Journal of Technology and Teacher Education, 7, pp. 57–74 (1999)
2. Allan, Q. G.: The TELEC secondary learner corpus: a resource for teacher development. Computer learner corpora, second language acquisition and foreign language teaching, pp. 195–212 (2002)
3. Berns, M.: Expanding on the Expanding Circle: Where do WE go from here? World Englishes, 24(1), pp. 85–93 (2005)
4. Biber, D., Conrad, S., Reppen, R., Byrd, P., Helt, M.: Speaking and Writing in the University: A Multidimensional Comparison. TESOL Quarterly, 36(1), pp. 9–48 (2002)
5. Biber, D., Conrad, S., Reppen, R., Byrd, P., Helt, M., Clark, V., Cortes, V., Csomay, E., Urzua, A.: Representing Language Use in the University: Analysis of the TOEFFL 2000 Spoken and Written Academic Language Corpus. Test of English as a Foreign Language (2004)
6. Boersma, P.: Praat, a system for doing phonetics by computer. Glot international, 5(9/10), pp. 341–345 (2002)
7. Carson-Berndsen, J., Gut, U., Kelly, R.: Discovering regularities in non-native speech. Language and Computers, 56(1), pp. 77–89 (2006)
8. Edwards, J. A., Lampert, M. D.: Talking data: Transcription and coding in discourse research. Hillsdale: Erlbaum (1993)
9. Gilquin, G., De Cook, S., Granger, S.: LINDSEI Louvain International Database of Spoken English Interchange. Handbook and CD-ROM. Louvain-laNeuve: Presses universitaires de Louvain (2010)
10. Gut, U.: The LeaP corpus. http://www.phonetik.unifrieburg.de/leap/ LeapCorpus.pdf (2004)
11. Gut, U.: The LeaP corpus A multilingual corpus of spoken. Multilingual corpora and multilingual corpus analysis, 14, pp. 3–23 (2012)
12. Hellwig, B., Van Uytvanck, D., Hulsbosch, M.: EUDICO Linguistic Annotator (ELAN) version 3.6 manual, http://www.lat-mpi.eu/tools/elan (2008)

13. Herment, S., Tortel, A., Bigi, B., Hirst, D., Loukina, A.: AixOx, a multi-layered learners' corpus: automatic annotation. In Proceedings of the 4th International Conference on Corpus Linguistics, Jaèn, Spain (2012)
14. House, J.: Developing pragmatic competence in English as a lingua franca. In: Knapp, K., & Meierkord, C, Lingua Franca Communication, Frankfurt: Peter Lang, pp. 245–269 (2002)
15. Housen, A.: A corpus-based study of the L2-acquisition of the English verb system. Computer learner corpora, second language acquisition and foreign language teaching, 6, pp. 2002–2077 (2002)
16. Hua, C., Qiufang, W., Aijun, L.: A Learner Corpus-ESCCL. In Proceedings of the Speech Prosody Conference, pp. 155–158 (2008)
17. Ishikawa, S.: A new horizon in learner corpus studies: The aim of the ICNALE project. In Weir, G., Ishikawa, S., & K. Poonpon, Corpora and language technologies in teaching, learning and research, Glasgow, UK: University of Strathclyde Publishing, pp. 3–11 (2011)
18. Ishikawa, S.: Basic Corpus Linguistics (in Japanese), the original title is Beshikku Kopasu Gengogaku. Tokyo: Hitsuji Shobo (2012)
19. Ishikawa, S.: The ICNALE and sophisticated contrastive interlanguage analysis of Asian learners of English. In: Ishikawa, S. (Ed.), Learner corpus studies in Asia and the world, Kobe, Japan: Kobe University, 1, pp. 91–118 (2013)
20. Ishikawa, S.: Design of the ICNALE-Spoken: A new data base for multi-modal contrastive interlanguage analysis. In Ishikawa, S. (Ed.), Learner corpus studies in Asia and the world, Kobe, Japan: Kobe University, 2, pp. 63–76 (2014)
21. Jenkins, J.: English as a Lingua Franca. Attitude and Identity, Oxford, Oxford University Press (2007)
22. Kwon, Y. E., Lee, E. J.: Lexical bundles in the Korean EFL teacher talk corpus: A comparison between non-native and native English teachers. The Journal of Asia TEFL, 11(3), pp. 73–103 (2014)
23. Luo, D., Yang, X., Wang, L.: Improvement of Segmental Mispronunciation Detection with Prior Knowledge Extracted from Large L2 Speech Corpus. In Interspeech, pp. 1593–1596 (2011)
24. Mattheoudakis, M.: Learner Corpora of English: Glimpses into learners' L2 development. In Proceedings of 4th Postgraduate Student Conference: Assessing and Analyzing Discourses, Faculty of English Language and Literature National and Kapodistrian University of Athens (2014)
25. MacWhinney, B.: The CHILDES project: The database. Psychology Press, Vol. 2 (2000)
26. MacWhinney, B.: The CHILDES project. Tools for analyzing talk – Electronic version, Part 1: The CHAT transcription format, http://childes.psy.cmu.edu/ manuals/chat.pdf (2008)
27. Menzel, W., Atwell, E., Bonaventura, P., Herron, D., Howarth, P., Morton, R., Souter, C.: The ISLE corpus of non-native spoken English. In: Proceedings of LREC 2000: Language Resources and Evaluation Conference, European Language Resources Association, 2, pp. 957–964 (2000)
28. Meunier, F.: Learner corpora and pedagogical applications. The Routledge Handbook of Language Learning and Technology, pp. 376 (2016)
29. Milde, J. T., Gut, U.: A prosodic corpus of non-native speech. In Bel, B., & Marlien, I. (Eds.), Proceedings of the Speech Prosody 2002 Conference, Aix-en-Provence: Laboratoire Parole et Langage, pp. 503–506 (2002)
30. Myles, F., Mitchell, R.: French learner language oral corpora (FLLOC) (2007)
31. Palacios, I.: The Santiago University Corpus of Learner English. Santiago, University of Santiago de Compostela, http://www.sulec.es (2005)

32. Muñoz, C.: Age and the Rate of Foreign Language Learning. Clevedon: Multilingual Matters (2006)
33. Reder, S., Harris, K., Setzler, K.: The multimedia adult ESL learner corpus. TESOL Quarterly, 37(3), pp. 546–557 (2003)
34. Reinhardt, J. S.: Directives usage by ITAs: An applied learner corpus analysis. Doctoral dissertation, Pennsylvania State University (2007)
35. Reinhardt, J.: Directives in office hour consultations: A corpus-informed investigation of learner and expert usage. English for Specific Purposes, 29(2), pp. 94–107 (2010)
36. Salakhyan, E.: The Tübingen Corpus of Eastern European English (TCEEE): From a small-scale corpus study to a newly emerging non-native English variety. A Journal of English Linguistics, Jan Kochanowski University Press, 1, pp. 143–157 (2012)
37. Thorne, S., Reinhardt, J., Golombek, P.: Mediation as objectification in the development of professional discourse: A corpus informed curricular innovation. In Lantolf, J., & M. Poehner (Eds.) sociocultural theory and the teaching of second languages, London: Equinox, pp. 256–284 (2008)
38. Tortel, A.: ANGLISH. Une base de données comparatives de l'anglais lu, répété et parlé en L1 & L2. TIPA, Travaux interdisciplinaires sur la parole et le langage, 27, pp. 111–122 (2008)
39. Wen, Q. F., Wang, L. F., Liang, M. C.: Spoken and written English corpus of Chinese learners. Foreign Language Teaching and Research Press (2005)
40. Yang, H., Wei, N.: Construction and data analysis of a Chinese learner spoken English corpus. Shanhai Foreign Languse Eduacation Press (2005)