First Stage in the Construction of METU Turkish English Exam Corpus (METU TEEC)

Çiler Hatipoğlu

Abstract

This paper presents and discusses the first stages in the construction of the Middle East Technical University Turkish English Exam Corpus (METU TEEC), which has been compiled by a research team at Middle East Technical University (METU), Ankara, Turkey. The corpus consists of 1914 Linguistics and ELT exam papers (955483 words in total) written in timed circumstances with no access to reference material by the students at the Foreign Language Education (FLE) department at METU, Ankara between January 2005 and December 2012. The corpus is intended to cater for the needs of both theoreticians/researchers and practitioners/pedagogues; therefore, each of the scripts included in the corpus is tagged with rich meta information about the informants (e.g., age, gender, years of learning English), the exam (e.g., course, type of question, academic year) and exam writers (e.g., educational background, experience). This system allows the compilation of sub-corpora according to the needs and interests of the researchers and practitioners. The first sections of the paper present the context and the needs for such a corpus while the latter sections discuss the development of the coding and tagging systems employed for this corpus.

Keywords: Specialised corpus development, learner corpus, exam English, text type, Turkey and Turkish.

Introduction

English has become "an integral component of all levels of national education in Turkey" (Doğançay-Aktuna, 2005, p. 254) and now it is "the most commonly taught foreign language in Turkish schools" (Bayyurt, 2010, p. 163). People with different ages, backgrounds and aims invest time and effort trying to learn the language. English is "the only foreign language that has become a compulsory subject at all levels of education, featuring predominantly in language policy" (Kırkgöz, 2009, p. 667). What is more, there are both state and private secondary schools and universities where the medium of instruction is English, and their number increases every year. In those institutions, due to the nature of the educational environment, during lectures, classroom discussions and in all exams, students have to use English and their educational success and/or failure depends on whether or not they learn to "adjust to a wide range of tasks in the university accomplished through language" (Biber, 2006, p. 1).

Despite the popularity of English in the country and "despite the time, money and effort spent on foreign language education in Turkey, a low foreign language proficiency level has remained a serious problem" (Işık, 2008, p. 15). Both practitioners and researchers agree that "English language teaching/learning is

Çiler Hatipoğlu, Prof. Dr., Middle East Technical University, Faculty of Education, Department of English Language Education, Ankara, ciler@metu.edu.tr

problematic in Turkey" (Kızıldağ 2009, p. 189). Therefore, a large number of studies trying to uncover the reasons behind these problems have been conducted since the late 1990s (Aktaş, 2005; Işık 2008; Oğuz, 1999; Paker, 2007; Şallı-Çopur, 2008; Tılfarlıoğlu & Öztürk, 2007) and factors such as language planning, selected foreign language teaching methodologies, student interest and motivation, learning environment and learning materials have been listed. One factor that comes to the forefront in those discussions, though, is the language teacher and his/her knowledge of the target language. Language teachers are central to the success of the educational process since the collection of their underlying Beliefs, Assumptions and Knowledge (i.e., the BAK, Sowden, 2007; Woods, 1996, p. 196) heavily affects what happens in class. What happens in language classes in Turkey is very important since here English is taught as a foreign language (i.e., it is taught and learned in a place where it is not typically used as the medium of ordinary communication). As in other foreign language learning contexts, native speakers of Turkish learning English are surrounded by their own native language and "have to go out of their way to find stimulation and input in the target language. These students typically receive input in the new language only in the classroom..." (Oxford & Shearin, 1994, p. 14). This, in turn, means that the teachers are role models for their students and their knowledge and skills in the foreign language frequently determine whether their students would become motivated and successful language learners and skilful communicators in the target language or not.

Taking into consideration the documented problems in teaching and learning English in the country, the growing number of English medium institutions in Turkey and the importance of the language competence of language teachers for successful language teaching in the country, it was decided to embark on a project aiming to create a specialised learner corpus including samples coming from non-native pre-service English language teachers in Turkey.

It was decided to focus specifically on the English produced by non-native pre-service teachers of English since in EFL contexts they are usually the ones who "impart language competence to learners" (Thomas, 1987, p. 34). All researchers concur that without competent language teachers with solid knowledge of the target language and culture, and the ability to teach this language accurately and confidently (Barnes, 2002; Demirel, 1989, 1990) success in language teaching is a difficult objective. High competence in the target language is a "prerequisite for the other competencies since incomplete knowledge of language interferes with effective language teaching" (Şallı-Çopur 2008, p. 11). Therefore, identifying the features of the English utilised by pre-service English language teachers will give them, their trainers, curriculum developers and maybe even teaching material writers a chance to see what their strengths and weaknesses are, and, if needed, some changes in the teaching and learning procedures utilised in the English language teacher training programs in the country can be made. This, in turn, may lead to more confident and better qualified teachers and teacher-training programs in the country. In addition, the creation of such specialised corpora can "contribute to rehabilitating learner output by providing researchers with substantial sources of tightly controlled computerised data which can be analysed at a range of levels using increasingly powerful linguistic software tools" (Granger, 2002, pp. 6-7). On the other hand, since "diversity is a feature of global use amongst both native and non-native speakers" (Weil & Pullin, 2011, p. 28) of English nowadays, it would be interesting to uncover the features of “exam Turkish English”. This corpus will allow researchers to identify and discuss the strategies that native speakers of Turkish utilise to explore and negotiate meaning when the aim is to produce mutual understanding in a language which is not their native tongue. It will also show whether or not future English language teachers are able to deal with the complex contexts within which they find themselves in an appropriate and enriching manner.

So, this paper presents and discusses the first stages of the development of the specialised Turkish English Exam Corpus (TEEC), which has been compiled by a research team at Middle East Technical University (METU), Ankara, Turkey. The corpus consists of 1914 exam papers written by students at the Foreign Language Education (FLE) department at METU, Ankara. Only exam papers were included in the corpus since the aim was to collect spontaneous data which are the real representation of the English of the pre-service English language teachers, according to Ellis (2001) and Selinker (1972) (see Section 2.2 for further details). Even though data only from the FLE students at METU were collected for this first version of the METU TEEC, it is believed that the results of the project can shed light on the needs of the students in all EFL departments in Turkey since the departments of FLE in Turkey are required to follow a standardized curriculum prescribed by The Council of Higher Education (YÖK) (also see Hatipoğlu, 2010).

The specific purposes for creating a Middle East Technical University Turkish English Exam Corpus (METU TEEC) were the following:

(i) to compile objective data that will allow theoreticians and researchers to explore and describe the “exam English” (i.e., a specialised variety English) of pre-service English language teachers who are native speakers of Turkish. According to Altenberg and Granger (2002), Granger (1998) and Mouranen (2003), the compilation of such corpora is imperative for the creation of valid as well as reliable SLA theories and language development projects since these allow experts to uncover the orthographic, grammatical, lexical, lexico-grammatical, discoursal, pragmatic and rhetorical features of the English utilised by specific communities of practice;

(ii) to facilitate and encourage research into the potential applications of learner corpora to pedagogical materials and learning aids since, when compared with other learners, teaching materials (e.g., books, exam companions) for advanced learners are underdeveloped or scarce to find;

(iii) to generate linguistic profiles for advanced language learners who had chosen ‘teaching English’ as their profession and, if needed, to enable curriculum designers to set informed targets for facilitating the language development of such specialised groups of learners;

(iv) to allow for diachronic analysis of the English used by this specific group of learners (i.e., this community of practice);

Corpus Design/Design Criteria

Tono (2003, p. 800) argues that “since people are interested in different aspects of learner language, it is quite natural that the design of learner corpora will vary from project to project”. He introduces, however, three basic categories of criteria that should be considered when building specialised learner corpora: (1) Learner-Related, (2) Language-Related and (3) Task-Related (see Table 1).

In the next paragraphs, the features of METU TEEC will be presented following the framework introduced by Tono (2003).

Table 1. Design considerations for building learner corpora (Tono 2003, p. 800)

	Types of Features		
	Learner-Related	**Language-Related**	**Task-Related**
1	**internal-cognitive** [age/cognitive style]	**mode** [written/spoken]	**data collection** [cross-sectional/ longitudinal]
2	**internal-affective** [motivation/attitude]	**genre** [letter/diary/fiction/essay]	**elicitation** [spontaneous/prepared]
3	**L1 background**	**style** [narration/argumentation]	**use of references** [dictionary/source text]
4	**L2 environment** [ESL/EFL]/ [level of school]	**topic** [general/leisure/ etc]	**time limitation** [fixed/free/homework]
5	**L2 proficiency** [standard test score]		

Learner-Related Variables

The learners who have contributed data to the METU TEEC form a rather homogeneous group. All of them are either native speakers of Turkish or speak it as one of their mother tongues. They have been learning English as a foreign language for 10-12 years and almost all of them are graduates of Anatolian Teacher Training High Schools (ATTHS) in Turkey. They are all undergraduate students at the Department of Foreign Language Education (FLE) at METU, Ankara (i.e., university undergraduates specializing in English Language Teaching) and since they chose to study in this department, it can be argued that all of them are motivated language learners with a positive attitude towards the language. Their level of proficiency can be described as advanced due mainly to the following three reasons:

(1) METU, Ankara, where the data were collected, is one of Turkey's most competitive universities. It is a state university where the language of instruction is English. To enter the FLE Department of METU, candidates have to take the English Test (LYS) (a test including 80 MCI, usually based on the topics included in the curriculum of Anatolian and Private High Schools), which is a part of the University Placement Exam (YGS) battery prepared and administered by the Student Selection and Placement Centre (ÖSYM) set in Ankara. Since there is a central matriculation system in the country, all university candidates take the same exam and only the top 1% of the applicants taking the LYS test are accepted to the FLE Department at METU.

(2) All METU entrants have to take the METU English Proficiency Test (EPE) before being allowed to start their undergraduate studies. If students are successful in the EPE (i.e., score at least 60 out of 100 which is equivalent to 6.0 on the IELTS exam, 75 on TOELF IBT and B1 level in the Common European Framework of Reference for languages) they are permitted to continue their studies at the University. If they fail they are required to spend at least one semester - but normally one year - in the Department of Basic English of the University, where they receive full-time English language training.

(3) Once in the department, students take courses such as *Contextual Grammar, Advanced Reading and Writing, Listening and Pronunciation* and *Oral Communication* in the first and second semesters of their training and the aim of these courses is to refine the language skills of future English language teachers.

Language-Related Variables

Papers from 66 written exams (i.e., 1914 exam papers totalling 955,483 words) taken between January 2005 and December 2012 by the students at FLE department at METU, Ankara were included in the corpus (see Table 2).

Table 2. Distribution of exams and exam papers according to academic years

Years	**Exams**		**Exam Papers**	
	N of Exams	**%**	**N of Exam Papers**	**%**
2004-2005	3	4.5	73	3.8
2005-2006	7	10.6	142	7.4
2006-2007	13	19.7	380	19.9
2007-2008	6	9.1	135	7.1
2008-2009	4	6.1	135	7.1
2009-2010	11	16.7	297	15.5
2010-2011	19	28.8	682	35.5
2011-2012	3	4.5	70	3.7
TOTAL	66	100.0	1914	100.0

This corpus was intended to be a specialized Turkish English Exam Corpus and for now very specific data related only to two main subjects - Linguistics and ELT taught in the FLE departments- were gathered. Nonetheless, due to the nature of the courses and the topics covered in those courses, the exam questions covered a wide variety of topics (e.g., ***Linguistics:*** phonetics, phonology, morphology, discourse, pragmatics, semantics, historical linguistics, first and second language acquisition; ***ELT:*** foreign language testing and evaluation, translation, theories of reading, types of reading, author positioning). There was a rich variety of question types included in the scrutinised exams (e.g., matching, multiple choice items, short answers, True/False).

However, since the aim of the project was to examine the language produced by trainee English language teachers, only data coming from the following types of questions were gathered: *compare and contrast, data analysis, definitions, descriptions, essays, short answers* and *translations*.

The settings where the data were collected were the exam halls/classrooms and the communicative purpose of the performed activity (i.e., answering exam questions) was to show/convince the examiners that the testees could, for instance, define, describe and/or name a phenomenon, compare and contrast events/theories, analyse a text/set of data and explain, classify and arrange events/data.

Task-Related Variables

Since the only contexts where the use of English is officially required in Turkey are the English medium institutions and since previous research has shown that in the majority of these educational contexts students' success or failure is determined by what they do on exams such as quizzes, midterms and finals (Hatipoğlu, 2011), it was decided to focus on the features of "exam English" utilised by native speakers of Turkish training to become English language teaches. The exams from which the data were collected are briefly defined and described below:

Quizzes: Only 3 out of 66 (4.5%) exams included in the METU TEEC are quizzes. They were short announced tests lasting from 15 to 30 minutes and their aim was to check students' understanding of a specific topic, data analysis technique, or a reading/listening text.

Midterm Exams were tests scheduled by the course instructors. Their number during the academic term, as well as their length and content were determined by the objectives of the individual course. While Linguistics courses in the METU TEEC generally included both theoretical and practical questions some ELT midterm exams included only text and data analysis questions. Thirty-seven (56.1%) of the exams in METU TEEC are midterm exams.

Final Exams at METU are usually three-hour long cumulative tests given to the students at the end of each of the academic terms (i.e., Fall, Spring and Summer). Their dates and times are scheduled by the Registrar's Office of the university. The purpose of these tests is to assess each student's knowledge of the topics covered during the term. The bulk of the final exams administered at FLE METU included both theoretical and practical questions and they were, usually, the most important exams for the students since, in the majority of the courses, they carried the largest weight in the students' course grade. METU TEEC comprises 26 (39.4%) final exams.

All the data included in the corpus were collected in timed circumstances with no access to reference material by the students since the aim was to collect spontaneous data. Ellis (2001:670) argues that it is possible to distinguish three broad types of data in SLA: (a) language use data, (b) metalingual judgements and (c) self-report data. The first of those types of data is defined as the learners' attempts to use the L2 in either comprehension or production and it has two sub-categories "natural use" and "elicited language use". Language use data is classified as "natural" if no control is exerted on the learners' performance and "elicited" if they result from a controlled experiment. The second data type is usually elicited by grammatical judgement tests and usually assesses learners' intuitions about the L2 (e.g., students are asked to judge the grammaticality of various groups of sentences); finally, the self-report data are obtained through interviews, questionnaires or think aloud activities. Selinker (1972, p. 210) criticizes research methodologies which rely on students' performance during meaningless activities in the second-language classroom and argues that spontaneous data are needed to study learners' "real" interlanguage.

Following Selinker (1972) and keeping in mind the main objective of the project (i.e., to uncover the "real" characteristics of the English utilised by future English language teachers) only data coming from contexts requiring the natural use of English were included in the METU TEEC.

Corpus Development

Compilation of Exam Scripts

One of the most important steps in the project was the collection of the exam scripts. Meetings were arranged with colleagues teaching the courses whose exams were to be included in the corpus, and they were informed about the aims and expected outcomes of the project. Students were also informed about the project and were asked to fill in a consent form if they wanted their work to be included in the corpus. After the original scripts were collected, they were photocopied for data entry and scanned as PDF files for later reference. The original exams were returned to the course instructors.

At the moment METU TEEC comprises only linguistics and ELT courses in which English is not an object of study in itself. Nevertheless, in order to achieve as much representativeness as possible data coming from a wide variety of linguistics and methodology courses in the curriculum were collected and included in the corpus. It was also ensured that data produced by students in all levels (i.e., pre-service English language teachers in their first, second, third or fourth year of education at the university) were incorporated in the corpus (see Table 3).

As many researchers have pointed out, there is no ideal size for a specialised learner corpus; nevertheless, when we compare METU TEEC with other available specialised learner corpora (e.g., International Corpus of Learner English (ICLE) with 3 million words; Japanese English as a Foreign Language Learner (JEFLL) with 500,000 words; Janus Pannonius University Corpus (JPU) with 400,000 words) and use Aston's (1997) criteria (i.e., small corpora are the ones in the 20,000-250,000 word range while large corpora examples are BNC or ANC), then, METU TEEC can be classified as a "medium size"/adequate size corpus with its almost one million words. Since the final aim is to collect a corpus which includes exam data from all of the subjects taught at the FLE Departments in Turkey, the sampling technique used for the first version of the corpus was convenient non-probability sampling. The use of this sampling technique allows researchers to gather data that are more representative of a specific group (i.e., trainee English language teachers) because when a specific group is targeted the gathered answers are similar to what the rest of the population of this group will answer (Gorard, 2003).

Table 3. Distribution of exams and exam papers according to academic years

Years	Exams		Exam Papers	
	N of Exams	%	N of Exam Papers	%
YEAR 1	30	45.5	789	41.2
YEAR 2	14	21.2	443	23.1
YEAR 3	7	10.6	173	9.1
YEAR 4	15	22.7	509	26.2
TOTAL	66	100.0	1914	100.0

Pre-Processing Stages

As a first step of the digitalisation of METU TEEC scripts, it was decided to save each exam paper as a separate plain text file (extension .TXT), with a part of its meta-information coded in its file name. It was decided to save texts in TXT format, since this is the standard format usually used for texts included in corpora, and it is more practical to use a simple text editor to work with texts rather than a standard word-processing program. The meta information included in the title of the files at this stage included the "Unique file number (UFN)_Unique informant number (UIN)_Year when the exam was administered (EY)_The course code (CC)_The type of exam (ET)_The term when the exam was administered (EAT)" (e.g., 0000000001_000001_2006-2007_FLE146_Midterm_Summer). A group of research assistants and students were trained and then asked to digitise the scripts. While entering the data they were asked to do the following:

(i) Enter students' original work with absolute fidelity (i.e., spelling, punctuation, grammar and any other mistakes/problem should not be corrected, as they have to appear in the digital file)

(ii) Ignore teachers' corrections and comments if any

(iii) Save each exam paper using the "UFN_UIN_EY_CC_ET_EAT" formula.

Alansary et al. (2007) argue that corpora are not simply collections of texts but rather representations of language, and the design of the corpora depends upon what they are meant to represent. In turn, their designs determine the kinds of research questions that can be answered using the created corpora. The coding of the files in the current corpus (METU TEEC) started with the unique file number followed by the unique number given to each of the informants and the EY, CC, ET and EAT criteria since, even though one of the main aims of the corpus was to enable researchers to observe the individual development of each informant, another major goal for compiling this corpus was to enable comparison and contrast of data belonging to different subjects/groups across categories such as courses, types of questions, semesters, years of education at university. What is more, when compiling a corpus, it is usually expected that the names of the files and the way they are stored reflect the hierarchical structure of the corpus. After a number of trials it was determined that this header structure successfully reflected the hierarchy in the corpus.

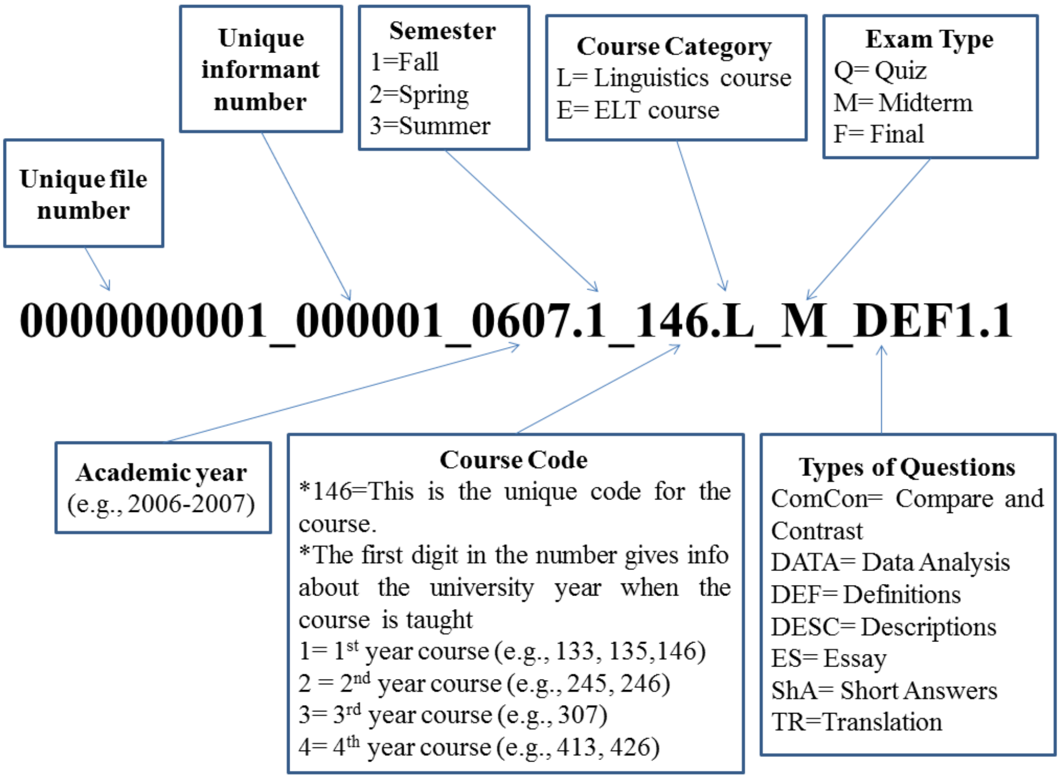

Unique informant number
Semester
1=Fall
2=Spring
3=Summer
Course Category
L= Linguistics course
E= ELT course
Exam Type
Q= Quiz
M= Midterm
F= Final
Unique file number
0000000001_000001_0607.1_146.L_M_DEF1.1
Academic year
(e.g., 2006-2007)
Course Code
*146=This is the unique code for the course.
*The first digit in the number gives info about the university year when the course is taught
1= 1st year course (e.g., 133, 135,146)
2 = 2nd year course (e.g., 245, 246)
3= 3rd year course (e.g., 307)
4= 4th year course (e.g., 413, 426)
Types of Questions
ComCon= Compare and Contrast
DATA= Data Analysis
DEF= Definitions
DESC= Descriptions
ES= Essay
ShA= Short Answers
TR=Translation

Figure 1. File header coding system in METU TEEC

Alansary et al. (2007) warn corpora compilers, however, that the creation of a corpus is a “cyclical” process and because of that, there should be constant evaluations and re-evaluations of the corpus and the tools related to it as the corpus is being compiled. Therefore, it was decided to test the efficiency and limits of applicability of the header system created for the corpus. After piloting the devised header system and keeping in mind that texts coming from different registers (i.e., tasks) “tend to exhibit a wide range of linguistic variation” (Biber & Conrad, 2009, p. 3), one more piece of information was added to the file names in METU TEEC – the type of question that the students had to answer on the exam (e.g., compare and contrast, data analysis, definition, short answers). In order to make the search easier, it was decided to separate each exam file into sub-sections made of different exam questions and the new files were stored with a name now consisting of seven units of information: UFN_UIN_EY_CC_ET_EAT_TYPE OF QUESTION (QT). If the questions had a number of sub-questions or there were 2 questions requiring the same type of answers/analysis, then numerical information such as 1.1, 1.2, 2.1 were added after the code for the type of question (see Figure 1). So, the file name of a Midterm exam of a Linguistics course taken by Subject 000001 in the 2006-2007 academic year would be as shown in Figure 1.

Tagging

Metadata is defined as 'data about data' (Turner, 2002) and the compiled learner corpora around the world differ from each other with respect to how much and what type of metadata is added to the raw text. The International Corpus of Learner English (ICLE) (Granger, 2003) and the Cambridge Learner Corpus (CLC) include POS and error tags, for instance, while the Indianapolis Business Learner Corpus (IBLC) (Connor & Precht, 1998) and the Janus Pannonius University Corpus (JPUC) (Jozsef, 1998) comprise plain texts and do not have any further linguistic annotation. According to Tono (2003, p. 4), whether a mark-up is used or not depends on the aims of the corpus developers; however, if a tagging scheme is to be utilised it should include at least two aspects:

(a) linguistic category classification (e.g. [grammar] - [verb] - [morpheme] - [tense])

(b) target modification taxonomy (e.g. [omission/ addition/ misinformation]) (James, 1998).

Since the aim in creating this corpus was to identify the characteristics of the exam English used by pre-service English language teachers (i.e., advanced learners of English) it was decided that the corpus would be more useful to potential users if it were tagged for features such as orthography, punctuation, grammar (i.e., word formation, agreement, tense, mood, word order) as well as discoursal, pragmatic and rhetorical characteristics.

The next step in the process was to choose the type of error-tagging system (i.e., flat or multi-layered) to be employed for METU TEEC and the program/system to be used to integrate the corpus.

The majority of the currently available error-tagged learner corpora employ a kind of flat file format. Within this system every word is followed by "a part-of-speech tag (following the STTS tagset, cf. Schiller et al., 1995) and its lemma" (Lüdeling et al., 2005, p. 4) as shown in Example (1):

Example 1 (source: Verbmobil corpus; http://verbmobil.dfki.de/)

I/PP/I have/VBP/have got/VBN/get ,/,/, Monday/NP/Monday or/CC/or Tuesday/NP/Tuesday off/RB/off

This system has, however, two important drawbacks according to Lüdeling et al. (2005, p. 4-5). First, the number and category of annotation layers must be determined in the corpus design phase since it is not easily possible to add layers in the implementation stages. Moreover, each token (i.e., an occurrence of any given word form) must be given a value for every annotation layer in the whole corpus. This system is not suitable for error annotation because it cannot be determined how many errors have to be tagged with any given token prior to the tagging process. Second, since the tag and its lemma immediately follow the word to which they belong/refer, it is not possible to join adjacent cells and annotate errors related to sequences of words as in the example shown in Figure 5 below.

Because of the abovementioned drawbacks related to the flat file annotation system, it was decided to use the multi-layer standoff (MLS) annotation (Carletta et al., 2003) in this corpus. In the MLS annotation there are a number of independent layers (e.g., Layer 1: the original text, Layer 2: annotation related to orthography, Layer 3: annotation related to grammar), which are connected to each other. Furthermore, the term "standoff" refers to the fact that, differently from the flat annotation models, here the annotation is not inserted into the text. The original text is available for reference and/or search whenever needed.

The number of layers in the MLS system is determined by the needs of the corpus compiles (See Figure 3 below). The original text is in a reference line and each related annotation is coded in an independent level, with pointers to the reference line.

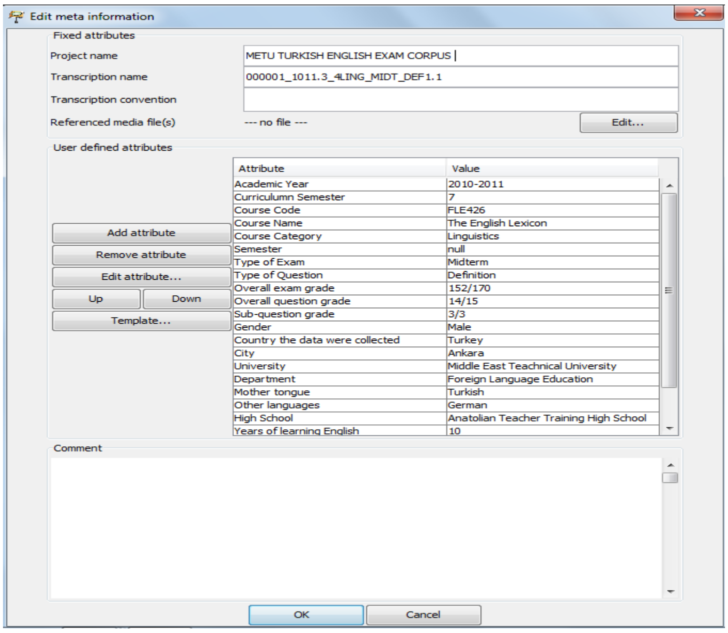
Since the system developed by Schmidt (2004) in his EXMARaLDA partiture editor (i.e., Extensive Markup Language for Discourse Annotation; http://exmaralda.org/) allows the use of the MLS system in a straightforward way, it was decided to use this program for the METU TEEC project. The time line and multiple annotation layers are always visible in EXMARaLDA and the program allows the addition, deletion, merge and split of different cells, so that errors are represented appropriately in the corpus.

Examples from METU TEEC

The aim of the METU TEEC project was to create a corpus catering for the needs of both theoreticians/researchers and practitioners/pedagogues, therefore, each of the texts entered in the EXMARaLDA partiture editor was accompanied by rich meta information that can be described in three categories (see Figure 2):

(1) Information related to the informant/test taker: age, gender, years of learning English, knowledge of any other (foreign) languages, educational background of the test taker (i.e., which high school he graduated from), his/her university and department;

(2) Details related to the exam (some of which are also represented in the header of the files): in which academic year the exam was administered (e.g., 2005-2006, 2010-2011), semester in the curriculum when the course is taught (e.g., 1st, 7th), the academic semester when the data were gathered (e.g., Fall, Spring, Summer), name and code of the course (FLE 426: The English Lexicon), course category (e.g., Linguistics, Methodology, Translation, Testing), type of exam (e.g., Quiz, Midterm, Final), type of question (e.g., comparison and contrast, data analysis, definitions) and the question itself (so that the researchers are able to see how different questions are interpreted and answers by the test-takers), the overall grade of the students for the exam (e.g., 152/170), the grades for each of the questions (e.g., 14/15) and sub-questions (2/3) in the exam;



Fixed attributes

Project name	METU TURKISH ENGLISH EXAM CORPUS
Transcription name	000001_1011.3_4LING_MIDT_DEF1.1
Transcription convention	
Referenced media file(s)	--- no file ---

User defined attributes

Attribute	Value
Academic Year	2010-2011
Curriculumn Semester	7
Course Code	FLE426
Course Name	The English Lexicon
Course Category	Linguistics
Semester	null
Type of Exam	Midterm
Type of Question	Definition
Overall exam grade	152/170
Overall question grade	14/15
Sub-question grade	3/3
Gender	Male
Country the data were collected	Turkey
City	Ankara
University	Middle East Teachnical University
Department	Foreign Language Education
Mother tongue	Turkish
Other languages	German
High School	Anatolian Teacher Training High School
Years of learning English	10

Comment

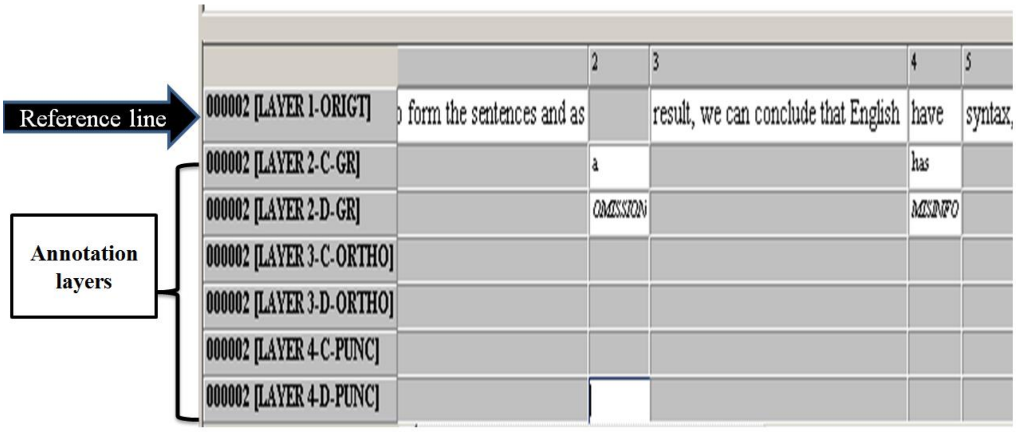
Figure 2. Meta information available for the files in METU TEEC

(3) Information related to the test writer: age, gender, years of experience in teaching (e.g., 5, 11), educational background (e.g., BA: English Language Teaching, MA: English Language Teaching, PhD: Linguistics)

It is believed that this rich meta information which allows the compilation of sub-corpora according to the needs of the researchers will be able to answer to the needs of a wide group of researchers and practitioners with a variety of interests.

The annotation of all of the texts in METU TEEC is not complete yet and the design and refinement of the tagsets to be employed in the corpus continue. However, the following section will describe the basic annotation system designed for METU TEEC and will provide a number of examples from the corpus data.

For this first version of the corpus, we focused on three error types - Grammar (as a general level), Orthography and Punctuation - and three different steps were followed for the annotation of each of the errors: (1) identification and isolation; (2) supplying the target form; (3) classification of the problem (see Figure 3).

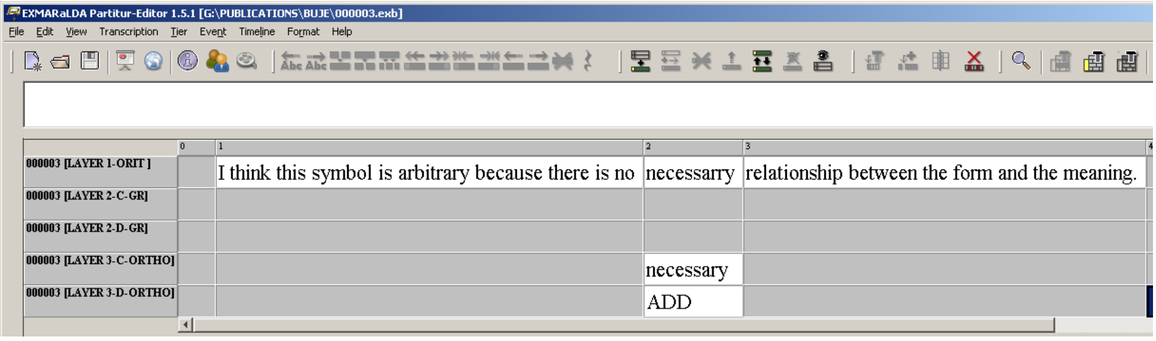


		2	3	4	5
000002 [LAYER 1-ORIGT]	ɔ form the sentences and as		result, we can conclude that English	have	syntax,
000002 [LAYER 2-C-GR]		a		has	
000002 [LAYER 2-D-GR]		OMISSION		MISINFO	
000002 [LAYER 3-C-ORTHO]					
000002 [LAYER 3-D-ORTHO]					
000002 [LAYER 4-C-PUNC]					
000002 [LAYER 4-D-PUNC]					

Figure 3. Annotation System in METU TEEC

The annotation procedures related to the first two stages mentioned above (i.e., (1) identification and isolation of the problem, (2) supplying the target form) are a combination of computerized and manual work. To classify the language use of informants related to grammar, orthography and punctuation as (in)correct a wide variety of reference resources such as grammar books (e.g., *Advanced Grammar in Use* by Hewings*;* A *Universal Grammar of English* by Quirk and Greenbaum*; Communicative Grammar of English* by Leech and Svartvik; *Longman Grammar of Spoken and Written English* by Biber et al.; *The Grammar Book* by Celce-Murcia and Larsen-Freeeman) and dictionaries (e.g., *Macmillan English Dictionary for Advanced Learners* (MEDAL), *Longman Dictionary of Contemporary English* (LDCE), *Oxford Advanced Learner's Dictionary* (OALD) and *The BBI Dictionary of English Word Combinations)*were used. For the classification of the identified problems, on the other hand, the scheme devised by Dulay et al. (1982) including categories such as "*omission", "addition", "misinformation"* and *"misordering"* was utilised.

To be able to follow this annotation system three layers, each with two sub-annotation levels, were needed: Layer 2 represented Grammar, Layer 3: Orthography and Layer 4: Punctuation (see Figure 3). In each of the layers we added tiers C and D, where C stands for the correct/target form and D stands for Description of the mistake as described by Dulay et al. (1982).



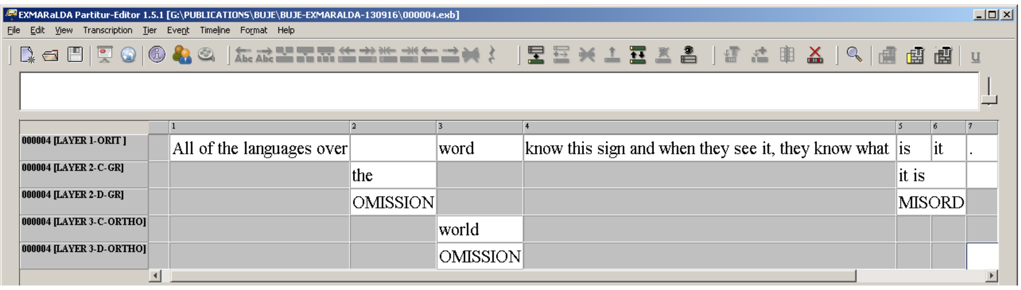
EXMARaLDA Partitur-Editor 1.5.1 [G:\PUBLICATIONS\BUJE\000003.exb]

File Edit View Transcription Tier Event Timeline Format Help

	0	1	2	3
000003 [LAYER 1-ORIT]		I think this symbol is arbitrary because there is no	necessarry	relationship between the form and the meaning.
000003 [LAYER 2-C-GR]				
000003 [LAYER 2-D-GR]				
000003 [LAYER 3-C-ORTHO]			necessary	
000003 [LAYER 3-D-ORTHO]			ADD	

Figure 4. Coding of an *"Orthographic Addition mistake"*

So, when we look at Figures 3 and 5, we see examples of the coding of *omission* (i.e., the indefinite and definite articles "a" and "the" were not used when required) and misinformation (e.g., *have* was used with a singular third person subject in the expression "*we can conclude that English has syntax*") grammar mistakes in the corpus; while Figures 4 and 5, display examples of the annotation of "orthographic addition/omission" and "grammatical misordering" mistakes, respectively.



EXMARaLDA Partitur-Editor 1.5.1 [G:\PUBLICATIONS\BUJE\BUJE-EXMARALDA-130916\000004.exb]

File Edit View Transcription Tier Event Timeline Format Help

	1	2	3	4	5	6	7
000004 [LAYER 1-ORIT]	All of the languages over		word	know this sign and when they see it, they know what	is	it	.
000004 [LAYER 2-C-GR]		the			it is		
000004 [LAYER 2-D-GR]		OMISSION			MISORD		
000004 [LAYER 3-C-ORTHO]			world				
000004 [LAYER 3-D-ORTHO]			OMISSION				

Figure 5. Coding of a *"Grammatical Misordering Mistake."*

The annotation system described above has already showed that it can accommodate the data included in the METU TEEC and the first preliminary results of the analysis show that the most common mistakes related to the analysed areas in the corpus are the following:

(1) Grammar: the article system in English (i.e., the use of the definite and indefinite article in English), wh-clauses;
(2) Orthography: consonant/vowel omission and consonant doubling;

(3) Punctuation: the use of semi-column in English.

Conclusions

On November 16, 1929, in an article entitled "English for the Turks", the correspondent for the London *Nation and Athenaem* wrote:

> …almost everybody,… not only in Constantinople, but throughout Anatolia, is learning English as hard as he can go. … The Ministry of Public Instruction has introduced English as a regular part of the school routine in all the secondary schools throughout the country. … On all sides, and every day, one hears such expressions as 'I want to learn English' and 'How long will it take me to learn English?' (Bear, 1992, p. 24).

Today, more than eight decades later, this demand for English continues to grow not only in Turkey but also around the world. "As it does, so does the need for competent teachers of English as a foreign language" (Bear, 1992, p. 24). Studies done in Turkey (e.g., Demirel, 1989, 1990; Gürbüz, 2006) and other countries (e.g., Lipton, 1996) repeatedly show that one of the most important competencies/attributes of successful language teachers is their "superior level of proficiency in all foreign language skills" (Lipton, 1996, p. 39). Therefore, one of the most important objectives of teacher education programs should be to facilitate trainees to gain "competence to impart competence in language" (Thomas, 1987, p. 34).

Language teachers in EFL contexts are role models for their students and their skills and knowledge in the foreign language determine whether students continue to work on their language development or not (Hanson, 2011). Therefore, it is important to uncover and describe in detail the characteristics of the English utilised by trainee language teachers so that, if needed, changes in the teachers training programs, teaching materials and, methods and techniques are made. It is believed that METU TEEC is a source which will allow theoreticians and practitioners to do that (i.e., examine in detail what pre-service English teachers do with English). We trust that this project is a valuable contribution to the ELT, teacher training, corpus linguistics fields since as far as the author is aware, up to now no such specialised corpus has been created in Turkey.

It is also hoped that this project will have wide-ranging effects on the teaching and learning of English in the country. If the teachers are models, then it is expected that the data coming from this project will not only affect the students at the English language teaching departments in Turkey but would also have a spill over effect on language education at the primary, secondary and high school level in Turkey. Language teacher trainers, curriculum designers, material writers and the language teachers themselves will have detailed information related to the characteristics of the English employed by this special group of language learners and, if needed, any "solid changes" could be made.

The corpus focuses on advanced interlanguage reflecting the wish of the corpus compilers to compensate for the dearth of pedagogical materials for advanced learners when compared to lower proficiency levels. Each file in the corpus is accompanied by rich meta information which will allow for the development of international comparative studies leading to the identification of some of the features of "Turkish English Exam Language".

References

Björkman, B. (2013). *English as an academic lingua franca*. Berlin: De Gruyter Mouton

Cogo, A., & Dewey, M. (2012). *Analysing English as a lingua franca*. London: Continuum.

Aktaş, T. (2005). Yabancı dil öğretiminde iletişimsel yeti. *Journal of Language and Linguistic Studies*, *1*(1), 89-100.

Alansary S., Nagi M., & Adly N. (2007). Building an International Corpus of Arabic Progress of Compilation Stage. In *Proceedings of 7th International Conference on Language Engineering* (pp. 337-366). Cairo, Egypt.

Altenberg, B., & Granger, S. (2002). The grammatical and lexical patterning of make in native and non-native student writing.*Applied Linguistics, 22*, 173-189.

Aston, G. (1997). Small and large corpora in language learning.In B. Lewandowska-Tomaszczyk & J. Melia (Eds.), *Practical applications in language corpora* (pp. 51-62). Łódź: Łódź University Press.

Barnes, A. (2002). Maintaining language skills in pre-service training for foreign language teachers. In H. Trappes-Lomax & G. Ferguson (Eds.), *Language in language teacher education* (pp. 199-214). Amsterdam: John Benjamins B.V.

Bayyurt, Y. (2010). Author positioning in academic writing. In S. Zyngier & V. Viana (Eds.), *Avaliaçoes E Perspectivas: Mapeando Os Estudos Empiricos Na Area deHumanas* [Appraisals and Perspectives: Mapping Empirical Studies in the Humanities] (pp. 163-184). Rio de Janeiro: The Federal University of Rio de Janeiro.

Bear, J. (1992). Context and content in English language teacher education. In A. J Mountford & H. Umunç (Eds.), *Tradition and innovation: ELT and teacher training in the 1990's*: Volume 2 (pp. 24-34). Ankara: The British Council.

Biber, D. (2006). *University language: A corpus-based study of spoken and written register*. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Biber, D., & Conrad, S. (2009). *Register, genre, and style.* Cambridge: Cambridge University Press.

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (2007). *Longman grammar of spoken and written English (Sixth Impression).* Essex: Pearson Education Limited.

Carletta, J., Kilgour, J., O'Donnell, T., Evert, S., & Voormann, H. (2003). The NITE object model library for handling structured linguistic annotation on multimodal data sets.*Proceedings of the EACL Workshop on Language Technology and the Semantic Web (3rd Workshop on NLP and XML, NLPXML-2003),* Toulouse.

Celce-Murcia, M., & Larsen-Freeman, D. (1998). *The grammar book: An ESL/EFL teacher's course (Second Edition)*. Boston: Heinle & Heinle Publishers.

Connor, U., & Precht, K. (1998). Business English: learner data from Belgium and the U.S. In J. Hung & S. Granger (Eds.), *Proceedings of the First Annual Symposium on Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching, 14-16 December 1998* (pp. 25-33). Hong Kong: Department of English, The Chinese University of Hong Kong.

Demirel, Ö. (1989). Yabancı dil öğretmenlerinin yeterlikleri. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi*, *4*, 5-26.

Demirel, Ö. (1990). Yabancı dil öğretmenlerinin yeterlikleri. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi, 5*, 133-161.

Doğançay-Aktuna, S. (1998).The spread of English in Turkey and its current sociolinguistic profile.*Journal of Multilingual and Multicultural Development*, *19*(1), 23-39.

Dulay, H. C., Burt, M. K., & Krashen, S. (1982). *Language Two.*New York: Oxford University Press.

Ellis, R. (2001). *The study of second language acquisition*.Oxford: Oxford University Press.

Gorard, S. (2003). *Quantitative methods in social science*. London: Continuum.

Granger, S. (1998). The International Corpus of Learner English: A new resource for foreign language learning and teaching and second language acquisition research. *TESOL Quarterly, 37*(3), 538-546.

Granger, S. (2002). A bird's-eye view of learner corpus research. In S. Granger, J. Hung, J. & S. Petch-Tyson (Eds.), *Computer learner corpora, second language acquisition and foreign language teaching* (pp. 3-33). Amsterdam/ Philadelphia: John Benjamins.

Granger, S. (2003). Error-tagged Learner Corpora and CALL: A Promising Synergy. *CALICO Journal*, *20*(3), 465-480.

Gürbüz, N. (2006). Differing perceptions of pre-service English teachers' strengths and weaknesses in the practicum: A case study in Turkey. *English language teacher education and development*, *9*, 39-46.

Hanson, J. (2011). Teacher reflection and identity – teaching a language from within an L2 cultural identity, or teaching from within L1 culture about L2. *The Journal of Language Teaching and Learning*, *1*, 1-38

Hatipoğlu, Ç. (2010). Summative evolution of an undergraduate 'English Language Testing and Evaluation' course by future English language teachers. *English Language Teacher Education and Development (ELTED),* 13 (Winter 2010), 40-51.

Hatipoğlu, Ç. (2011, July). Indirectness in L2 exam papers: Characteristics of the English that native speakers of Turkish use on subject exams. Paper presented at the *Sixth International Symposium on Politeness: Corpus Approaches,* METU, Ankara, Turkey.

Hewings, M. (1999). *Advanced grammar in use: a self-study reference and practice book for advanced learners of English*. Cambridge: Cambridge University Press.

Işık, A. (2008). Yabancı dil eğitimimizdeki yanlışlar nereden kaynaklanıyor? *Journal of Language and Linguistics, 4*(2), 15-26.

James, C. (1998). *Errors in language learning and use*. London: Addison Wesley Longman.

József, H. (1998). *Advanced writing in English as a foreign language: A corpus-based study of processes and products*. Unpublished PhD dissertation. Janus Pannonius University, Pécs, Hungary.

Kırkgöz, Y. (2009). Globalization and English language policy in Turkey. *Educational Policy, 23*(5), 663-684.

Kızıldağ, A. (2009). Teaching English in Turkey: Dialogues with teachers about the challenges in public primary schools. *International Electronic Journal of Elementary Education, 1*(3), 188-201.

Leech, G., & Svartvik, J. (2003). *A communicative grammar of English (Third edition).*Essex: Pearson Education Limited.

Lipton, G. (1996). FLES* Teacher preparation: Competencies, content and complexities. In Z. Moore (Ed.), *Foreign language teacher education* (pp. 37-58). USA: University Press of America.

Lüdeling, A., Walter, M., Kroymann, E., & Adoplphs, P. (2005).Multi-level error annotation in learner corpora.*Proceedings of the Corpus Linguistics 2005 Conference*. Birmingham, UK, 14-17 July 2005 (Retrieved from: https://linguistik.huberlin.de/institut/professuren/korpuslinguistik/forschung/falko/pdf/FALKO-CL2005.pdf)

Mouranen, A. (2003). The corpus of English as a Lingua Franca in academic settings.*TESOL Quarterly, 37*(3), 513-527.

Oğuz, E. (1999). *İlköğretimde yabancı dil (İngilizce) öğretimi sorunları* [The Problems of foreign language (English) teaching in elementary schools].Unpublished Master's Thesis. Kocaeli University: Kocaeli, Turkey.

Oxford, R., & Shearin, J. (1994). Language learning motivation: Expanding the theoretical framework. *The Modern Language Journal*, *78*(1), 12-28.

Paker, T. (2007). Problems of teaching English in schools in Çal Region and suggested solutions. *21. yüzyıla girerken geçmişten günümüze Çal yöresi: Baklan, Çal, Bekilli*. *Çal Yöresi Yardımlaşma ve Dayanışma Derneği Yayını, 3*, 684-690.

Quirk, R., & Greenbaum, S. (1976). A *universal grammar of English.*Essex: Longman.

Schiller, A, Teufel, S., & Thielen, C. (1995) *Guidelines für das Tagging deutscher Textcorpora mit STTS Technical report.* IMS Stuttgart / Seminar für Sprachwissenschaften, Tübingen.

Schmidt, T. (2004). EXMARaLDA - ein System zur computergestützten Diskurstranskription. In A. Mehler & H. Lobin (Eds.), *Automatische Textanalyse.Systeme und Methoden zur Annotation und Analyse natürlichsprachlicher Texte* (pp. 203-218). Wiesbaden: Verlag für Sozialwissenschaften

Selinker, L. (1972). Interlanguage.*IRAL, 10*(3), 209-231.

Sowden, C. (2007). Culture and the 'good teacher' in the English language classroom. *ELT Journal, 61*(4), 304-310.

Şallı-Çopur, D. (2008). *Teacher effectiveness in initial years of service: A case study on the graduates of METU Foreign Language Education program*. Unpublished doctoral dissertation. Middle East Technical University: Ankara, Turkey.

Thomas, A. L. (1987). Language teacher competence and language teacher education. In R. Bowers (Ed.), *Language teacher: An integrated programme for ELT teacher training: ELT Documents 125* (pp. 33-42). Oxford: Modern English Publications in associations with the British Council.

Tılfarlıoğlu, F. Y., & Öztürk, A. R. (2007). An analysis of ELT teachers' perceptions of some problems concerning the implementation of English Language Teaching Curricula in elementary schools. *Journal of Language and Linguistic Studies, 3*(1), 202-217.

Tono, Y. (2003). Learner corpora: Design, development and application. *Proceedings of the Corpus Linguistics 2003 Conference* (pp. 800-809). Lancaster, UK, 28-31 March 2003.

Turner, T. (2002). What is metadata? *Kaleidoscope, 10*(7), 1-3.

Weil, M., & Pullin, P. (2011). English as a lingua franca in education - Internationalisation speaks English. *Education Permanente, 1*, 28-29.

Woods, D. (1996). *Teacher cognition in language teaching*. Cambridge: Cambridge University Press.

ODTÜ Türkçe İngilizce Sınav Derleminin Oluşturulmasındaki İlk Aşamalar (METU TEEC)

Özet

Bu çalışmada, Orta Doğu Teknik Üniversitesi (ODTÜ) Ankara, Türkiye'den bir araştırma ekibi tarafından derlenen Orta Doğu Teknik Üniversitesi Türkçe İngilizce Sınav Derlemi'nin (METU TEEC) oluşturulmasındaki ilk aşamalar anlatılmaktadır. Derlem, ODTÜ İngiliz Dili Eğitimi bölümü (FLE) öğrencileri tarafından Ocak 2005 ve Aralık 2012 arasında belli bir sürede, hiçbir referans kullanılmadan yazılan 1914 Dilbilimi ve ELT sınav kağıdından (955483 sözcük) oluşmaktadır. Derlemin gerek kuramcılar/araştırmacılar gerek uygulamacı/pedagogların ihtiyaçlarına yanıt vermesi amaçlanmıştır; bu nedenle derlemdeki her metin, katılımcılar (ör. yaş, cinsiyet, İngilizce öğrenme süresi), sınav (ör. ders, soru çeşidi, akademik yıl) ve sınav yazarları (ör. eğitim geçmişi, deneyim) hakkında detaylı ara bilgiler açısından etiketlenmiştir. Sistem, araştırmacılar ve uygulayıcıların ihtiyaçları ve ilgilerine göre alt-derlemler toplanmasına olanak vermektedir. Yazının ilk bölümlerinde böyle bir derlem için gerekli ortam ve şartlara yer verilirken, sonraki bölümlerde bu derlemde kullanılan kodlama ve etiketleme sistemlerini tartışılmaktadır.

Anahtar sözcükler: Gelişmiş derlem oluşturulması, öğrenci derlemi, sınav İngilizcesi, metin çeşidi, Türkiye.