ASIALEX
The Asian Association for Lexicography

廣东外語外貿大學
GUANGDONG UNIVERSITY OF FOREIGN STUDIES

KEY RESEARCH INSTITUTE IN UNIVERSITY

ASIALEX 2017

June 10–12, 2017 · Guangzhou · China

Proceedings of the 11th International Conference of the Asian Association for Lexicography

Lexicography in Asia: Challenges, Innovations and Prospects

Edited by Hai Xu
Center for Linguistics and Applied Linguistics
Guangdong University of Foreign Studies

Preface

All of us in China are proud to host the Asian Association for Lexicography (ASIALEX) Conference again after it has traveled around nine Asian countries and regions in the span of twenty years. The 11th Conference of ASIALEX (ASIALEX 2017, Guangzhou, June 10-12, 2017), organized by the National Key Research Center for Linguistics and Applied Linguistics at Guangdong University of Foreign Studies thus represents a happy opportunity for ASIALEX to celebrate the 20th anniversary of its founding.

Besides receiving felicitations from the Presidents of our global sister associations AFRILEX, AUSTRALEX, DSNA, and EURALEX, we have invited four world-renowned lexicographers as our keynote speakers:

- Prof. Jianhua Huang of Guangdong University of Foreign Studies, the First President of ASIALEX,
- Dr. Michael Rundell, Editor-in-Chief of Macmillan Dictionary,
- Prof. Andrea Abel of EURAC Research, President of EURALEX, and
- Dr. Julia Miller of Adelaide University, President of AUSTRALEX.

We have organized two advanced workshops, on Sketch Engine and DPS5, which will be run by Mr. Miloš Jakubíček, CEO of Lexical Computing, and by Mr. Holger Hvelplund, Vice President of Digital Solutions, IDM, respectively.

The theme of ASIALEX 2017 is Lexicography in Asia: Challenges, Innovations and Prospects. We think that it is timely to recognize our achievements in lexicographic research and practice in the past 20 years in Asia, and to look ahead to see how we can respond to the challenges of the revolutions in corpus linguistics and digital lexicography we are currently facing. In the four keynote speeches, Huang and Abel speak on the common theme of dictionary user orientation/participation in the digital age, and Rundell and Miller discuss extended units of meaning or phraseology, which lexicographers are increasingly aware of as representing the norm, rather than the exception, in language. All the issues the keynote speakers address are cutting-edge concerns, and most certainly deserve our special attention.

The enthusiasm of scholars and publishers from Asia and beyond that has greeted

this conference has been unexpectedly high. As one of the largest conferences in its series, ASIALEX 2017 hosts approximately 160 participants from 75 institutes over 24 countries and regions in Asia, Europe, Africa and North America. We received an astounding number of 130 abstract submissions. This volume of proceedings, which is 915 pages long, consists of 64 full papers and 49 abstracts, which are roughly divided into the sections digital lexicography, general-purpose lexicography, cognitive approaches to lexicography, bilingual lexicography, pedagogic lexicography, specialized lexicography, and historical lexicography. We are truly indebted to the contributors and the abstract reviewers for their hard work in bringing together such a remarkable collection.

While preparations for this grand event were under way, we sadly lost two great lexicographers who were highly influential in both China and the world, Professor Gusun Lu of Fudan University, who passed away on July 28, 2016, and Professor Boran Zhang of Nanjing University, who passed away on May 26, 2017. They both made enormous contributions to our field. To honour their great achievements, we have therefore set up a special session in their memory, and also dedicate this volume to these two great colleagues.

Finally, I would like to thank my PhD students, Yongfang Feng, Huilian Hu, Lingling Li, and Ziyue Chen, for assisting me in editing the proceedings. Ms. Yongfang Feng also painstakingly proofread the whole text. I am also grateful to my colleagues Prof. Martin Weisser and Dr. Vincent Ooi who helped revise some parts of the text.

Hai Xu

Chair, the 11th International Conference of the Asian Association for Lexicography (ASIALEX 2017)

June, 2017

CONTENTS

Lexicological Module of Do It Yourself Corpora for Turkish

Bülent ÖZKAN
Mersin University
ozkanbulent@gmail.com

Abstract

The aim of this paper is to introduce the lexicological module of a corpus platform, which is flexible according to the research questions of scholars, and which is specific to the scholar; is user friendly, and corpus-database. Considering this perspective, it is expected to provide a corpus platform in which the results of the research can be obtained in a functional way. In addition to these, in lexicographic studies, the system can present a corpus database for users as a corpus output.

The study, which is named as *To Built Do It Yourself Corpora for Turkish (DIYCT)*, supported by TÜBİTAK* 1005 New National Ideas and Products Research Support Group. In this study the *Lexicological Module of Do It Yourself Corpora for Turkish* will be introduced. Firstly the outline of compiling a corpus by using *DIYCT* will be introduced and using *Lexicological Module* will be present to the researchers by explaining adding application like etymological knowledge, collocations, run-ons, spelling and if necessary voice and picture file adding and also other steps of compelling a dictionary.

Keywords: *lexicography, corpus linguistic, DIY corpora*

1. Introduction

It has been proved that corpora are important resources for linguistics studies. Almost in all linguistics disciplines, corpora have achieved to open new areas of research or to bring new insights to many traditional research questions (Meyer, 2004).

Currently, applications of corpus linguistics are used mainly in lexicography and lexical studies in parallel with applied linguistics. Additionally, these applications are used in other fields such as grammatical studies, register variation and genre analysis, historical studies, translation studies, diachronic studies, language change, language learning and teaching, semantics, pragmatics, sociolinguistics, discourse analysis, stylistics and literary studies, forensic linguistics etc. (McEnery, 2006: 80-122).

* The Scientific and Technological Research Council of Turkey

From the earlier text collections to large linguistics database, named as corpus, the studies of lexicography evolved to the empirical perspective. “Today, advances in computer technology have given several advantages to the corpus-based lexicographic research over earlier work. First of all, computers have made possible to the collection and storage of big chunks of texts and so that analyses are not limited to sentence-length excerpts.” (Biber, 2002: 22). With the ability of designing large corpora is provided to describe the language for lexicographer. In addition to this, “… using corpora allows dictionary-makers to extract all authentic, typical examples of the usage of a lexical item from a large body of text in a few seconds.” (McEnery, 2006: 80).

The structure of corpus design provides particular knowledge about lexical item that are examined. According to metadata of the corpus or structured datum, researcher can get this knowledge automatically such as frequency, co-occurrence, and collocations, key words in context (KWIC); register, genre and domain and part-of-speech…etc. Researcher also can build semantic patterns of lexical units while querying lexical items from a corpus.

On the other hand, in corpus linguistics literature, many researchers (Kennedy, 1998; McEnery, T. et. al. 2006; Sinclair, 1991; McEnery et. al. 1996; Barnbrook, 1996) have underlined the importance of determination of the motivation of corpus design and according to this, they have emphasized the importance of research questions. Moreover, the corpus have evaluated as data sets, which are used in linguistics researches in corpus linguistics literature (Nesselhauf, 2005; Gries, 2006; Kawaguchı, 2004; Dale, 2000; Scott, 2006; Sterkenburg, 2003).

Experts, suggest to use specific softwares or web aplications in lexicological corpus linguistics studies for standart-simple outputs such as frequency, KWIC etc. (Sinclair, 1991; Stevens, 1995; Todd, 2001); but these outputs cannot be adequate for lexicological studies. There are same specific patterns to form a headword in a dictionary. Some of these are *spelling, pronunciation, inflections, word class, senses, definition, examples, usage, run-ons, etymology* etc. (Jackson, 2002: 26-27; Hanks, 2003: 56-57). At this point, the important issue is to decide which of these patterns will be take place in a headword. The determinative approach is the aim of lexicological studies -in other word, research questions- and according to this, decision of dictionary-makers.

Even in this case, a dictionary-maker needs to use a corpus tool, which is adequate for completing her/his studies. Unlikely, lack of a database and/or a storage support is the biggest disadvantages of corpus tools. Another difficulty that dictionary-makers face to face is the learning period and adaptation period of these tools.

2. DIY Corpora Project for Lexicological Issues

The aim of the project DIYCT is provide a database-supported corpus, which can be shaped according to the research questions, and this corpus is specific and user friendly

for the scholar. Considering this perspective, this corpus platform can generate flexible reports according to tagged corpus units for the linguistics researcher. By using DIYCT, linguistics researcher can built a Turkish corpus and also can tag this corpus via the *Lexicological Module, Semantic Module, Syntactic Module, Morphological Module, Discourse Analysis Module, and Learner Corpora Modules* (http://kkd.mersin.edu.tr/index.php?dil=en).

After building a corpus in *Lexicological Module*, researchers can form the headword patterns, according to their research questions and can tag the headwords pattern in the *Lexicological Module*. Researchers can tag the patterns of headwords to the units that are determined via the *Lexicological Module*.

As soon as the corpus is built via DIYCT, the platform automatically process the texts as lemma, deduplication, and makes morphological analysis, frequency analysis, parse the sentences, shows the n-gram and collocation computing in a few seconds.

After these steps, the data that will be used for lexicological research becomes available for headwords tagging. These processes can be summarised as below:

I. Building a corpus

- Definition of layers and metadata of the corpus
- Uploading the texts to corpus

II. Standard corpus processes

- Lemmatization and stemming
- Deduplication
- Morphological analysis
- Frequency analysis
- Parsing the sentences
- N-gram and collocation computing

III. Lexicological Module

- Lemmatization and stemming
- Definition of flexible tags of the headwords (spelling, pronunciation, inflections, word class, senses, definition, examples, usage, run-ons, etymology etc.)
- Data processing
- Dictionary-makers can report the headwords as output from the module via flexible tagging.

3. Lexicological Module Processes

As it is mentioned above (III. *Lexicological Module*) this process consists of three steps. The first step is lemmatization and stemming, second one is definition of flexible tags headwords, and the last one is data processing. These stages are shown in Figure 1.

Figure 1 Processes of lexicological module.

- *Lemmatization and stemming process:*

This process is the main source of headwords of dictionary via lemmatization and stemming interface. Because of agglutinative structure of Turkish, stemming is important to determine the headwords of dictionary.

In conclusion, respectively lemmatization, stemming, and headwords steps can be structured for a dictionary via the interface of DIYCT software (Figure 2). The stemming list is shown in Figure 3.

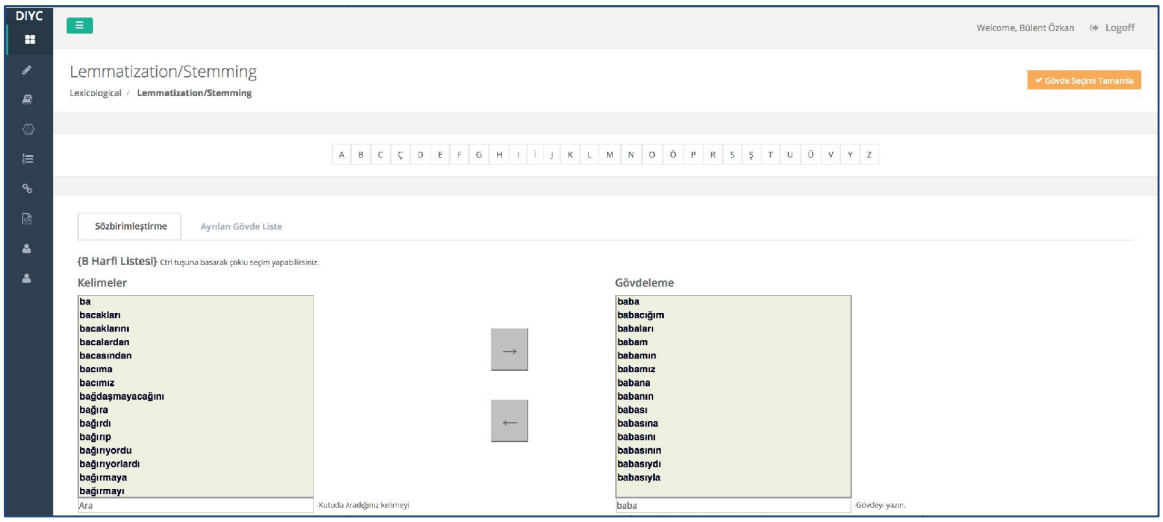

Figure 2 Interface of lemmatization and stemming

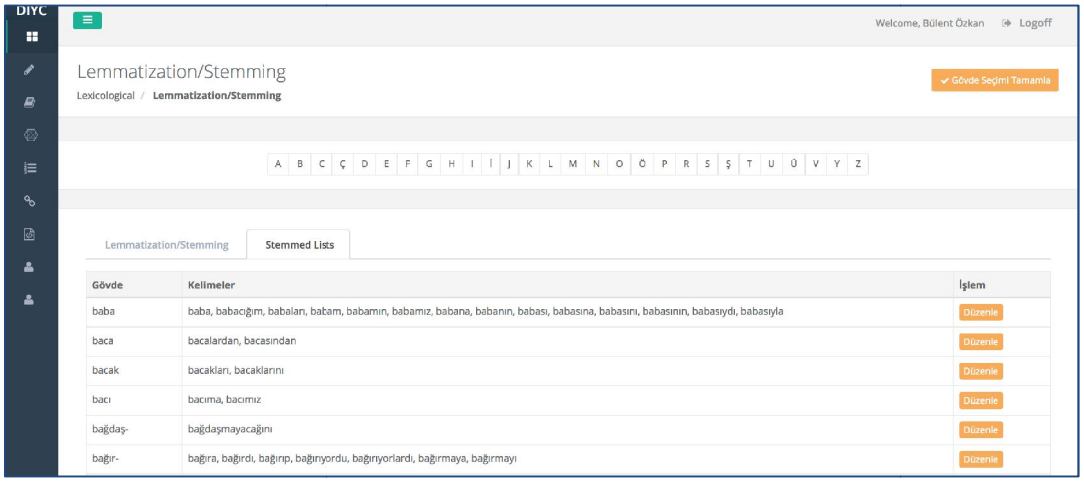

Figure 3 Stemming list of headwords.

- *Definition of flexible tags of the headwords (spelling, pronunciation, inflections, word class, senses, definition, examples, usage, run-ons, etymology etc.):*

The flexible tagging interface of DIYCTL (Figure 4) can determine the dictionary structure tags such as *spelling, pronunciation, inflections, word class, senses, definition, examples, usage, run-ons, etymology* etc. after the lemmatization and stemming processes in the basis of research question and/or dictionary researcher's (dictionary-maker) aim. On the other hand, other tagging such as voice and picture files can be added through this interface.

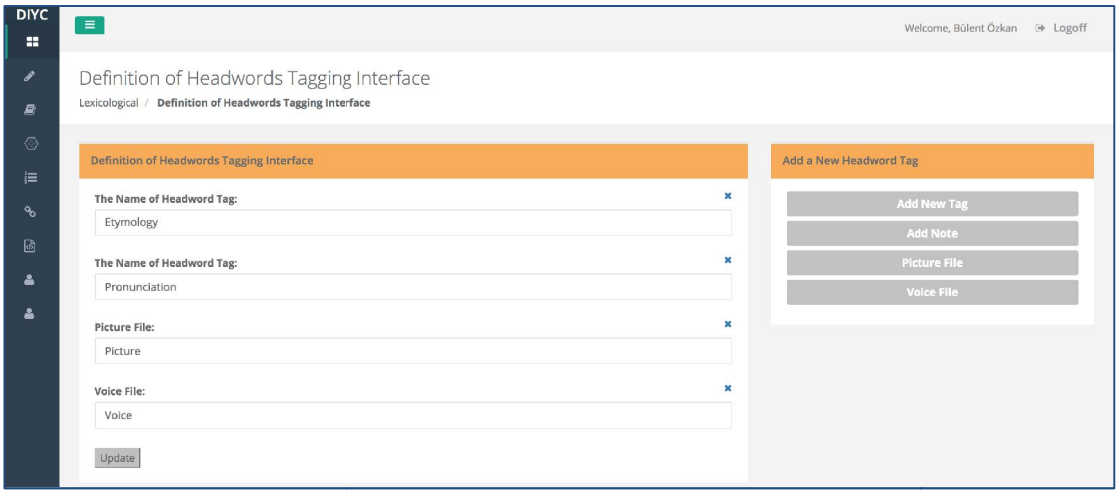

Figure 4 Interface of flexible tagging.

- *Data processing:*

In this stage, the stemmed headwords are listed for data processing. The data processing sample of "baba" (father) headwords is shown in Figure 5. The tags that have been determined at the *Definition of flexible tags* stage, like meanings of headwords and choosing examples (see also Figure 6.); part of speech tagging, pronunciation/spelling of headwords, compound word structure, etymology, picture

file (see also Figure 7.) and collocational structures (see also Figure 8.) can be tagged via these interfaces.

Figure 5 Data processing of "baba" (father) headwords.

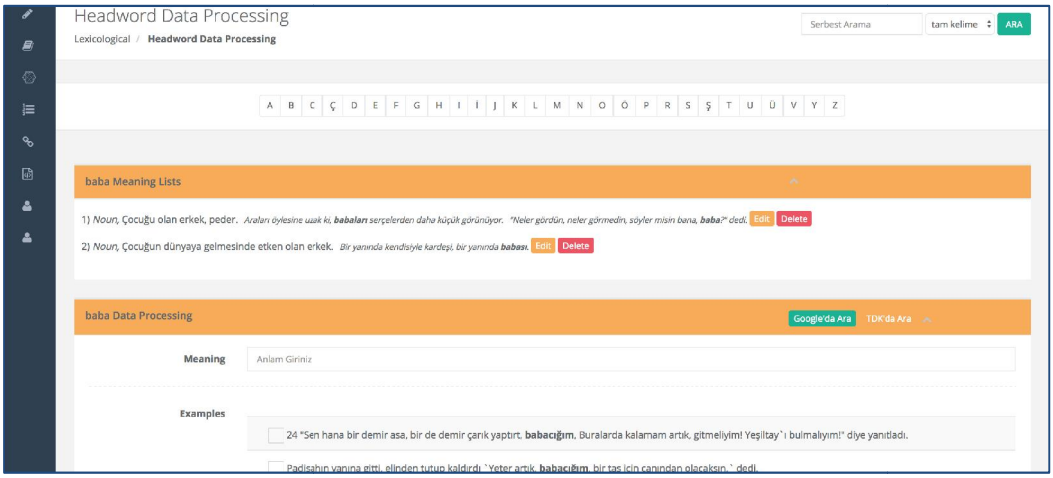

Figure 6 Interface of meanings and examples tagging.

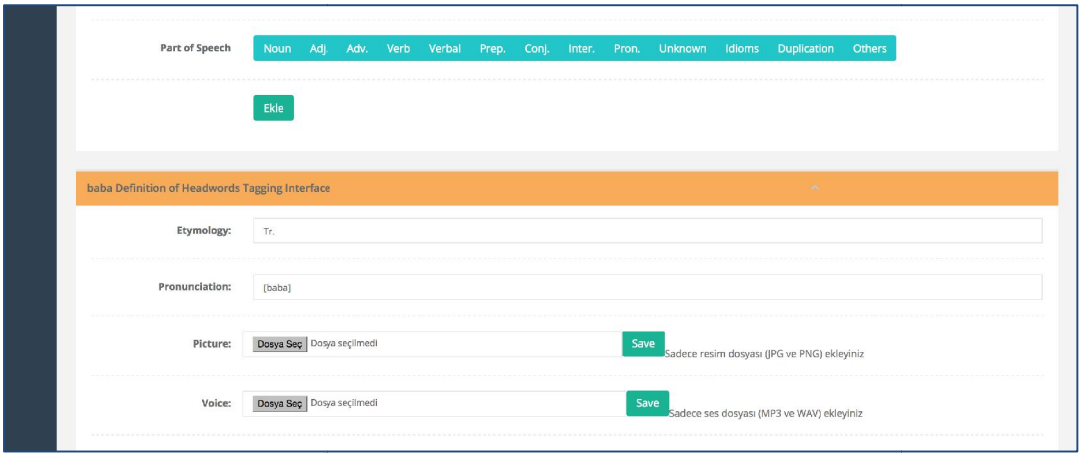

Figure 7 Interface of etymology, part of speech, pronunciation/spelling, etc.

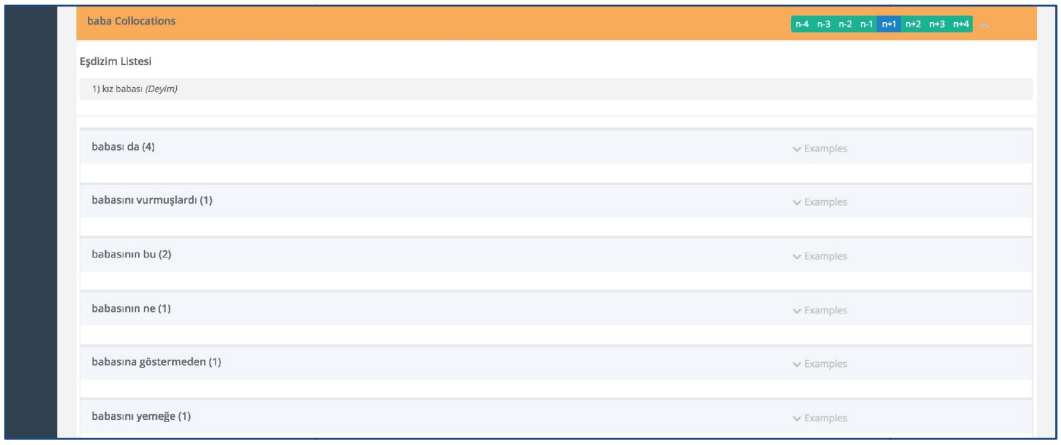

Figure 8 Interface of collocational structure (n-4 .... n+4).

- *Flexible report:*

Consequently, the data processing of headwords, researcher can take flexible report from the *Lexicological Module* of DIYCT as an output (see Figure 9). System allows flexible tagging for dictionary-makers to built headwords and according to this flexible tagging, the determined headwords are always available as an output (doc. docx etc.) for the researcher.

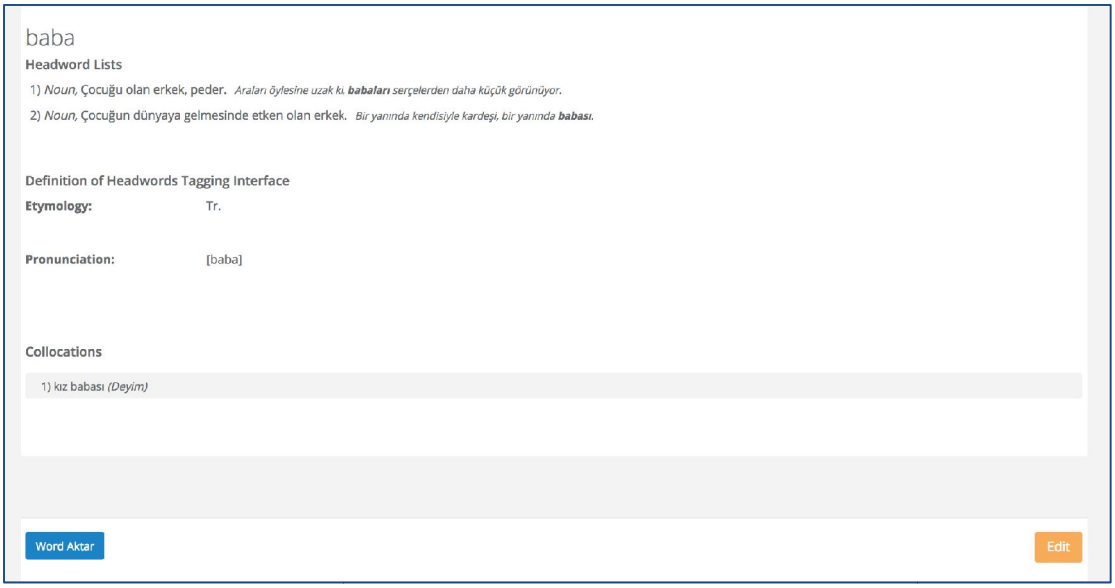

Figure 9 Flexible headwords report of lexicological module of DIYCT.

Acknowledgments

This study, which is presented and named as *To Built Do It Yourself Corpora for Turkish*, supported by TÜBİTAK 1005 *New National Ideas and Products Research Support Group*. Many thanks for this contribution to TÜBİTAK.

References

Baker, P., Andrew H. and Tony M. 2006. *A Glossary of Corpus Linguistics.* Edinburgh University Press.

Barnbrook, G. 1996. *Language and Computers: A Practical Introduction to the Computer Analysis of Language*. Edinburgh: Edinburgh University Press.

Biber, D. et. al. 2002. *Corpus Linguistics, Investigating Language Structure and Use*. Cambridge: Cambridge University Press. UK.

Burnard, L. 2004. *Metadata for Corpus Work. Developing linguistic corpora: A guide to good practice*. Oxford: Oxbow Books. (Available online from http://ahds, ac. uk/linguistic-corpora/.

Dale, R. et. al. 2000. Handbook of Natural Language Processing . New York: NY. USA.

Gries, S. T. et. al. 2006. Corpora in Cognitive Linguistics: Corpus-Based Approaches to Syntax and Lexis. Berlin: DEU.

Hanks, P. 2003. *Lexicography. Computational Linguistics*. Ed. Ruslan Mitrov. Oxford: University Press.

http://kkd.mersin.edu.tr/index.php?dil=en

Jackson, H. 2002. *Lexicography: An Introduction*. Routledge: USA.

Kawaguchi, Y. et. al. 2005. *Corpus-Based Approaches to Sentence Structures*. Philadelphia: PA. USA.

Kennedy, G. 1998. *An Introduction to Corpus Linguistics*. New York: Addison Wesley Longman Limited.

McEnery, T., Wilson A. 1996. *Corpus Linguistics.* Edinburg: Edinburg University Press.

McEnery, T., Richard X. and Yukio T. 2006. *Corpus-Based Language Studies an Advanced Resource Book*. Routledge: New York.

Meyer, C. F. 2004. *English Corpus Linguistics an Introduction*. Cambridge: Cambridge University Press. UK.

Nesselhauf, N., 2005. *Collocations in a Learner Corpus*. Philadelphia: University of Heidelberg: USA.

Özkan, B. et. al. (2016). Result report of Do İt Yourself Corpora for Türkish Languacge. Project number: 114E791. Mersin/TURKEY

Scott, M. 2006. *WordSmith Tools* [available at http://lexically.net/wordsmith/index.html]

Scott, M. et. el. 2006. *Textual Patterns: Key Words and Corpus Analysis in Language Education*. Philadelphia. PA. USA.

Sinclair, J. 1991. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Sterkenburg, P. 2003. *A Practical Guide to Lexicography* . Philadelphia, PA, USA: John Benjamins Publishing.

Stevens, V. 1995. Concordancing with language learners: Why? When? What? *CAELL Journal.* vol. 6(2). 2-10.

Todd, W. R. 2001. Building and using your own corpus and concordance. *Thai TESOL Bulletin* vol. 14 no. 2.