Journal of
Language and Translation
Volume 11, Number 3, 2021 (pp. 215-223)

Corpora and Translation Studies: Implications and Applications

Mehrdad Vasheghani Farahani [1]*, Hossein Vahid Dastjerdi [2]

[1]Ph.D., Leipzig University, Leipzig, Germany

[2]Associate Professor, English Department, University of Isfahan, Isfahan, Iran

Received: June 18, 2020 Accepted: January 15, 2021

INTRODUCTION

Concurrent with the advent of Mona Baker's seminal and trendsetter paper (1995) on the constructive role of corpora in translation research, we have witnessed an exponential revolution in this fast-growing strand of research (Fang, 2020; Vasheghani Farahani and Kazemian, 2021; Zanettin, 2012). What Baker envisaged in Corpus-based Translation Studies borrowed its roots from Descriptive Translation Studies (to study translation in accordance with target text boundaries) and was concomitant with the time when corpora were in their incipient stages of fruition in Applied Linguistics (Laviosa, 2013). As a matter of principle, the critical nexus between translation and corpora was underpinned based on the assumption that "hypotheses are tested by examining language in use rather than concocted examples" (Laviosa, 2013, p. 228).

By compiling a small English-Arabic parallel corpus, Baker established her theory of translation universals which can be summarized into four translation idiosyncrasies: explicitation, normalization, simplification, and leveling out. By explicitation, Baker meant adding information in the target text to change implicit information into explicit. Normalization is the process through which the norms of the source text tend to conform to those of the target text. Simplification consists of the fact that the target language becomes more simplified in terms of lexicogrammatical patterns, syntactic structures, and word selection for the ease of understanding of the target reader. Leveling out means "the tendency of translated text to gravitate towards the center of a continuum" (Baker, 1996, p.84).

Corpus by definition refers to an ensemble of texts which are compiled according to preplanned criteria and which are stored

* Corresponding Author's Email:
mehrdadfarahani1365@gmail.com

electronically in a systematic way (Paquot & Th. Gries & 2020). To put it differently, a corpus is a collection of texts in machine-readable form which can be automatically analyzed and processed by software package(s) (Timmis, 2015). Corpora must have some specifications in order to fully fit the research. As a matter of fact, corpus must be carefully designed which means that the data gathering regime must be clearly set beforehand (Egbert, Biber & Gray, 2022). By the same tokens, a corpus must be balanced; meaning that it must be compiled out of data proportional to such variables as text type, time of the data, genre, and the language pairs (for parallel corpora). Furthermore, a corpus must be necessarily representative which refers to the fact that it must be reflecting the language, to a great extent, it belongs to (Olohan, 2004).

Although most of the research in the domain of Corpus Linguistics falls within the scope of Contrastive Analysis (see for example Patterson, 2018 and Brock, Solano & Segundo, 2022) and Language Teaching and Learning (see for example 2016 Alonso-Ramos; 2016; Brezina & Flowerdew, 2017 and Charles & Frankenberg-Garcia, 2021), in the last decade and in line with the technology advances, corpora have found their way in Translation Studies. The methodology of corpora in translation includes data gathering, creating parallel corpora, generating hypotheses based on research questions, analyzing the (parallel) concordance lines, uncovering translation patterns, answering the research questions, and testing the hypothesis. In this thematic paper, two questions will be addressed: what contributions does Corpus Linguistics make in Translation Studies, and how one can use them in translation? What kind of corpora are used in translation and what techniques are utilized in this area of research?

What is a parallel Corpus?

As their names imply, parallel corpora are the most prototypical ones for conducting research in translation and interpretation (Kenning, 2010). As a matter of principle, parallel corpora are defined as at least two texts which have been aligned sentence by sentence or paragraph by paragraph in such a way that they can be compared and contrasted directly. Parallel corpora can be either uni-directional or bi-directional. In the former, the relationship between source and target texts is one way (for example from English to Persian), in the latter; however, the nexus is two ways (for example From English into Persian and from Persian to English). In the same fashion, parallel corpora can be either bilingual or multilingual. Bilingual, as is conspicuous, consists of language A (source text) and language B (target language), whereas, in a multilingual parallel corpus, there is one source language and more than one target language. Bilingual parallel corpora are a good benchmark for comparing translation into one target language; however, the multilingual parallel corpora are the mainstay for comparing more than one target text simultaneously.

English language (Source Text) → Persian Language (Target Text)

Figure 1

Uni-directional Parallel Corpus

English language (Source Text) ↔ Persian Language (Target Text)

Persian Language (Target Text) ↔ English language (Source Text)

Figure 2

Bi-directional Parallel Corpus

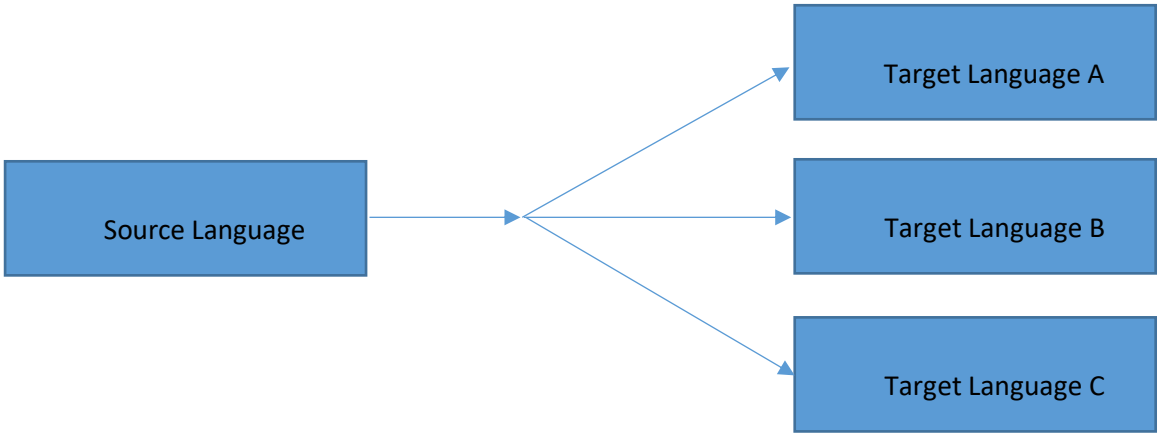
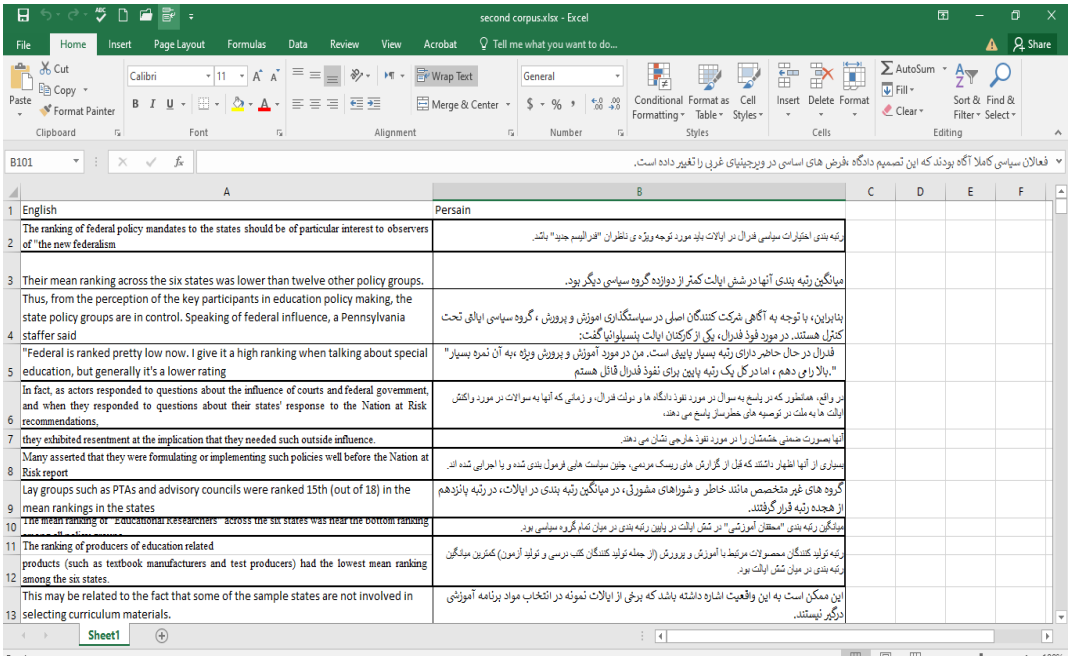


Figure 3

Multilingual Parallel Corpora

How to Create Parallel Corpora

Creating a parallel corpus is, inherently, an arduous kind of task that necessitates careful consideration and a full-fledged effort. The fundamental step in creating parallel corpora is the issue of text availability. To put it differently, as long as no translation is available from a so-called source language, creating a parallel corpus is impossible (Laviosa, 2002). Apart from text availability, the process of alignment is of paramount importance for compiling a parallel corpus. Alignment is the process in which the source and target languages are set against each other at sentence or paragraph levels. Although there are various online and web-based alignment tools (such as vanilla aligner, Microsoft bilingual sentence aligner, ABBYY, and Ugarit Aligner), it is still advisable to do the alignment manually or semiannually as the automatic alignment still needs a lot to be desired. It is advisable if texts are aligned in an EXCEL file for, they can be better processed in corpus software.



	A	B
1	English	Persain
2	The ranking of federal policy mandates to the states should be of particular interest to observers of "the new federalism	رتبه بندی اختیارات سیاسی فدرال در ایالات باید مورد توجه ویژه ی ناظران "فدرالیسم جدید" باشد.
3	Their mean ranking across the six states was lower than twelve other policy groups.	میانگین رتبه بندی آنها در شش ایالت کمتر از دوازده گروه سیاسی دیگر بود.
4	Thus, from the perception of the key participants in education policy making, the state policy groups are in control. Speaking of federal influence, a Pennsylvania staffer said	بنابراین، با توجه به آگاهی شرکت کنندگان اصلی در سیاستگذاری اموزش و پرورش ، گروه سیاسی ایالتی تحت کنترل هستند. در مورد نفوذ فدرال، یکی از کارکنان ایالت پنسیلوانیا گفت:
5	"Federal is ranked pretty low now. I give it a high ranking when talking about special education, but generally it's a lower rating	فدرال در حال حاضر دارای رتبه بسیار پایینی است. من در مورد آموزش و پرورش ویژه ،به آن نمره بسیار "بالا را می دهم ، اما در کل یک رتبه پایین برای نفوذ فدرال قائل هستم
6	In fact, as actors responded to questions about the influence of courts and federal government, and when they responded to questions about their states' response to the Nation at Risk recommendations,	در واقع، همانطور که در پاسخ به سوال در مورد نفوذ دادگاه ها و دولت فدرال، و زمانی که آنها به سوالات در مورد واکنش ایالت ها به ملت در توصیه های خطرساز پاسخ می دهند،
7	they exhibited resentment at the implication that they needed such outside influence.	آنها بصورت ضمنی خشمشان را در مورد نفوذ خارجی نشان می دهند.
8	Many asserted that they were formulating or implementing such policies well before the Nation at Risk report	بسیاری از آنها اظهار داشتند که قبل از گزارش های ریسک مرتبی، چنین سیاست هایی فرمول بندی شده و یا اجرایی شده اند.
9	Lay groups such as PTAs and advisory councils were ranked 15th (out of 18) in the mean rankings in the states	گروه های غیر متخصص مانند خاطر و شوراهای مشورتی، در میانگین رتبه بندی در ایالات، در رتبه پانزدهم از هجده رتبه قرار گرفتند.
10	The mean ranking of "Educational Researchers" across the six states was near the bottom ranking	میانگین رتبه بندی "محققان آموزشی" در شش ایالت در پایین رتبه بندی در میان تمام گروه سیاسی بود.
11	The ranking of producers of education related	رتبه تولید کنندگان محصولات مرتبط با آموزش و پرورش (از جمله تولید کنندگان کتب درسی و تولید آزمون) کمترین میانگین
12	products (such as textbook manufacturers and test producers) had the lowest mean ranking among the six states.	رتبه بندی در میان شش ایالت بود.
13	This may be related to the fact that some of the sample states are not involved in selecting curriculum materials.	این ممکن است به این واقعیت اشاره داشته باشد که برخی از ایالات نمونه در انتخاب مواد برنامه آموزشی درگیر نیستند.

Figure 4

A parallel corpus in the EXCEL file

Parallel Concordance Lines

One of the basic objectives of corpus-based translation is to look for differences and similarities between the source and target text(s). The similarities and differences of the source and target languages can vary in many aspects running the gamut of syntactic, lexicogrammatical, equivalence, stylistic, semantic, and pragmatic levels. This analysis of the source and target texts entails parallel concordance lines which are a set of source and target texts set against each other either horizontally or vertically (Mikhailov & Cooper, 2016). The reading of parallel concordance lines and making conclusion(s) is composed of two phases quantitative and qualitative. The quantitative phase entails a descriptive analysis of the source and target texts. In other words, in quantitative analysis, the unique features are counted through such techniques as keywords in the context, type-token ratio, word lists, and n-grams in order to compare the source and target languages. The qualitative analysis, on the other hand, is an interpretive phase that necessitates a close reading of the selected parallel concordance lines in order to draw conclusions. The close reading and analysis of the parallel concordance lines are done by the researcher in an effort to find answers to the questions as well as refute or accept the hypothesis.

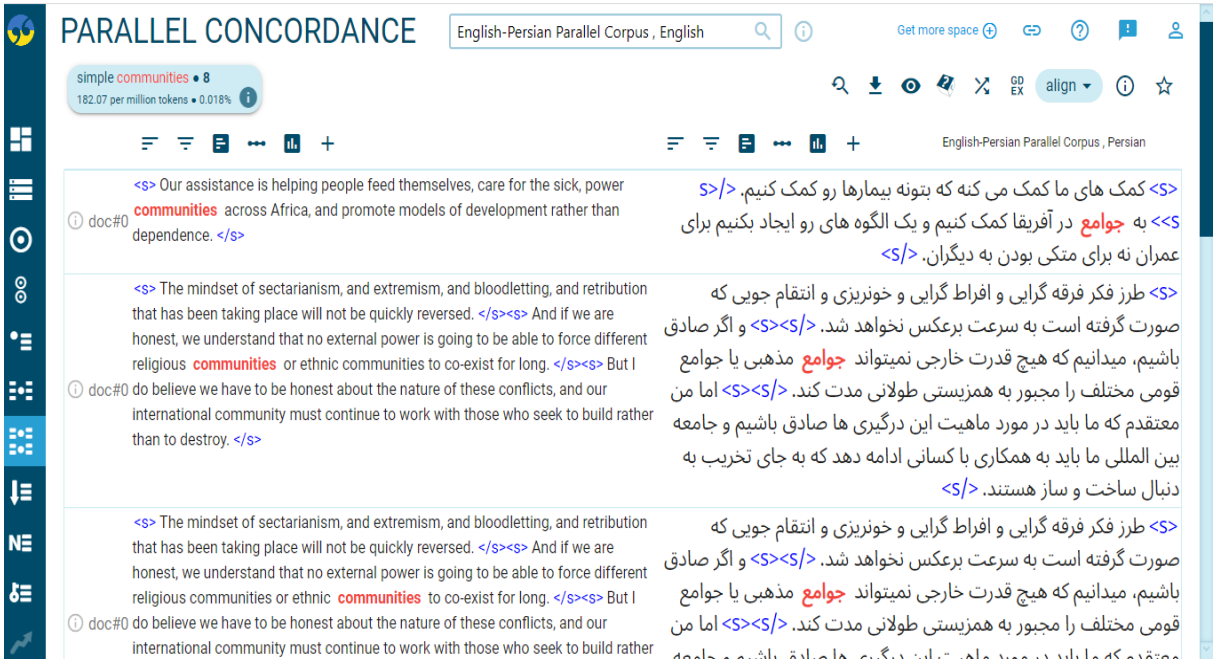


Figure 5

A screenshot of English-Persian parallel concordance lines

Combination of Parallel Corpora and Reference Corpora

There are various types of corpora pertaining to research questions and research boundaries. In the field of translation, in order to reach richer data and in an effort to utilize data triangulation (Malamatidou, 2017), sometimes, researchers in translation make use of more than one corpus. In most cases, the second corpus which is used in line with the parallel corpus is called the reference corpus. A reference corpus is a collection of authentic non-translated texts created for the purpose of comparing translations (translationese) with non-translations and to posit the similarities and differences between them. Compared to parallel corpora, reference corpora are usually larger (at least three times) and diverse in nature as their collection and compilation is less arduous than those of the parallel corpora. This is mostly due to the fact that there is only one language in creating reference corpora and no parallel concordance line and alignment is mandatory. Having said that the largeness of the reference corpora makes it possible for the researcher(s) to have a broader look at the language and phenomenon they are investigating.

Creating Parallel Corpora Out of Spoken Data

A scan of the literature review will prove the fact that most of the research in the domain of Corpus-based Translation Studies has been geared toward written corpora and there is a noticeable focus on written corpora (see for example Uzar, 2002; Bulter, 2008; Mousavi

Razavi, 2011 and Biel, Engberg, Ruano & Sosoni, 2019).

Notwithstanding the prominence of written corpora, spoken corpora are utilized in translation research (2002). Creating parallel corpora from spoken data is by far a nebulous kind of task when compared to that of the written parallel corpora. This arduousness stems from the fact that spoken data are raw data that has to be processed before they can be utilized as written data for corpus research. In recognition of the fact that spoken discourse including simultaneous and consecutive interpretation is a unique discourse with its own idiosyncrasies, creating a parallel corpus out of spoken resources is receiving the attention of researchers in interpreting studies (Halliday, 2004).

Before a spoken corpus can be analyzed, it must undergo various steps for the matter of practicality. Spoken data must be gathered and or recorded, transcribed, annotated, aligned, and analyzed. As far as the recording of the data is concerned, it is vital for the researcher to record the whole spoken materials. It is of paramount importance to use high-quality recording software for the spoken data must be optimally accurate and exhaustive (Knight and Adolphs, 2006). After the recording of the materials, it is essential to add metadata to better recognize the spoken data. The metadata includes such information as the time of the departure, the person(s) involved in the speaking, language pairs, and the place and duration of each recording. In the same manner, it is vital to add punctuation to the spoken corpus so as to identify the sentences clearly. Once the data are gathered, they have to be aligned at the sentence and/ or paragraph levels for comparing source and target languages.

What Corpora Cannot Tell Us?

Corpora, regardless of their types and versatility, have some limitations which need to be taken into consideration. Corpora represent only the data that they contain and every conclusion is made based on inputted data. In other words, one must be cautious in generalizing the conclusions beyond the data put into the corpus. Moreover, corpora will not be able to analyze the data on their own. Indeed, it is the responsibility of the researcher(s) to analyze the corpus data beyond the (parallel) concordance lines. In addition, corpora cannot be analyzed without extensive use of corpus and computerized software. In this regard, without having a deep knowledge of the corpus software, the results will become truncated and abortive. The last limitation consists of the fact that corpora are about the what of the matter not how of the matter. This means that the researcher must be able to interpret the concordance lines and the extracted data from the corpus with intuitive knowledge.

CONCLUSION

Corpora have a wide range of applications in language-related studies. Among the many applications of corpora are parallel corpora which are the most appropriate type of corpora for translation research. Despite the fact that creating parallel corpora is a complicated, time-

consuming, and even expensive process, thanks to technological advances, they are finding ways to translate research. Although corpora have a wide range of applications in research in translation, they can be extensively used in translation practice and in such activities as term extraction, finding appropriate equivalences, translation teaching, and translation quality assessment. Alonso-Ramos, M. (2016). Spanish learner corpus research: Current trends and future perspectives. John Benjamins Publishing Company.

References

Baker, M. (1995). Corpora in translation studies: An overview and some suggestions for future research. *Target: International Journal of Translation Studies, 7*(2), 223-243. DOI: 10.1075/target.7.2.03bak.

Baker, M. (1996). Corpus-based Translation Studies: The Challenges That Lie Ahead. In Terminology, LSP and Translation. Studies in Language Engineering in Honour of Juan C. Sager, H. Somers, ed., 175-186. Amsterdam: Benjamins.

Biel, Ł., Engberg, J., Ruano, R. M., & Sosoni, V. (2019). *Research methods in legal Translation and interpreting: Crossing methodological boundaries*. London and New York: Routledge.

Brezina, V., & Flowerdew, L. (2017). Learner corpus research: new perspectives and applications. Bloomsbury Publishing.

Brock, A., Solano, R. M., & Segundo, P. R. (2022). Anglicisms and corpus linguistics: Corpus-aided research into the influence of English on European languages. Peter Lang Gmbh, Internationaler Verlag Der Wissenschaften.

Bulter, C. S. (2008). The subjectivity of basically in British English – a corpus-based study. *Mouton Series in Pragmatics and Corpus Linguistics*, 37-63. DOI: 10.1 515/9 78311 019902 4.37.

Charles, M., & Frankenberg-Garcia, A. (2021). Corpora in ESP/EAP writing instruction: Preparation, exploitation, analysis. Routledge.

Egbert, J., Biber, D., & Gray, B. (2022). Designing and evaluating language corpora: A practical framework for corpus representativeness. Cambridge University Press.

Feng, H. (2020). Form, meaning and function in collocation: A corpus study on commercial Chinese-to-English translation. Routledge.

Halliday, M. A. K. (2004) The Spoken Language Corpus: A Foundation for Grammatical Theory, in K. Aijmer and B. Altenberg (eds) Advances in Corpus Linguistics. Amsterdam: Rodopi, pp. 11–38.

Kenning, M. (2010). What are parallel and comparable corpora and how can we use them? In Anne O'Keeffe, A. & McCarthy, M. The Routledge

Handbook of Corpus Linguistics (1st ed., pp. 487 -500). London: Routledge.

Knight, D. and Adolphs, S. (2006) Text, Talk and Corpus Analysis [academic online module, restricted access], University of Nottingham, UK.

Laviosa, S. (2002). *Corpus-based Translation Studies. Theory, Findings, Applications.* Amsterdam: Rodopi.

Laviosa, S. (2013). Corpus linguistics in translation studies. In Millán, C. & Bartrina, F. (2013). Routledge Handbook of corpus linguistics (1st ed.). London: Routledge.

Malamatidou, S. (2017). Corpus triangulation: Combining data and methods in corpus-based translation studies. Routledge.

Mikhailov, M., & Cooper, R. (2016). Corpus linguistics for translation and contrastive studies: A guide for research. Routledge.

Mousavi Razavi, M. S. (2011). On Fronted Themes in Translation of Dramatic Texts into Persian: A Corpus-based Study of Markedness in Translation. *Translation Studies Quarterly*, *9* (33). Retrieved from https://journal.translationstudies.ir/ts/article/view/466.

Olohan, M. (2004). Introducing Corpora in Translation Studies. London and New York: Routledge.

Paquot, M., & Gries, S. T. (2020). A practical handbook of corpus linguistics. Springer.

Patterson, K. (2018). Understanding metaphor through corpora: A case study of metaphors in Nineteenth-century writing. London: Routledge.

Timmis, I. (2015). Corpus linguistics for ELT: Research and practice. London: Routledge.

Teubert, W. (2002). The role of parallel corpora in translation and multilingual lexicography. In Atenberg, Bengt/ Granger, Sylviane (eds.), Lexis in Contrast. Amsterdam/Philadelphia: John Benjamins, 189-214.

Uzar, R. S. (2002). A corpus methodology for analyzing translation. *Cadernos de* Tradução, *9*(1), 235–263.

Vasheghani Farahani, M., & Kazemian, R. (2021). Speaker-audience interaction in spoken political discourse: A contrastive parallel corpus-based study of English-Persian translation of Metadiscourse features in TED talks. *Corpus Pragmatics, 5*(2), 271-298. https://doi.org/10.1007/s41701-021-00099-z.

Zanettin, Federico. (2012). *Translation-driven corpora: Corpus resources for descriptive and applied translation studies* (1st ed.). Routledge: London.

Biodata

Dr. Mehrdad Vasheghani Farahani holds a Ph.D. in English Translation Studies from Leipzig University, Germany. His area of research interest includes such areas as Corpus

Linguistics, Translation Studies, and Corpus-based Translation Studies. He has published extensively in international journals.
Email: *mehrdadfarahani1365@gmail.com*

Dr. Hossein Vahid Dastjerdi is an associate professor of applied linguistics and has taught courses of variegated character, including translation courses. He has been a fellow of the English Centers at the universities of Isfahan and Shiraz where he has investigated issues related to materials preparation for GE. and ESP. courses. He is the author of a number of books in this respect. He has also published a good number of articles on discourse, testing, and translation in local and international journals. He is Editor-in-Chief of the International Journal of Foreign Language Teaching and Research and a member of the editorial board of some Iranian and non-Iranian journals. Dr. Vahid's current research interests include testing, materials development, the metaphoricity of language, discourse analysis, pragmatics, and critical discourse analysis.
Email: *h_vahid@yahoo.com*