Do Speakers Produce Different Referring Expressions in Their Native Language Than A Non-native Language?

Imtiaz Hussain Khan, Muazzam Ahmed Siddiqui
King Abdulaziz University
Jeddah, Saudi Arabia
ihkhan@kau.edu.sa, maasiddiqui@kau.edu.sa

ABSTRACT: *This article explores empirically how Arabic speakers produce referring expressions in their native and non-native languages. Twenty six native Arabic speakers, who were also fluent in English, took part in the empirical study. Participants were presented with objects in visual domains and they were asked to describe the (marked) target object by typing a description which can uniquely identify the target to their addressee. The data reveal that there are no significant differences between referring expressions produced in native and non-native languages.*

Keywords: Referring Expressions, Native and Non-Native Differences, Under-Specification, Over-Specification, Arabic Referring Expressions

Received: 12 February 2015, Revised 8 March 2015, Accepted 18 March 2015

© 2015 DLINE. All Rights Reserved

1. Introduction

A Noun Phrase (NP) is called a referring expression if its communicative purpose is to identify an object to an addressee. We use referring expressions every now and then in both written and oral communication. Interestingly, different referring expressions can be used to describe the same object. For example, consider the visual context/scene in Figure 1 where there are three persons: two with glasses, and one without glasses. Suppose we want our addressee to meet with the person in the middle (without specifying his name). We could say: please meet with *the bearded man*. In this context, just saying *the man* would not help our addressee meet the (right) intended person. In this case, there is a choice though: one can describe the intended person in many different ways, including, for example, *the bearded man*, *the bearded man with glasses*, *the man with long hair*, or even *the man in the middle*. All these referring expressions are valid and can serve the purpose. So, it is important to decide which of these three referring expressions is more appropriate in a given context.

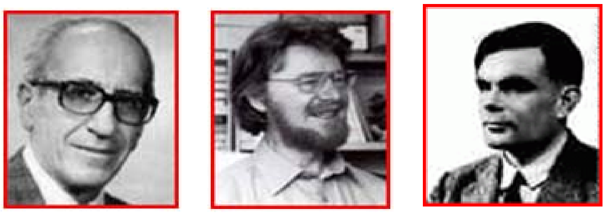


Figure 1. A referential domain

Generation of Referring Expressions (GRE) is an important component of most Natural Language Generation (NLG) systems (Reiter & Dale, 2000). In GRE, the focus is on developing algorithms which generate referring expressions to distinguish an entity (or a set of entities) from other entities in a given context. In the past three decades, various GRE algorithms have been developed, which mainly focused on generating descriptions which are as short as possible (for example, (Dale, 1992), (Gardent, 2002)) or almost as short as possible (Dale & Reiter, 1995); for details, see (Krahmer & Van Deemter, 2012).

Research in GRE has attracted considerable attention from both computational linguists and psycholinguists. The former researchers mainly focus on developing algorithms to produce efficient referring expressions in a given situation, while the latter intend to investigate empirically how human process such expressions. Empirical studies so far mostly confined to native speakers disregarding how the non-native speakers process such expressions. Some interesting work (Bortfeld & Brennan, 1997; Wu & Keysar, 2007; Luk, Xiao, & Cheung, 2012) is also reported in literature which takes native and non-native phenomena into account, paying little attention whether or not native speakers process referring expressions differently in their native language than a non-native language. For example, (Bortfeld & Brennan, 1997) studied the difference between native and non-native speakers in reference production and found that non-native speakers adjust themselves quickly to adapt to native speakers in a collaborative communication task. In the other study, (Wu & Keysar, 2007) found that Chinese participants were much more effective than English speakers at taking the speaker's perspective into account (see also (Luk, Xiao, & Cheung, 2012)). According to Wikipedia, the population of non-native English speakers is almost double as compared to native English speakers. Therefore, it seems imperative to see how non-native English speakers, e.g. Arabic speakers, process referring expressions. Arabic is the fourth most widely spoken language in the world (Nwesri, Tahaghoghi, & Scholer, 2005), with a rich morphology and syntax (Farghaly & Shaalan, 2009; Habash, 2010). Arabic Natural Language Processing (ANLP) is still in infancy, and there is very little work which focuses on Arabic Natural Language Generation (ANLG). One piece of work on ANLG is by (Shaalan, Monem, Rafea, & Baraka, 2008) who used a grammar-based approach to generate Arabic text, by using interlingua-based spoken dialogue. Their work provides a potential inroad for futuristic research in ANLG.

This study investigated empirically how Arabic speakers produce referring expressions, both in their native language (Arabic) and (non-native language) English. The purpose of this study is twofold. First, to investigate how referring expressions are processed in Arabic. The second purpose is to identify if there are any differences between referring expressions produced in the native language and the ones produced in the non-native language. To the best of our knowledge, this is the first study which compares the production of referring expressions in English and Arabic, focusing on Arabic speakers.

The rest of this article is organised as follows. A background overview of GRE is presented in Section 2. Section 3 describes the empirical study. The results are reported and discussed in Section 4. The article concludes in Section 5.

2. Background and Related Work

GRE is an active research topic among NLG researchers. Different GRE algorithms have been proposed in past three decades, and they have very similar formalization of the problem. They assume that a knowledge base consisting of a finite set *O* of objects is available; the objects in *O* are characterized as having a finite set *P* of properties. These properties are used to build descriptions, which are usually represented as sets of properties or logical formulae. With these assumptions, the GRE task can informally be stated as follows. Given an intended referent $r \in O$ (i.e., the object to be identified) and a set *D* (where $D \subseteq O$) of distractors (i.e., other objects that can be confused with the target referent), the task is to find a description (henceforth Distinguishing Description, DD) that allows a hearer to identify its referent uniquely (Dale, 1992).

One of the most widely studied GRE algorithm is Dale and Reiter's Incremental Algorithm (Dale & Reiter, 1995). The algorithm aims at generating DDs which are humanlike, and which can be generated efficiently. Unlike Full-Brevity Algorithm (Dale, 1992), which aims to construct the shortest possible description (also called a *minimal description*), the Incremental Algorithm may, sometime, generate over-specified descriptions (i.e., a description which contains more properties than absolutely required) just as speakers often do (Pechmann, 1989). Given an intended referent *r* and a set *D* of distractors, the algorithm iterates through a pre-determined ordered list *P* of properties, adding the next available property to the description of *r* only if it is true of *r* and at the same time it rules out some of the distractors that have not already been ruled out. The distractors that are ruled out are removed from *D*. The algorithm terminates when a DD for *r* is constructed or list of properties *P* is exhausted.

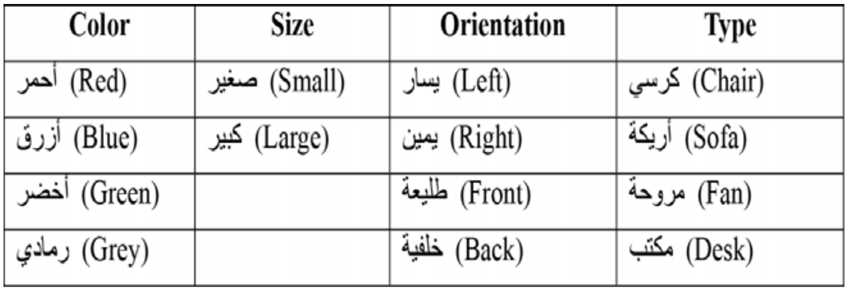
GRE community has seen substantial evaluation studies to evaluate different GRE algorithms proposed in literature (Mellish & Dale, 1998; Gupta & Stent, 2005; Van Deemter, Van der Sluis, & Gatt, 2006; Belz & Kilgarriff, 2006; Belz & Reiter, 2006; Paris, Colineau, & Wilkinson, 2006; Viethen & Dale, 2006; Gatt & Belz, 2008); for a detailed discussion see (Krahmer & Deemter, 2012). These studies generally focused on how close the output of a GRE algorithm is to human when they produce referring expressions in similar situations. The existing studies are mostly speaker-oriented, focussing on the degree of human-likeness, but there are few hearer-oriented studies as well, which focus on the effectiveness of referring expressions (Khan, Deemter, & Ritchie, 2012). The limitation of the existing evaluation studies is that they are mostly confined to English or other European languages; we are not aware of any studies which take Arabic referring expressions into account. Moreover, the existing evaluation studies are also lacking in how speakers produce referring expressions in non-native languages.

3. The Experiment

In this study, participants were presented with visual domains of objects and they were asked to describe the target object by typing a description similar in spirit to (Gatt, Van der Sluis, & Van Deemter, 2007). The participants produced referring expressions, at a broader level, in two different experimental conditions: 1) Arabic Expressions Condition, in which participants were asked to describe the intended object in Arabic, and 2) English Expressions Condition, in which they had to describe the same object in English. Participants were native Arabic students who took part in the experiment as part of their coursework.

3.1 Material and Design

Materials were constructed using the same pictures of furniture items which were used in the development of the TUNA corpus (Gatt, Van der Sluis, & Van Deemter, 2007). These items differed along four dimensions as shown in Table 1.



Color	Size	Orientation	Type
أحمر (Red)	صغير (Small)	يسار (Left)	كرسي (Chair)
أزرق (Blue)	كبير (Large)	يمين (Right)	أريكة (Sofa)
أخضر (Green)		طليعة (Front)	مروحة (Fan)
رمادي (Grey)		خلفية (Back)	مكتب (Desk)

Table 1. Attributes and values of the furniture items

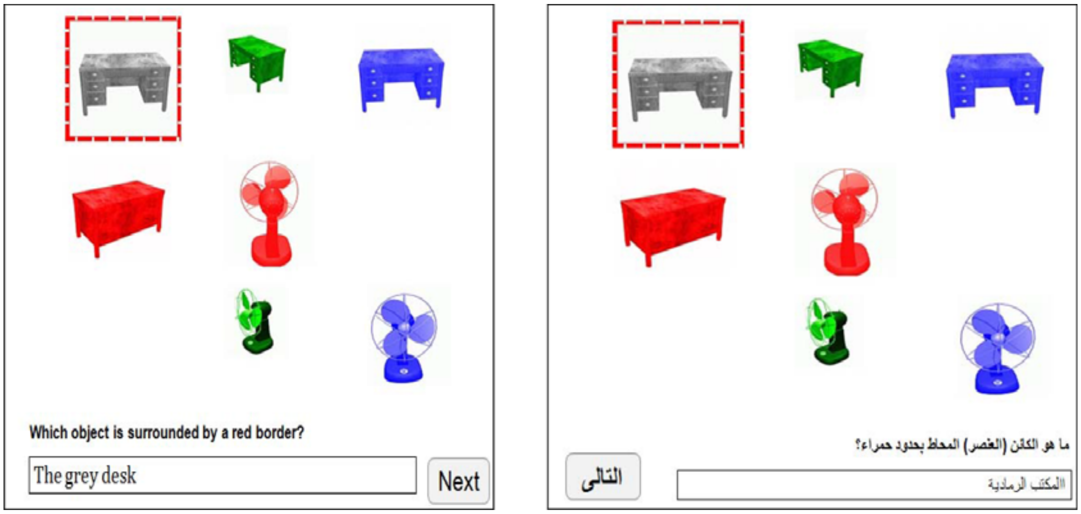
A trial in this experiment consisted of a collection of seven pictures, one target and six distractor objects as shown in Figure 2. The target referent was surrounded by a red border so that it could be easily distinguished from its distractors. Like the original TUNA experiment, the construction of trials was balanced in the sense that, for each possible combination of the properties, there was one trial in the relevant domain where that combination of properties was minimally required to distinguish the referent. For example, there was a trial in which the target could be distinguished using color only, another one in which the target could be distinguished using a combination of color and size, and so on. This arrangement gave us seven different combinations of minimal descriptions, henceforth fine-grained experimental conditions: CM: Color minimal condition, OM: Orientation minimal condition, SM: Size minimal condition, CSM: Color+Size minimal condition, COM: Color+Orientation minimal condition, SOM: Size+Orientation minimal condition, and COSM: Color+Orientation+Size minimal condition (for details, see (Khan, 2015)). There were total fourteen experimental trials, seven trials in the English Expressions Condition and seven in the Arabic Expressions Condition. The participants were advised (see below) to complete all the trials.

3.2 Participants and Procedure

Twenty six senior-year undergraduate native (Saudi) Arabic students took part in the experiment as part of their coursework. The participants were students of the Department of Computer Science, King Abdulaziz University. The experiment was carried out in a conducive laboratory at the University and it lasted for approximately 25 minutes for each participant.

Before running the experiment, the participants were briefed about the purpose and format of the experiment. The main part of the instructions was as follows:

In this experiment, we collect descriptions that people give when they have to describe an object, for example "The blue chair facing front". You will be asked to describe pictures in such a way that your addressee can uniquely identify them. .... In each collection of pictures, one object is marked by a red boarder. Beneath these pictures, you will see a question: Which object is



Which object is surrounded by a red border?

The grey desk

ما هو الكائن (العنصر) المحاط بحدود حمراء؟

المكتب الرمادية

Figure 2. An example experimental trial ((a): English Expressions Condition, (b): Arabic Expressions Condition)

surrounded by a red border?. Your task is to write the answer[1] *in the text box provided in a way as you are communicating to a person ...*

After introduction, the participants were presented with the experimental trials, one trial at a time. Each participant was presented with all fourteen trials in different pseudorandom order, alternating the two experimental conditions. They were allowed to withdraw from the experiment at any stage. Out of twenty six participants, only twenty two completed the experiment.

4. Results and Discussion

A total 308 (= 22 x 14) descriptions were recorded, 154 in each experimental condition (i.e. English Expressions Condition and Arabic Expressions Condition). Results were analyzed according to whether a participant provided under-specified, minimal or over-specified description. The results are shown in Table 2. Table 2 also shows the percentage of redundant attributes used in each condition. A one-way ANOVA test was used to test whether the apparent differences in the descriptions under the two experimental conditions are statistically significant or not.

The results in Table 2 (and also Figure 3) show that participants produced a large proportion of over-specified descriptions. Interestingly, they produced relatively more over-specified descriptions in English Expressions Condition (41.15%) than in the Arabic Expressions Condition (37.86%). ANOVA analysis, however, revealed that these apparent differences in the two main experimental conditions are not statistically significant: $F(1, 42) = 0.38$, $p = 0.54$. A further fine-grained analysis revealed that participants produced a large proportion of minimal descriptions in the Color-Minimal condition, whereas they produced relatively smaller number of minimal descriptions in both Size-Minimal and Orientation-Minimal conditions; these proportions are evident in both English Expressions Condition and Arabic Expressions Condition.

The present study revealed some interesting results. First, no significant differences were observed in English Expressions Condition and the Arabic Expressions Condition. These results can be interpreted as speakers' choice of referring expressions is not influenced by the language, i.e., speakers produce the same referring expressions in both their native and non-native languages. Second, speakers produced a large proportion of over-specified descriptions in both their native and non-native languages. These results are consistent with the existing psycholinguistic findings that speakers often produce over-specified referring expressions (Pechmann, 1989; Koolen, Gatt, Goudbeek, & Krahmer, 2011). Interestingly, the proportion of over-specified descriptions is very high in both Size-Minimal and Orientation-Minimal conditions, indicating that speakers use visually salient attributes more often than the less salient attributes (e.g. color is more salient than size and orientation). Again, this interpretation

[1] In the English Expressions Condition, the participants were explicitly asked to provide descriptions in English, whereas in the Arabic Expressions Condition, they were advised to provide descriptions in Arabic.

Condition	English Expressions Condition			Arabic Expressions Condition		
	Minimal	Under-specified	Over-specified	Minimal	Under-specified	Over-specified
CM	82.61	3.18	14.21	89.13	1.28	9.59
OM	17.85	5.09	**77.06**	21.82	4.61	**73.57**
SM	31.64	5.73	**62.63**	34.05	4.96	**60.99**
CSM	80.16	2.47	17.37	84.93	3.13	11.94
COM	78.02	2.91	19.07	74.82	4.01	21.17
SOM	29.81	6.38	63.81	36.92	4.37	58.71
COSM	49.35	16.72	33.93	57.63	13.29	29.08
Mean (Std Dev)	52.78 (27.34)	6.07 (4.93)	41.15 (26.13)	57.04 (26.76)	5.09 (3.82)	37.86 (26.05)

Table 2. Results of experimental study (the number of descriptions in %)

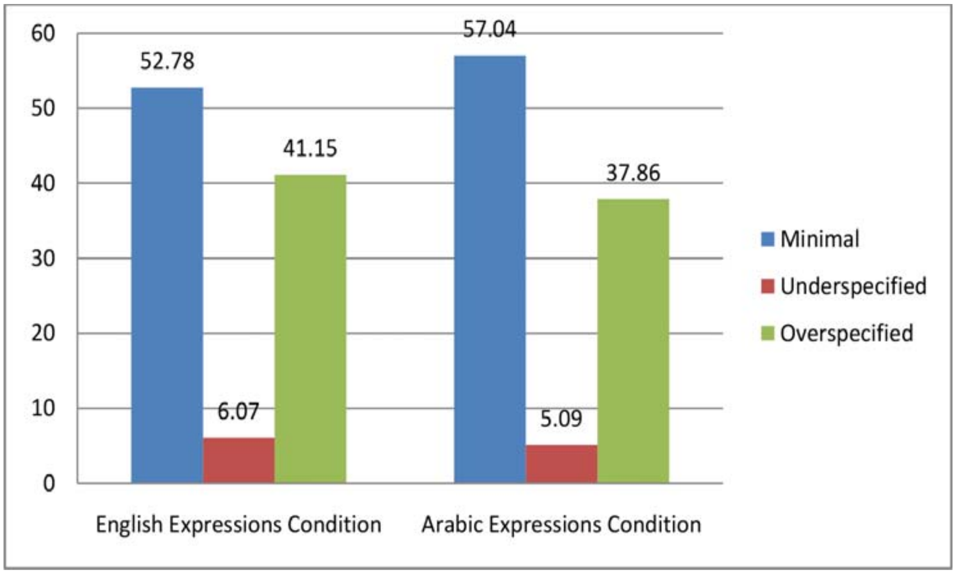


Figure 3. Percentage of minimal, under-specified and over-specified descriptions (averaged over seven sub-conditions)

of data is consistent with the existing research (Pechmann, 1989). Third, the present study also revealed that participants use relational attributes fairly often. Interestingly enough, the use of relational attributes was more often in complex descriptions, that is, relatively longer descriptions, in both (main) experimental conditions. These observations were supported by Fleiss' Kappa scores; participants showed reasonably high agreement on their choice of referring expressions in both experimental conditions ($k = 0.78$, $z = 5.61$, $p < 0.01$).

Finally, it is important to mention here that in the present study the participants produced the actual linguistic descriptions in words, rather than constructing description using semantic properties. In the data analysis, we derive the properties from linguistic descriptions to avoid any linguistic variation. For example, the two expressions '*the small red chair*', and 'the small chair with the red color' are both counted as having two properties[2].

[2] It is important to note that this study was focused on content determination task (of GRE) only and did not take into account different linguistic realizations of a property where the same property can be expressed using a single word in one language and two words in another.

5. Conclusion

The present study examined empirically how speakers produce referring expressions in the native (Arabic) and non-native (English) languages. Experimental trials were constructed with objects (furniture items) in visual domains and participants were asked to describe the (marked) target object by typing a description which can uniquely identify the target to its addressee, in two experimental conditions: Arabic Expressions Condition and English Expressions Condition. The data revealed that speakers produce similar referring expressions in both experimental conditions, meaning that the choice of referring expressions is not influenced by speakers' language. The results also showed that Arabic speakers produce over-specified referring expressions fairly often; these results corroborate earlier findings in the production of referring expressions.

Reference

[1] Belz, A., Kilgarriff, A. (2006). Shared-task Evaluations in HLT: Lessons for NLG. *Proceedings of the* 4th *International Conference on Natural Language* (p. 133-135). Sydney, Australia: Association for Computational Linguistics.

[2] Belz, A., Reiter, E. (2006). Comparing Automatic and Human Evaluation of NLG Systems. *Proceedings of the* 11th *Conference of the European Chapter of the Association for Computational Linguistics* (p. 3-7). Trento, Itly: Association for Computational Linguistics.

[3] Bortfeld, H., Brennan, S. E. (1997). Use and acquisition of idiomatic expressions in referring by native and non native speakers. *Discourse Processes*, 23 (2) 119-147.

[4] Dale, R. (1992). *Generating Referring Expressions: Building Descriptions in a Domain of Objects and Processes.* MIT Press.

[5] Dale, R., Reiter, E. (1995). Computational interpretations of the Gricean Maxims in the Generation of Referring Expressions. *Cognitive Sciene*, 18, 233-263.

[6] Farghaly, A., Shaalan, K. (2009). Arabic Natural Language Processing: Challenges and Solutions. *ACM Transactions on Asian Language Information Processing,* 8 (4), 1-22.

[7] Gardent, C. (2002). Generating Minimal Distinguishing Descriptions. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* (p. 96-103). Philadelphia, USA: Association for Computational Linguistics.

[8] Gatt, A., Belz, A. (2008). Attribute Selection for Referring Expressions Generation: New Algorithms and Evaluation Methods. *Proceedings of the 5th International Natural Language Generation Conference* (p. 50-58). Salt Fork, Ohio, USA: Association for Computational Linguistics.

[9] Gatt, A., Van der Sluis, I., Van Deemter, K. (2007). Evaluating algorithms for the generation of referring expressions using a balanced corpus. *Proceedings of the Eleventh European Workshop on Natural Language Generation* (p. 49-56). Saarbruecken, Germany: Association for Computational Linguistics.

[10] Grice, H. P. (1975). Logic and conversation. (P. Cole, & J. L. Morgan, Eds.) *Syntax and semantics, 3,* p. 41-58.

[11] Gupta, S., & Stent, A. (2005). Automatic Evaluation of Referring Expression Generation Using Corpora. *Proceedings of the Workshop on Using Corpora for Natural Language Generation,* (p. 1-6). Brighton, UK: Association for Computational Linguistics.

[12] Habash, N. (2010). *Introduction to Arabic Natural Language Processing (Synthesis Lectures on Human Language Technologies).* (G. Hirst, Ed.) Morgan & Claypool Publishers.

[13] Khan, I. H. (2015). Production of referring expressions in Arabic. *International Journal of Speech Technology*(*To Appear*).

[14] Khan, I. H., Deemter, K. V., Ritchie, G. (2012). Managing Ambiguity in Reference Generation: The Role of Surface Structure. *Topics in Cognitive Science* , 4 (2) 211-231.

[15] Koolen, R., Gatt, A., Goudbeek, M., Krahmer, E. (2011). Factors causing overspecification in definite descriptions. *Journal of Pragmatics* , 43, 3231-3250.

[16] Krahmer, E., Deemter, K. v. (2012). Computational generation of referring expressions: A survey. *Computational Linguistics,* 38 (1) -173-218.

[17] Krahmer, E., Van Deemter, K. (2012). Computational Generation of Referring Expressions: A Survey. *Computational Linguistics* 38 (1) 173-218.

18] Luk, K., Xiao, W., Cheung, H. (2012). Cultural effect on perspective taking in Chinese–English bilinguals. *Cognition*, 124 (3) 350-355.

[19] Mellish, C., Dale, R. (1998). Evaluation in the Generation of Referring Expressions. *Computer Speech and Language*, 12 (4) 349-373.

[20] Nwesri, A. A., Tahaghoghi, S. M., Scholer, F. (2005). Stemming Arabic conjunctions and prepositions. 12*th International Conference on String Processing and Information Retireval.* (p. 206–217). Buenos Aires, Argentina: Springer.

[21] Paris, C., Colineau, N., Wilkinson, R. (2006). Evaluations of NLG Systems: Common Corpus and Tasks or Common Dimensions and Metrics? *Proceedings of the 4th International Conference on Natural Language Generation* (p. 127-129). Sydney, Australia: Association for Computational Linguistics.

[22] Pechmann, T. (1989). Incremental speech production and referential overspecification. *Journal of Linguistics*, 27, 89-110.

[23] Reiter, E., Dale, R. (2000). *Building Natural Language Generation Systems.* Cambridge University Press.

[24] Shaalan, K., Monem, A. A., Rafea, A., Baraka, H. (2008). Generating Arabic text from interlingua. *Proceedings of the 2nd workshop on computational approaches to Arabic script-based languages*, (p. 137-144). Stanford, USA.

[25] Van Deemter, K., Van der Sluis, I., Gatt, A. (2006). Building a SemanticallyTransparent Corpus for the Generation of Referring Expressions. *Proceedings of the 4th International Conference on Natural Language Generation* (p. 130-132). Sydney, Australia: Association for Computational Linguistics.

[26] Viethen, J., Dale, R. (2006). Towards the Evaluation of Referring Expression Generation. *Proceedings of the 4th Australasian Language Technology Workshop* (p. 115-122). Sydney, Australia: Association for Computational Linguistics.

[27] Villringer, A., Chance, B. (1997). Non-invasive optical spectroscopy and imaging of human brain function. *Trends Neuroscience* 20 (10) 435-442.

[28] Wu, S., Keysar, B. (2007). The effect of culture on perspective taking. *Psychology Science,* 18 (7) 600-606.

Authors' Profiles

Imtiaz Hussain Khan is an assistant professor in Department of Computer Science at King Abdulaziz University, Jeddah, Kingdom of Saudi Arabia. He received his MS in Computer Science from the University of Essex UK in 2005 and PhD in Natural Language Generation from the University of Aberdeen UK in 2010. His areas of research are Natural Language Processing and Evolutionary Computation.

Muazzam Ahmed Siddiqui is an assistant professor at the Faculty of Computing and Information Technology, King Abdulaziz University. He received his BE in electrical engineering from NED University of Engineering and Technology, Pakistan, and MS in computer science and PhD in modeling and simulation from University of Central Florida. His research interests include text mining, information extraction, data mining and machine learning.