# Compiling and Analysing a Corpus of Transcribed Spoken Gulf Pidgin Arabic Based on Length of Stay in the Gulf

**Najah Albaqawi, Michael Oakes**
**Research Group in Computational Linguistics, University of Wolverhampton**

**N.Albaqawi@wlv.ac.uk, Michael.Oakes@wlv.ac.uk**

## Abstract

The focus of this paper is on how to compile and analyse a transcribed spoken Gulf Pidgin Arabic (GPA) corpus with a specific focus on the influence of length of stay in the Gulf on foreign expat female speakers of GPA. GPA is a simplified contact variety of the Arabic language used in the Gulf states for communication between native Arabic speakers and foreign workers and among the workers themselves. This study provides a quantitative analysis of language variation in GPA based on five morpho-syntactic features that are related to the length of stay in the Gulf: definiteness and indefiniteness, coordination, copular verbs, pronouns, and agreement in the verb phrase and in the noun and adjective phrase. Digital recorders and planned interviews were used for collecting accurate naturalistic data. Through a comparative corpus-based analysis of approximately 72,000 words spoken by GPA female participants, evidence from this corpus data indicates that length of stay in the Gulf seems to have a little effect on informants 'choice between GPA linguistic variants. Newcomers and long-term resident GPA female speakers in the Gulf shift towards Gulf Arabic (GA), the lexifier language, in only two features: definiteness and use of conjunction markers.

## 1 Introduction

The field of corpus linguistics has gained huge popularity in recent years. It has become one of the most wide-spread methods of linguistic investigation not only among the experts, but also many researchers who would not consider themselves to be corpus linguists have begun to apply methods of corpus linguistics to their linguistic statements and assumptions. Joseph (2004:382) states: 'we seem to be witnessing as well a shift in the way some linguists find and utilize data – many papers now use corpora as their primary data, and many use internet data'.

GPA has received relatively little attention in the literature apart from a few descriptive works such as Albakrawi (2013); Albaqawi (2016); Alghamdi (2014); Almoaily (2008, 2012); Alshammari (2010); Al-Azraqi (2010); Al-Zubeiry (2015); Avram (2014, 2015); Gomaa (2007); Hobrom (1996); Næss (2008); Salem (2013); Smart (1990); and Wiswal (2002). In this paper, we are particularly interested in what a corpus of GPA spoken data, ideally in the form of recordings aligned with an orthographic Arabic transcription, might tell us about the use of language. Length of stay in the Gulf and GPA language variation will be examined in this study from a sociolinguistic point of view, since the study of linguistic variation in contact languages can make a valuable contribution to the field of sociolinguistic variation and change. Traditionally, researchers in sociolinguistics were not interested in using corpora in their

investigations (Baker, 2010:1) until 1996 when McEnery and Wilson suggested a first possible relation with corpora. They showed the value of supplementing the qualitative analysis of language with quantitative data (McEnery & Wilson, 2003). In 2006, McEnery et al. also indicated that along with the speed of information processing, there are specialised software which can classify and select words to look at their frequencies between major classes, for example, male and female usage.

This paper will start with a brief definition of pidgin and the situation of pidgin Arabic in Saudi Arabia. This will be followed by a discussion of compiling and analysing a spoken variety of Arabic, GPA, and the difficulties associated with that. Then we will analyse the impact of the number of years of residency by Asian female workers located in the Gulf as a potential factor conditioning language variation in GPA. The final section will provide some conclusions and suggestions for future studies on GPA.

## 2 Definition of Pidgin

In this section we will try to give a simple definition of pidgin and creole, regardless of the diverging views in defining these two contact language types.

**Pidgin**: Pidgin is defined by Velupillai (2015) as "a language that emerges when groups of people are in close and repeated contact, and need to communicate with each other but have no language in common". McWhorter (2001, 2004) also defined pidgins as the languages that result from maximal contact and adult language learning, and their speakers use them as "transitory tools" for passing exchanges. If people use this simplified version of language, pidgin, as an everyday language, a pidgin can become a real language, a creole.

Usually a pidgin language is a blend of the vocabulary of one major language (i.e. language of the dominant group which is referred to as the 'lexifier' or 'superstrate', in our case GA) with the grammar of one or more other languages (i.e. languages which are spoken by groups with lesser social status to the lexifier speakers which are referred to as the 'substrate languages'. In our case they are from the following six different language groups: Tagalog, Punjabi, Sinhala, Malayalam, Sunda, and Bengali).

## 3 The Situation of Pidgin Arabic in the Gulf States

The situation in which GPA was developed is a textbook case of the situations that create a pidginised variety. Sakoda and Siegel (2003:1) write:

> Nowadays, the term "pidgin" has a different meaning in the field of linguistics. It refers to a new language that develops in a situation where speakers of different languages need to communicate but don't share a common language.

According to their definition, the situation in the Gulf States is ideal for the birth of a new contact language as the Arab Gulf States are located in the centre of the Old World[1]. Following the October 1973 "oil boom," the Arab Gulf States (GCC)[2] have experienced radical social, political and demographic changes in a very short time. This has led to an extremely rapid increase in the demand for foreign labour. The number of foreign labourers in the Gulf countries, especially the Kingdom of Saudi Arabia, rapidly increased, amounting to almost 4.4 million in 1985, a more than three-fold increase within a single decade. Also, the kingdom is the biggest economy in the Arab world, endowed with the world's second largest proven oil reserves. This makes Saudi Arabia a major hub for population movements (De Bel-Air, 2014). Saudi Arabia, as stated by Avram (2013b), has a multilingual setting as do all Gulf countries; Gulf Arabic (GA) is a form of Colloquial Arabic language spoken by the indigenous people of the Gulf Region. Migrant workers, who come from various linguistic backgrounds and usually do not speak Arabic, come into contact with GA speakers as well as speakers of other Arabic dialects, and there is an urgent need for communication between the two groups, "Arabic-speaking locals and expats on one hand and non-Arabic speaking expats on the other" (Almoaily, 2012, p. 1). Thus, a simplified form of Arabic has developed as a result of this

1 Some geographers use the term Old World to refer to Asia, Africa, and Europe (see Mundy, Butchart, and Ledger 1992).

2 Gulf Cooperation Council, which includes: Saudi Arabia, Kuwait, Bahrain, Qatar, United Arab Emirates, and Oman

contact which is known as 'Gulf Pidgin Arabic' (henceforth GPA). GPA is a reduced system of language that is used for communication between foreign workers and the native speakers of Arabic. Indeed, GPA and GA are two distinct forms of language, with lexical, phonological, syntactic, and morphological differences. At the level of phonology, Albaqawi (2016) who conducted a study that investigated the phonetic variation within GPA spoken by Asian migrant workers in the Gulf countries concluded that the basic GPA phonetic inventory is either reduced or simplified and differences in phonology are limited in GPA varieties. However, one vital question should be asked: Does a local speaker use GPA when he/she is speaking to a GPA speaker? To answer this question, Almoaily (2008) asked 77 Saudi respondents if they 'don't mind using GPA with speakers who are not fluent in GA'. Half of the Saudi respondents agreed to use GPA with non-Arabic speaking foreigners (especially among the younger generation of locals) and the other half either disagreed or strongly disagreed with this statement. He also claimed that locals' use of GPA when speaking to GPA speakers was higher than 50%.

However, this issue is still a controversial and it depends on the quantity and quality of input which GPA speakers are daily exposed to the superstrate language, GA.

## 4 Corpus and Methodology

When spoken language is addressed, traditionally, a corpus linguistics work starts with deriving an orthographic transcription from a recording of large stretches of speech. The main aim of building a spoken language corpus is to acquire large amounts of data reflecting the natural use of language. Thus, all the data in this research are collected via interviews with female informants who do not speak Gulf Arabic as their first language. In order to examine such data, a quantitative variationist analysis of GPA variability was used. In this study we attempt to provide a quantitative analysis which aims to discover the potential effect of the number of years spent in the Gulf on variability in GPA morpho-syntax.

### 4.1 The Corpus

The corpus consists of the speech of informants participating in interviews which were conducted in Saudi Arabia[3]. To test the influence of the length of stay on the GPA female speakers' language variation, face-to-face recorded interviews were conducted between the subjects and the interviewer (the first author) by using a high-quality digital voice recorder[4] and ranged from 16 to 27 minutes. The data-base consists of interviews with 72 GPA speaking female informants from six linguistic backgrounds (Malayalam, Punjabi, Bengali, Tagalog, Sunda, and Sinhala) as these substrate languages are the largest number of speakers in Saudi Arabia based on the results of the Population Census from De Bel-Air (2018). Half of the data was produced by informants who have spent five years or less in the Gulf while the other half had spent ten years or more in the Gulf at the time the researcher interviewed them. This paper seeks to investigate whether the long-term residents have actually shifted towards GA or not. The structural patterns of GA that were collected from the newcomers of each language group were compared with those of long-term residents (e.g. newly-settled Tagalog speakers vs. Tagalog speakers who spent more than a decade in the Gulf). In other words, we compared their proportional use of GA tokens of the morpho-syntctic phenomena investigated in this study: Arabic definiteness markers (i.e. the prefix *al-*), Arabic conjunction markers (these markers are mostly the free morphemes *wa* 'and', *laakin* 'but', and *aw* 'or'), object or possessive pronoun (i.e. subject pronouns in GA are the free morphemes , 1SG *ana* whereas object and possessive pronouns are always bound morphemes, 1SG -i.), copula (i.e. the GA copula *fi* is used overtly only in the past and future whereas it is covert in the present tense), and agreement in the verb phrase and the noun phrase with that produced by their newly settled counterparts. We opt to examine these morpho-syntactic features as we believe that they are adequate to test the proposed typological features (reduced inflection; reduction of agreement markers in verb and noun and adjective agreement, and reduced inventory of function words; copulas, definite and indefinite

3 All the interviews were conducted in the Saudi Central Province where Najdi Arabic – a sub-dialect of GA – is spoken

4 Olympus vn-7800pc

articles, and pronouns) that might be found in all pidgin and creole languages worldwide (irrespective of their input languages) see Almoaily (2013); Bakker (1995, 2003); Roberts and Bresnan (2008); Bakker, Daval-Markussen, Muysken, and Parkvall (2011); Sebba (1997); and Siegel (2004).

Counting the lexical features has been excluded for two reasons: First, the purpose of this paper is to examine the structure of GPA rather than its lexicon. Second, vocabulary studies are more related to developed languages. For example, Malmasi, et al. (2016) identified a set of four regional Arabic dialects (Egyptian, Gulf, Levantine, North African) and Modern Standard Arabic (MSA) which are all native languages unlike the GPA which has no natives.

General principles for the quantification of variability above the level of phonology are still a matter of debate (Macaulay, 2002). A number of researchers have come up with several approaches for the quantification of tokens. Some quantify them by the number of words as was done by Precht (2008) and Cheshire, Kerswill and Williams (2005). On the other hand, some researchers prefer to quantify them per minute or hour of speech in a sociolinguistic interview, as was done by Rickford and McNair-Knox (1994). In our case, we preferred to calculate the tokens per number of words as Almoaily (2013) suggested, irrespective of the length of the turn or the number of words produced in a minute of speech. Our reason was that the informants of our study have been exposed to GPA over a period ranging from eight months to twenty-five years, and newly arrived speakers are expected to speak slower than those who have spent more than ten years in the Gulf.

### 4.2 Transcribing the Interviews

The first author transcribed the interviews herself. It took nearly three hours to transcribe and revise only ten minutes of speech. She used Express Scribe Transcription Playback Software[5]and transcribed that segment of the interview manually[6] (since the transcriptions of the whole interviews are in Standard Arabic script).

### 4.3 Annotation of Counting the Tokens

In the corpus each variant of a variable is labelled with a unique code[7]. The example below shows a code and its meaning:

**Code:** (+روابط)/ ( - روابط)

**Meaning:** The conjunction marker is present (CONJ +)/ The conjunction marker is dropped (CONJ -).

In order to count and retrieve the tokens from the transcribed interviews, we used the AntConc software[8]. AntConc is one of the best tools for analysing a corpus. Froehlich (2015) refers to AntConc as a very good toolkit for finding patterns in language which would be difficult to identify just by reading the text. Figure 1 below shows how a transcribed interview appears with the AntConc program:

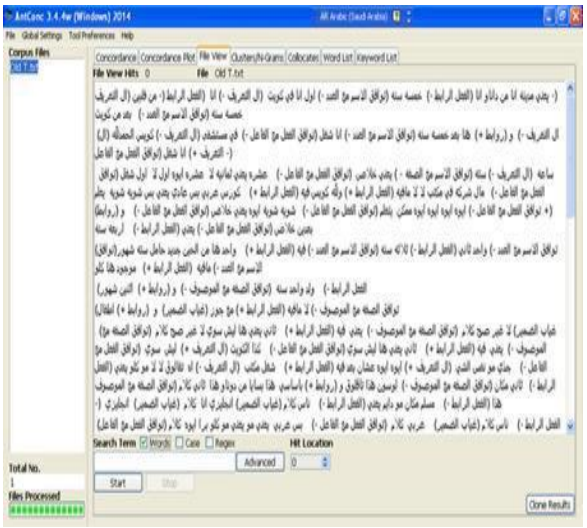

Figure 1: Old Tagalog corpus

We tried to find the frequency of occurrence for every linguistic feature chosen in the study (e.g. definiteness). The Concordance view showed whenever the chosen linguistic feature (e.g. definiteness) appeared in our corpus (e.g. Tagalog corpus newcomers) and some context of it (such as a window of x words). We did the same for all the corpus files that we had. We then calculated the percentage of tokens produced in every variant.

5 Professional audio player software designed to assist the transcription of audio recordings (Free, cross- platform).

6 Arabic transcription/dictation software tools for non- standard Arabic varieties or Arabic-based contact languages are inaccurate and thus were avoided in transcribing the data for the current study.

7 We used Arabic characters as a unique code.

8 A freeware corpus analysis toolkit for concordancing and text analysis.

To compare the use of the given variant by members of a sub- group with that of other sub-groups (e.g. newly-settled female Tagalog speakers vs. long-term female Tagalog residents), the researchers used statistical analysis to look at the differences between two corpora.

This was used to establish the significance of the effect of the years of residency in the Gulf on variation in GPA.

## 5 Issues in Compiling and Analysing a Spoken Corpus

This study depends on using a suitable corpus and since GPA is only a spoken variety of the Arabic language, there was no such corpus previously available in electronic form. In addition, the GPA corpus is different from Arabic Learner Corpus as in Alfaifi and Atwell (2013). In the Arabic Learner Corpus, the students are all trying to learn Standard Arabic, while in the GPA corpus, the target language, whether GA or GPA, is a matter of debate. Thus, we had to design and build our own corpus. A number of difficulties and challenges were associated with implementing such a corpus. These included, size, balance (choosing informants), representativeness, and annotation. We will discuss the question of annotation here.

**Annotation:** Annotating a corpus written in Arabic script presents challenges. Many dialects are written in different scripts, have no conventions for spelling and no large body of literature. In our case we have "code-mixed" text, interspersed with other languages (Arabic and English). As a first attempt, we labelled each variant of a variable with a unique Roman code (e.g. CONJ+ if the conjunction is used and CONJ- if the conjunction is dropped). This attempt failed because the AntConc software was not able to detect accurately the linguistic code switching within Arabic script text as Arabic script starts from right-to-left where English Roman script starts from left-to-right. To overcome these systematic changes in writing direction, we decided to retranscribe all our corpus files in a unified spelling system by using Arabic code instead of Roman code for the annotation (e.g. الفعل الرابط+ the copula is used and الفعل الرابط- the copula is dropped). This revised annotation works very well and it has been adopted in the main corpus.

## 6 Results and Discussion

### 6.1 New versus Old participants

Each language group was split into two groups based on their length of stay in Saudi Arabia, or any other GA speaking country (5 years or less—referred to as "New" or 10 years or more—referred to as "Old"). Chi-squared tests were run to establish the significance of the effect of years of residency in the Gulf on variation in GPA.

Results of simple concordance comparisons of the new and old participants are presented in Table 1. Comparing the percentages of occurrence of each variable gives us the opportunity to contrast the proportion of use of the GA variants as opposed to the proportion of use of the GPA variants.

These were variants in definiteness, conjunction markers, the copula, object and possessive pronouns and agreement in the VP and in the NP and in the ADJP presented in table 1 and Figure 2:

| GA Linguistics Feature | GPA Informants | |
|---|---|---|
| | new | old |
| **Definiteness** | 10.7% | 33.7% |
| **Conjunction Markers** | 12.9% | 41.5% |
| **Copula Fi** | 54.7% | 59.2% |
| **Object and Possessive Pronouns** | 18% | 22.9% |
| **Subject-Verb Agreement** | 5.2% | 8.4% |
| **Nominal Agreement** | 19.3% | 26.6.% |

Table 1: Concordance and percentage used in the corpus of the new and old participants

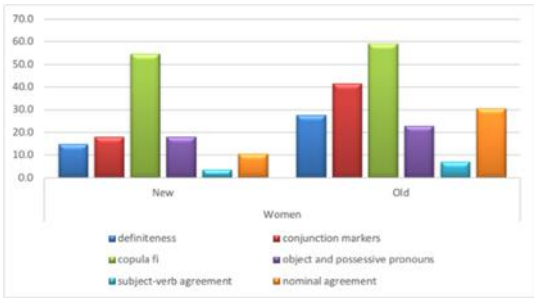


Figure 2: Data Comparison between New and Old GPA speakers

### 6.2 Variation in Definiteness

We noticed a possible link between the length of stay in GA speaking countries and the use of the definiteness marker *al-*. This shift towards GA was seen in all six language groups. The newly-arrived GPA speakers produced the definiteness marker in 10.7% of the cases, whereas the old members produced them in 33.7%. The chi-squared test revealed that the difference between the new informants and those who stayed longer in the Gulf in producing definiteness markers is significant at a p-value of 0.002.This noticeable shift towards using the GA definiteness marker among the long-term residents could potentially be a result of the fact that definiteness in GA is one of the morpho-syntactic features that are easiest to learn as it only involves adding the prefix *al* – or one of its allophones – to the target noun.

### 6.3 Variation in the Use of Conjunction Markers

The data reveals a major shift towards GA in the use of conjunction markers. This effect was seen in all six language groups. The newly-arrived GPA speakers produced conjunction markers in 12.9% of the cases, whereas the old informants produced them in 41.5%. The chi-square test reveals that the difference between the new informants and those who stayed longer in the Gulf in producing conjunction markers is significant at a p-value of 0.001. This significant difference could be due to the fact that learning GA conjunction markers is not hard. GA conjunction markers are free morphemes (e.g. *wa* 'and', and *aw* 'or'). This result is in parallel with Almoaily (2013)'s study of male GPA speakers.

### 6.4 Variation in the Use of the Copula

In GA, there is no copula in the present tense. Thus, the focus here is on the use of the copula *fi* in the present tense in GPA. If long-term residents are found to drop the copula more than the newcomers, this might be an indication of a shift towards GA. The data reveals that the relation between the years of stay and the shift towards GA seems to be slightly negative at a p-value = 0.35. Overall, there is no significant shift towards Gulf Arabic in the data of speakers participating in this study regarding the use of a copula, as new speakers dropped it on average 54.7% of the time and old speakers dropped it in 59.2% of the time.

### 6.5 Variation in the Use of the Object and Possessive Pronouns

GA personal pronouns inflect for number, person, and gender. In GPA, there are four variants for object and possessive pronouns: GA bound pronoun which agrees with the noun in person, number, and gender, GA bound pronoun which does not agree with the noun, free pronoun, and dropping the object or possessive pronoun. On average, newly-settled informants in all six language groups produced bound object and possessive pronouns in 24.2% of the cases, while the long-term residents produced them in 49.2% of the cases. The difference is significant at a p-value of 0.0001. Note that the newcomers produced tokens of pronouns in free forms 71% of the time and the old group counterparts produced them 75.5% of the time. In fact, this high rate of free object and possessive pronouns indicates that the overall shift is clearly not towards GA (bound pronouns) but GPA (free pronouns). Since this feature (free pronoun) is found in the informants' L1s, it could probably have some influence on GPA speakers and lead them to learn it at the first stage as reported in Almoaily (2012).

### 6.6 Variation in Subject-verb Agreement

In Gulf Arabic, the verb inflects for gender, number, tense, and person (Feghali, 2004). The data shows that there is a positive development related to the informant's length-of-stay in the use of verbs: members of the new group tend to drop verbs more frequently (35.6%) than their old group counterparts (8.4%). The rate of dropping the verb is significantly higher in the data of new informants at a p-value of 0.0002. However, it

seems that there is no development in acquiring agreement in the GA verbal system. Overall, the data revealed that all of the informants rarely produced the form of the verb that is used in GA (i.e. fully inflected verb forms that agree with the subject in number, gender, and person). Compare the overall percentage of new-comers who produced a fully inflected GA verb only in 5.2% of the total number of tokens, with that of old informants who produced it in 8.4%. Yet, the difference is not significant (p-value= 0.22). Note, the overall shift is clearly not towards GA, as the use of forms of verb markers which do not agree with the noun in gender, number, and person are predominant in the data of both new and old speakers.

### 6.7 Agreement in the NP and in the ADJP

In GA, the adjectives agree with the head noun in gender, number, and definiteness (Feghali 2004, Smart 1990, Almoaily 2008). The data show that there is a little positive improvement in the occurrence of nominal agreement by the participants who stay long in the Gulf as compared to their new counterparts. We have noticed that long-term residents produce a few more tokens of noun-adjective agreement in number and gender than their new counterparts. The new informants produced agreement tokens in 19.3% of the total number of cases, while their long-term counterparts produced it in 26.6 % of the total number of cases. Although the difference is not statistically significant (p-value = 0.08), and even though there is obviously a vast amount of variation within the groups, there seems to be a trend towards the acquisition of GA norms.

## 7 Conclusion

The main aim of this study was to examine how to build and analyse a spoken corpus for a sociolinguistic investigation. Indeed, we expected to face difficulties when deciding on size, balance, representativeness, and annotation of our spoken corpus. Compiling and analysing the corpus for this investigation were the most demanding task and time-consuming task (see section 5). First, choosing GPA speakers who meet certain criteria and convincing them to participate in the interview was not an easy task. Many simply refused to be interviewed and many others were too busy to take part in this study. Also, transcribing the interviews and choosing the appropriate transcription protocol for Arabic script presented greater challenges. The strategy we employed to overcome, or lessen the impact of these problems was by transcribing all our corpus files in a unified spelling system by using Arabic code instead of Roman code. It was very fruitful technique (see section 4).

The study also was aimed to investigate language variation in GPA resulting from the speakers' length of stay in the Gulf. The analysis suggests that this factor seems to have a little effect on informants 'choice between GPA linguistic variants. We expected long-residence speakers to produce more GA tokens than the newly-settled GPA speakers. They have made a significant shift to GA after spending ten years in the Gulf in two linguistic features only: definiteness and conjunction (p-value=0.002, p-value=0.001) respectively. There are some factors which we believe could have had an effect on the informants' choice between the selected features' variants. This could potentially be a result of the fact that most of the informants are female maids living with a local family who mostly use GA when communicating with them, which could play a major role on the process of acquiring a language. This in turn leads them to rapidly learn the language of the host community and effortlessly adopt the system of Gulf Arabic (the target language). Another effect on the informants' choice between the selected feature variants is that it may depend more on the amount of GA input that GPA speakers receive during their stay in the Gulf (rather than the language of origin), different learning abilities to learn a language, and motivation.

We conclude this study with a set of recommendations for future research on this pidgin language. First, we suggest considering the role of input in pidgin formation. Second, we will conduct a comparison study to investigate male and female GPA production and effect of the language of the origin. Finally, it would be fruitful to conduct and computationally analyse more data-based studies of Arabic-based pidgins which are less known in the literature of non-Indo-European pidgin languages.

## References


Abdullah Alfaifi and Eric S. Atwell. 2013. Arabic learner corpus v1: A new resource for arabic language research. Leeds.

Abdul-Qadir Wiswall. 2002. Gulf pidgin: An expanded analysis, *unpublished pro-seminar paper,* Ohio State University.

Andrei A. Avram. 2013b. On the periphery of the periphery: Gulf Pidgin Arabic. *Proceedings of 10th Conference of Association* Internationale *de Dialectologie Arabe.* Qatar University, Doha.

Andrei A. Avram. 2015. On the developmental stage of Gulf Pidgin Arabic. *In Arabic varieties: Far and wide. Proceedings of the 11th International Conference of AIDA* (p.87-98). Bucharest, Romania.

Andrei A. Avram. A. 2014. Immigrant Workers and Language Formation: Gulf Pidgin Arabic. *Lengua y Migración,* 6 (2). 7- 40.

Ashraf A. Salem. 2013. Linguistic Features of Pidgin Arabic in Kuwaiti. *English Language Teaching*, 6 (5). 105-110.

Brian Joseph. 2004. On change in Language and change in language. *Language*, 80(3), 381-383.

David. Evans. 2007. Corpus building and investigation for the Humanities. *University of Nottingham* http://www. corpus. bham. ac. uk/corpus-building. shtml.

Emad A. Alghamdi. 2014. Gulf Pidgin Arabic: A Descriptive and Statistical Analysis of Stability. *International Journal of Linguistics* 6, no. 6, p.110.

Françoise De Bel-Air. 2014. Demography, Migration and Labour Market in Saudi Arabia. *Gulf Research Center Knowledge for All.* Retrieved from http://gulfmigration.eu/media/pubs/exno/GLMM_EN_2014_01.pdf.

Françoise De Bel-Air. 2018. Demography, migration and labour market in Saudi Arabia.

Gisle Andersen. 2010. How to use corpus linguistics in sociolinguistics, in O'Keeffe A. & M. McCarthy (ads.), *The* Routledge handbook of corpus linguistics, 547-62. 1" ed. London: Routledge.

Gomaa, Y. 2007. Arabic Pidginization: The Case of Pidgin in Saudi Arabic. *Journal of the Faculty of Arts, 19*, 85-120, Assiut University, Egypt.

Habaka J. Feghali. 2004. Gulf Arabic: the dialects of Riyadh and eastern Saudi Arabia: grammar, dialogues, and lexicon. *Springfield, VA, Dunwoody Press.*

Hameed Y. Al-Zubeiry. 2015. Linguistic Analysis of Saudi Pidginized Arabic as Produced by Asian Foreign Expatriates. *International Journal of Applied Linguistics & English Literature*, 2(4), 47-53.

Heather Froehlich. 2015. Corpus analysis with AntConc. *Programming Historian*.

Hussien Albakrawi. 2012. The linguistic effect of foreign Asian workers on the Arabic Pidgin in Saudi Arabia. *language* 2, no. 9.

Jack R. Smart. 1990. Pidginization in Gulf Arabic: A first report. *Anthropological Linguistics,* 32, 83-118.

Jeff Siegel. 2004. Morphological Simplicity in Pidgins and Creoles, *Journal of Pidgin and Creole Languages,* 19: 139–62.

Jenny Cheshire, Paull Kerswill, and Ann Williams. 2005. Phonology, Grammar and Discourse in Dialect Convergence. In: Peter Auer, France Hinskens, and Paull Kerswill (eds). *Cambridge: Cambridge University Press.*

John H. McWhorter. 2004. The story of human language (Course Guidebook). Chantilly, VA: *The Teaching Company.*

John R. Rickford and Faye McNair-Knox. 1994. Addressee and Topic Influenced Style Shift: a quantitative sociolinguistic study. In: Douglas Biber and Edward Finegan, (eds). Oxford: *Oxford University Press.*

Kent Sakoda and Jeff Siegel. 2003. Pidgin grammar: An introduction to the creole English of Hawai'i. Honolulu, *Hawaii: Bess Press*.

Kristen Precht. 2008. Sex Similarities and Differences in Stance in Informal American Conversation, *Journal of Sociolinguistics*, 12 (1) 89-111.

Mark Sebba. 1997. Contact Languages: pidgins and creoles. *London, Macmillan*.

Mohammad Almoaily. 2008. A data-based description of Urdu Pidgin Arabic. Unpublished MA dissertation, *Newcastle University*.

Mohammad AlMoaily. 2012. Language Variation in Gulf Pidgin Arabic (Doctoral dissertation). *Newcastle University,* the United Kingdom.

Munira Al-Azraqi. 2011. Pidginisation in the eastern region of Saudi Arabia: Media presentation. In *Arabic and the Media*, pp. 159-173. Brill.

Najah S. Albaqawi. 2016. Unity and Diversity within Pidginized Arabic as Produced by Asian Migrant Workers in the Arabian Gulf.

Natalie Schilling-Estes. 2007. Sociolinguistic Fieldwork. In: Robert Bayley and Ceil Lucas, (eds.) Sociolinguistic Variation Theories, Methods, and Applications: *Cambridge: Cambridge University Press.*

Peter Bakker, Aymeric Daval-Markussen, Mikael Parkvall, and Ingo Plag. 2011. Creoles are Typologically Distinct from Noncreoles, *Journal of Pidgin and Creole Languages*, 26 (1) pp. 5-42.

Peter Bakker. 1995. *Pidgins*. In Jacques Arends, Pieter Muysken , and Norval Smith (eds) Pidgins and creoles an introduction.1995. *Amsterdam; Philadelphia: J. Benjamins.*

Peter Bakker. 2003. Pidgin inflectional morphology and its implications for creole morphology, *Yearbook of Morphology*, Part 1, 3-33.

R Core Team. 2013. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Robert Bayley and Dennis R. Preston. 1996. Second Language Acquisition and Linguistic Variation. *Amsterdam: Benjamins*.

Ronald Macaulay. 2002. Discourse Variation. In: Jack K. Chambers, Peter Trudgill, and Natalie Schilling-Estes (eds.) The handbook of language Variation and change. *New York: Blackwell Publishing Co.*

Sean Wallis. 2014. What might a corpus of parsed spoken data tell us about language. In *Complex Visibles Out There. Proceedings of the Olomouc Linguistics Colloquium* (pp. 641-662).

Sarah J. Roberts and Joan Bresnan. 2008. Retained Inflectional Morphology in Pidgins: A typological study, *Linguistic Typology* 12: 269–302.

Shervin Malmasi, Marcus Zampieri, Nikola Ljubešić, Preslav Nakov, Ahmed Ali, and Jörg Tiedemann. 2016. Discriminating between similar languages and arabic dialect identification: A report on the third dsl shared task. In *Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial3)* (pp. 1-14).

Tony McEnery, Richard Xiao, and Yukio Tono. 2006. Corpus- based language studies: An advanced resource book. Taylor & Francis.

T-Wilson McEnery and A Wilson. 1996. Corpus Linguistics Edinburgh: Edinburgh University Press.

T-Wilson McEnery and A Wilson. 2003. Corpus linguistics. The Oxford handbook of computational linguistics, 448-463.

Viveka Velupillai. 2015. Pidgins, Creoles and Mixed Languages. An Introduction. Amsterdam and Philadelphia: *John Benjamins*.

Wafi Alshammari. 2010. An Investigation into Morpho-syntactic Simplification in the Structure of Arabic Based Pidgin in Saudi Arabia (Master's theses). *Mu'tah University, Jordan*.

William Labov. 1972. Language in the Inner City; studies in the Black English vernacular. Philadelphia: *University of Pennsylvania Press*.