Who, When and Why: The 3 Ws of Code-Switching

Chapter · July 2020

DOI: 10.1007/978-3-030-51999-5_7

6 authors, including:

Alia el Bolock
The German University in Cairo & Ulm University

Injy Khairy Hamed
Universität Stuttgart

Yomna Abdelrahman
Universität Stuttgart

Thang Vu
Universität Stuttgart

Who, When and Why: The 3 Ws of Code-switching

Alia El Bolock[1,2], Injy Hamed[1,3], Yomna Abdelrahman[5], Ngoc Thang Vu[3], Cornelia Herbert[1], and Slim Abdennadher[1]

[1] German University in Cairo, Cairo, Egypt
[2] Ulm Univeristy, Ulm, Germany
[3] Stuttgart University, Stuttgart, Germany
[4] Bundeswehr University, Munich, Germany
{alia.elbolock,slim.adbennadher}@guc.edu.eg
{cornelia.herbert@uni-ulm.de}
{injy.hamed,ngoc-thang.vu}@ims.uni-stuttgart.de
{yomna.abdelrahman@unibw.de}

Abstract. With the rise of globalization, the use of mixed languages in daily conversations, referred to as "code-switching" (CS) has become a common linguistic phenomenon among bilingual/multilingual communities. It has become common for people to alternate between distinct languages or "codes" in daily conversations. This has placed a high demand on Natural Language Processing (NLP) applications to be able to deal with such mixed-language input. Researchers have lately achieved advancements in multilingual NLP applications, however, few work has been done to adapt these applications to users' CS behaviour. In this work, we take the first steps towards this goal by investigating the CS behavior and its correlation with the users' profiles in our case study on Egyptian Arabic-English code-switching. Although these factors have been investigated by linguists, the findings have been mostly made through theoretical studies. We provide empirical evidence based on a user study with 50 participants showing initial correlations between user traits and the CS frequency, which can be used to predict users' CS. Our findings imply that in the scope of our study, people (who) code-switch in specific discourse domains more than others (when) and depending on their background and social factors (why). The study also shows that to be able to properly investigate who code-switches more data needs to be collected and further analysis is required.

Keywords: Character Computing, Code-switching, Code-mixing, Personality, Arabic-English

1 Introduction

Verbal communication is a basic and main means of communication between individuals. The most known obstacle is the language barrier. Over the past years a phenomenon called *Code-switching* (CS) became common with the rise of mixed languages [24]. CS occurs when a bilingual/multilingual person alternates between two or more languages, intentionally or unintentionally, in the context of one conversation. It became an increasingly topical and important field of research in several fields such as, Natural Language Processing (NLP), Cognitive Science, and Socio-Linguistics. CS is an obstacle that could hinder communication between two speakers with the same mother-tongue. This is especially the case when interacting with technology, that attempts to automatically detect speech.

In this work, we aim to investigate whether it could be possible to predict user's CS behavior based on user knowledge. This information can then be used for user-profiling to build user-adaptive NLP applications that can process information more accurately. By knowing the factors that affect the user's CS behavior, we can predict the user's CS frequency based on given user information. This can then be integrated into a system to build a user-adaptive model that would perform better.

A lot of research investigates CS and its reasons. It has been often attributed to various psychological, social, and external factors [17, 28, 7, 26, 29, 3, 25, 18, 28, 14]. In this paper, we aim to advance the state of the art by investigating who is more prone to code-switching, as well as why and when people code-switch. Like any other behavior, the CS behavior of the same person could vary in different situations and depending on their different traits and states (i.e., their character) [11, 9, 13]. This paper tackles the 3 "W"s of CS: who, why and when, by investigating the correlation between 1) the personality of an individual, 2) the domain and topic of discourse, 3) the individual's background and history, and the CS behavior, respectively. We conducted a user study where we collected speech data as well as background and personality information from 50 participants to generate the needed dataset. We labeled the dataset to get insights into the correlations between the frequency of CS and the aforementioned attributes; who, when and why. This paper contributes the investigation of factors that affect individuals' CS behavior and the collected dataset of Arabic-English CS speech.

The remainder of the paper is structured as follows. We first present a summary of existing related work, and define the needed concepts in Section 2. Next, in Section 3, we describe our study design and the dataset generation. We then discuss the analysis and results in Section 4. Finally, we make conclusions and discuss the future work in Section 5.

2 Related Work

Many scientific fields have been interested in understanding the processes behind code switching, from its measurement to its modeling. Given that CS is a user-dependent behaviour, where each user code-switches differently, it would be useful to incorporate the different code switching frequencies in the NLP applications to make them user-adaptive. A pre-condition for this is the ability to understand the code switching behavior of users. In this section, we discuss the types of code switching and as well as the work done by socio- and psycholinguistics to determine the factors that affect the individual's CS behavior.

2.1 Code-switching (CS)

Code-switching (CS) is the act of using more than one language in a conversation. This phenomenon has become popular, especially among urban youth. It evolved as a result of several factors, including globalization, immigration, colonization, the rise of education levels, and international business and communication. CS can be seen in several bilingual/multilingual societies, such as: Cantonese-English in Hong Kong [19], Mandarin-Taiwanese in Taiwan [6], Mandarin-English in Singapore and Malaysia [20], Spanish-English in Hispanic communities in the United States [1], Turkish-German in Germany [5], Italian-French and German-Italian in Switzerland [2], Arabic-English in Egypt [16] and Arabic-French in Tunisia [27], Algeria [8] and Morocco [4].

In [23], the author categorized CS into the following three types:

- Inter-sentential switching: happens at clausal or sentential level where each clause or sentence is in one language or another. For example, "Du musst immer optimistisch bleiben. Every cloud has a silver lining." (You must stay optimistic. Every cloud has a silver lining).
- Intra-sentential switching: the most complex type among the three, can take place at clausal, sentential or even word level. For example, "Als Alice gesagt hat "that's great", waren sie alle zufrieden" (When Alice said "that's great", everyone was happy).
- Tag-switching: involves inserting a borrowed word in one language into an utterance that is otherwise entirely in another language. For example,"Das ist mein Lifestyle!" (This is my lifestyle!).

2.2 Reasons of Code-Switching

Researchers have investigated the factors that motivate code-switching. Bilinguals are driven towards code-switching whenever the second language is linguistically easier, for example, whenever a word is not accessible in the first language [17, 28, 7], and whenever some words are easier, more distinguishable and easier to use or the concepts involved are easier to express in that languages [7]. Ritchie & Bhatia [26] also reported that code-switching is affected by the Participant Roles and Relationship; whether bilinguals code-mix or not depends on whom they talk to. The authors also stated that the topic of the conversation affects the code-switching behavior [29]. This finding was also confirmed by Velásquez [29], where it was found that code-switching stood out in certain topics, including family, school, ethnicity, and friends. Code-switching is also affected by social factors, such as age, gender, religion, level of education and social class [3, 25, 26]. It was also found that code-switching can be done intentionally for the speaker's own benefit. Janet Holmes [18] has mentioned that code-switching can be used on purpose in order to attract attention and to persuade an audience. Nerghes [22] also reports on the impart of code-switching to reflect a certain socioeconomic identity which can give the speaker more credibility and reliability. Cheng [7] also reported that code-switching can be used to capture attention, appeal to the literate/illiterate, or exclude another person from the dialog. It has also been agreed by several researchers that a speaker may code-switch intentionally to express group solidarity [7, 28, 14, 25] or reflect social status [14]. As stated by Peter Auer [2], "Code-switching carries a hidden prestige which is made explicit by attitudes". In this paper we aim at further investigating some of the aforementioned factors as well as filling the gap by investigating the speaker's personality factors that affect code-switching, as well as the speaker's background and past experiences.

2.3 Temperament and Character Inventory

The Temperament and Character Inventory (TCI) [15] is a self-report questionnaire specifically designed to identify the intensity of its seven basic personality dimensions and the relationship between them. The seven dimensions interact together to make up the individual's personality. The TCI consists of four temperament dimensions: Novelty Seeking (NS), Harm Avoidance (HA), Reward Dependence (RD) and Persistence (PS) and three character dimensions: Self-Directedness (SD), Cooperativeness (CO), Self-Transcendence (ST). Each of these dimensions in turn has its own sub-dimensions. The TCI was chosen instead of the more commonly used Five Factor Model [21] as it offers more detailed sub-dimensions that are of interest for the designed virtual environments. Both models can be mapped to each other.

3 Study

To answer our research questions and get more insights about the correlations of code-switching behavior with personality, character and background, we conducted a user study in which we recorded the following:

1. Temperament and Character Inventory profile
2. Interview speech annotated with Arabic, English, idle utterances and domain of discourse per utterance
3. Background information from a questionnaire
4. Self-awareness and general code-switching questionnaire
5. Electroencephalography (EEG)recordings throughout the interview
6. Heart rate variability (HRV), heart rate (HR) and galvanic skin response (GSR) throughout the interview

Although the data from the bio-sensors were not used in the scope of this work, it can be used in future work to investigate the relation between CS and cognitive load, and to correlate it with user profile and imply whether CS is done intentionally or intuitively.

3.1 Study Timeline

In Figure 1, we present the overall plan of our work. Our study consisted of three phases: personality profiling (TCI), interview and post-questionnaire. In future work, we plan to use our collected data and findings, to build a predictive model using machine learning, that would learn to predict user CS frequency based on given information, including the user profile, background and conversation domain.

Prior to the interview, the participants signed a consent form and filled the TCI profile questionnaire. Then, in the interview, we collected spontaneous code-switched speech. The interview was semi-structured and consisted of a series of open-ended questions. The duration of the interview ranged from 25 to 35 minutes. To not affect users' behavior, we designed the questions to cover a wide range of topics e.g., family, university life, hobbies, quotes, songs, movies, books, and traveling. After finishing the interview questions, the participants were asked to describe seven different pictures. The idea behind this part of the interview is to control the answers and thus have a fair comparison between the different CS frequencies across users. In order to examine the effect of the language used by the interviewer, the questions in the first part of the interview were asked in monolingual Arabic and then in mixed code-switched Arabic-English. After the interview, the participants filled a questionnaire to provide background information (gender, age, educational level, history and travel experience), as well as feedback about their self-awareness of their code-switching behavior.

3.2 Participants and Procedure

We recruited 50 participants (25 males, 25 females) with an average age of 22.1 years (SD = 3.32) using university mailing lists. Participants were teaching assistants and students from different majors ranging from computer science to management and applied arts (17 graduates, 33 undergraduates). All participants are native Arabic speakers who have at least English as a second language. The participants were assigned to the TCI test before coming to the study. Upon arrival, participants were asked to sign a consent form, and we explained the procedure of the study. The aim of the study was explained after the study to avoid any biases. There were three speakers in each setup, two interviewers, a male and a female and the interviewee. The interviewees' speech was recorded using a noise cancellation microphone. Afterwards, the participants were asked to fill the questionnaire.

3.3 Data Analysis

To be able to answer our research question using the collected dataset, we first annotated the data as well as the dependent variables for analysis. Our independent variables included the TCI profile and the interview and questionnaire questions while the dependent variables included the bio-sensor data and the CS behavior of the participants. Two researchers annotated the recorded speech using

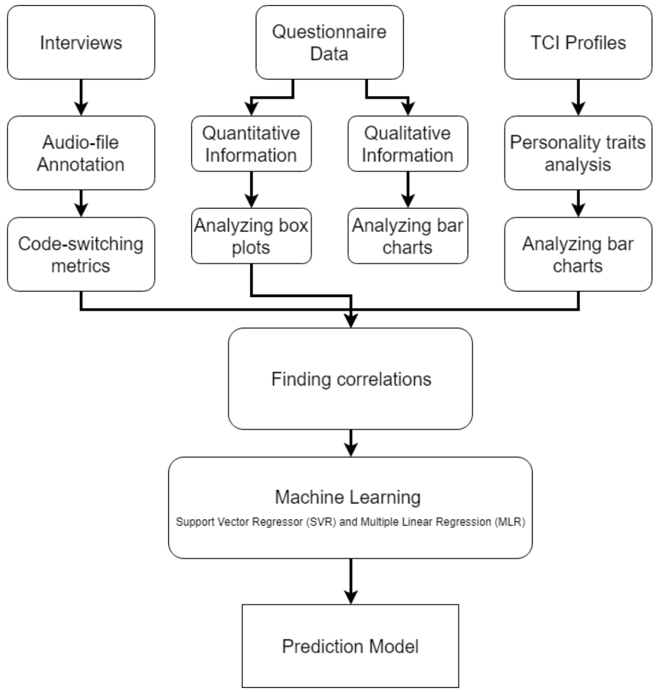

Fig. 1: Flow chart of the data analysis phase

the TranscriberAG annotation tool, to distinguish between English, Arabic and Idle/Interviewer segments. After identifying the boundaries of each language-switching point. The CS percentage was calculated using Eq. 1

$$\frac{\sum(EnglishSecondsPerUtterance \div SecondsPerUtterance)}{\#utterances} * 100 \tag{1}$$

We also divided the interview audio file into five separate domains of discourse based on the interview questions: Background, Travel, Interests and Hobbies, Personality and Character, and Pictures. We then calculated the CS percentage separately for each domain.

4 Results

We analyzed the data to investigate who, why and when participants code switch.

4.1 Who: Personality vs. Code-switching %

For the purpose of our study, we only analyzed the dimensions and sub-dimensions that may affect that may influence the speech behavior. These sub-dimensions include: impulsiveness, fear of uncertainty, shyness with strangers, sentimentality, openness to warm communication, eagerness of effort, perfectionism, self-acceptance, self-actualization, social acceptance, helpfulness and pure-hearted conscience. The character and personality of the participants are logistically distributed. Table 1 shows the Pearson's correlations between the CS percentage and the different relevant character and temperament sub-dimensions pinpointed by a TCI expert. It also shows the mean and average values for each of the sub-dimensions (values ranging from 1 to 5). All of the sub-dimensions show a weak correlation with the CS percentage or no correlation at all. This means that no single sub-dimension contributes to the code-switching on its own. The strongest correlations were found for sentimentality and self-acceptance. This could be due to the confidence acquired from self-acceptance [22]. The negative correlation with sentimentality is in-line with previous research, showing that individuals may code-switch to express certain feelings and attitudes [14] or to distant themselves from emotional events [17]. The weak correlations found in case of the personality traits can be explained by the fact that the situation was kept as a constant for all the collected data which means that the trigger points to affect CS based on personality were not present. This is because depending on the personality type and the situation the CS behavior would vary [10]. The older age group showed higher CS than the younger one. This observed changed behavior is found to relate to the occupation rather than the age. As the older target group were teaching assistants with the teaching language as English, hence, their natural language is code-switched more than students i.e., the younger age group.

Table 1: Results of correlation and statistics of the CS percentage and the TCI sub-dimensions

	Pearson's r	**Mean**	**SD**
CS %		22.92	13.08
Impulsiveness	-0.01	2.48	0.70
Fear of Uncertainty	-0.01	3.41	0.67
Shyness with Strangers	-0.10	3.04	0.85
Sentimentality	-0.25	3.77	0.53
Openness to Warm Comm.	-0.22	3.59	0.59
Eagerness of Effort	0.051	3.48	0.61
Perfectionist	0.11	3.54	0.52
Self-Acceptance	0.25	2.66	0.73
Self-Actualization	-0.15	3.35	0.49
Social Acceptance	-0.16	3.91	0.59
Helpfulness	-0.05	3.65	0.36

4.2 When: Domain of Discourse vs. Code-switching %

We compared the percentage of CS throughout the whole interview with the percentages of CS per domain of discourse. Table 2 shows the means and standard deviation of code-switching per domain. It also shows the r Pearson correlation coefficient between the general CS and CS within a specific domain, proving that CS varying according to the domain, (supported by 92% of the questionnaire responses to belonging questions and [29, 26]). The fact that describing interests and pictures has a high percentage of CS, could be explained by the fact that keywords describing the words often come to mind in the CS language (supported by 80% of the questionnaire responses to belonging questions and [17, 28, 7]).

Table 2: Results of correlation and statistics of the CS percentage per domain

	Background	**Traveling**	**Interests**	**Personality**	**Pictures**
Mean	23.8565418	17.41	28.15	24.15	26.47
SD	14.36	14.41	18.43	15.24	16.56
r	0.80	0.67	0.77	0.82	0.77

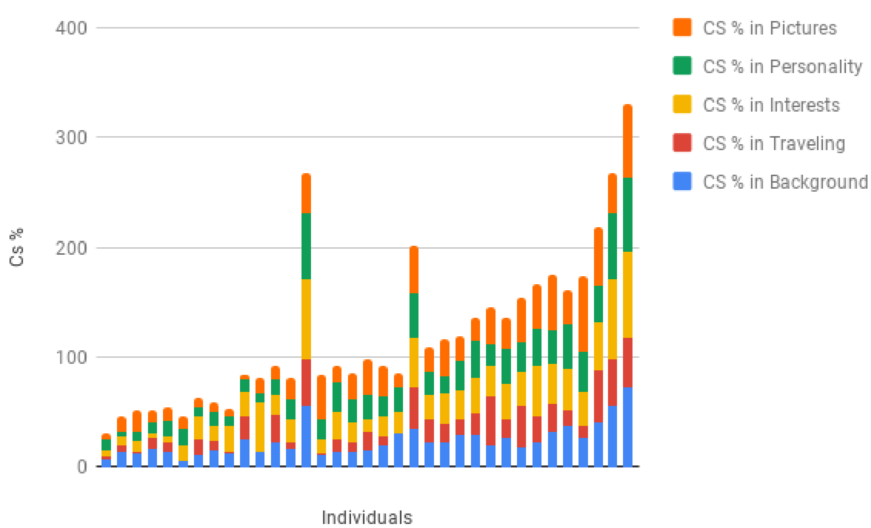

Fig. 2: Stacked step area chart of the CS percentages per domain of discourse.

Fig. 2 shows the comparison between the CS percentages per domain. It shows that while talking about the interests and describing pictures the CS is highest while it is lowest when talking about traveling. The pair-wise p value is < 0.00001 between the CS percentages of any two domains.

4.3 Why: Background Information vs. Code-switching %

We analyzed the background information and demographics of the participants with respect to their code-switching behavior during the interview. The data for this consideration is extracted from responses to the questionnaire questions dealing with the background and history of the participants. Males exhibited lower percentage of CS (mean = 20.79, SD = 13.65) as opposed to females (mean = 24.73, SD = 12.63). The older age group tends to code-switch more often (mean = 28.19, SD = 15.03) than the younger one (mean = 19.71, SD = 15.03). Unexpectedly, the longest stay in a foreign country of the subjects didn't show any correlation with the code-mixing percentage. Our analysis also showed a negative correlation between whether family spoke various languages and the CS percentage, where individuals from uni-lingual families tend to code-switch more. The participants' school type (national or international), showed a correlation with the CS percentage, with some out-liners. In Figure 3, we present some of the factors perceived by participants as gained benefits from code-switching.

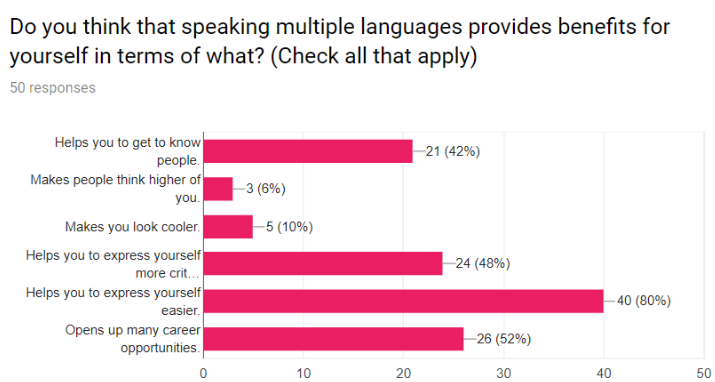

Fig. 3: Perceived benefits provided to the participants from CS.

4.4 Self-awareness

We investigated if participants could rate their own rate of code-switching behavior through self-awareness questions. The participants' answers showed their awareness of their CS behavior, as there is a high correlation ($p < 0.01$) between their reported level of code-switching and the calculated code-mixing percentage. The language considered as the mother tongue by the participants did not show a correlation as the participants who chose pure Arabic as their mother-tongue have very similar CS percentage (MEAN = 22.46) to those who viewed their mother-tongue as code-switched (MEAN = 24.81). On the other hand, the minimum and maximum values are higher in the participants who consider their mother tongue to be code-switched (Arabic-English). According to the subjects' answers on how often they code-switch the results showed a positive correlation between their own prediction of their CS frequency and the calculated code-mixing percentage. The participants' replies showed a correlation between how often do their friends' code-switch and the calculated CS percentage which proves that people adjust their language and thus CS behavior depending on the people they are surrounded with. Only 8 % believed that their CS behavior did not vary depending on the environment, which were all participants with low CS percentages.

4.5 Potential Use-Cases

Finding correlations between an individual's code-switching level and his/her character and background has numerous use cases. This correlation can be used to predict the code-switching percentage of an individual which in turn can be used to adapt the NLP applications to suit the individual's CS behavior and thus achieve higher accuracies. This is particularly useful for smart assistants as it would improve the processing of user's input. It can also be later used to have the smart assistant talk to the user in the same manner as the user making it more familiar and human like. This would be done by programming the smart assistant to mimic the user's talking style by enforcing the same percentage of code switching. Another possible use case is for Facebook Translate. During translation and dictation having a prediction of the a user's code-switching frequency can help improve the system accuracy. In general, User-profiling in terms of CS behavior can be used to generate user-adaptive models, which can help improve the performance of the NLP application involved. Our research can also be used to ease communication by using the profile of the people we are talking to, in order to predict their code-switching behavior. These profiles can be extracted from different daily cues to provide the needed features of a specific individual as proposed in [12].

5 Conclusion and Future Work

In this work, we proposed our investigation on the correlation between users' profiles and code switching behavior in the scope of Egyptian Arabic-English language. Through our review of related work, we concluded that no prior work explored the correlation between code-switching behavior and users' personality and background, as well as the lack of experimental studies supporting theoretical theories on the reasons of code-switching. We investigated the effects of personality (TCI profile), background (stay in foreign country, school type, language requirements, having multilingual family and friends), domain of discourse on the code-switching behavior. Our findings gave initial insights on the correlation between code-switching and the investigated factors. Our findings also highlight that users showed high level of awareness concerning their code switching behavior. This would pave the road to enhanced NLP applications, since users can feed to the system their percentage of code switching, and the system can adapt accordingly. In summary, our results reveal the feasibility of having a system that predicts the code-switching behavior based on their background, personality and context. Hence, we envision that our work can serve as an initial building block to understanding the code-switch behavior.

To further extend the results of this research, we need to collect a larger sample of interview speech with a wider range of participants. We need to conduct the study with participants of more diverse age groups, occupations, educational background and social classes, as our preliminary findings showed that these might have an effect on CS behavior. We also intend to include the personality traits from the Five Factor Model of Personality, as well as other factors affecting behavior, i.e. different character attributes [10]. A bigger dataset would enable us to use different machine learning techniques to predict the code-switching behavior from the features of the three measured cues of personality traits, background information and domain of discourse. Moreover, we can combine the collected data with bio-sensor information (EEG, HRV, eye trackers) to help measure the physiological state and cognitive load in various instances of CS. It would be interesting to investigate whether the cognitive load at CS points can be used to distinguish between intentional and intuitive CS. Also, although our findings are in-line with previous socio- and psycho-linguistic studies, we cannot generalize them to other language pairs, as CS is language-dependent, where the behaviour of users varies across language pairs. Therefore, it would be interesting to conduct the same study on different language pairs, and explore the similarities and differences of users' CS behaviour across languages and cultures. Finally, this work is to be extended to adapt NLP applications based on users' profiles.

References

1. Ardila, A.: Spanglish: an anglicized spanish dialect. Hispanic Journal of Behavioral Sciences **27**(1), 60–81 (2005)
2. Auer, P.: Code-switching in conversation: Language, interaction and identity. Routledge (2013)
3. BENGUEDDA-KESRAOUI, A.: Sociolinguistic Dimensions of Code switching: the Role of Social Factors in its Occurrence in an Algerian Context, Tlemcen Speech Community. Ph.D. thesis (2017)
4. Bentahila, A.: Language attitudes among Arabic-French bilinguals in Morocco. Multilingual Matters Clevedon (1983)
5. Çetinoglu, Ö.: A turkish-german code-switching corpus. In: LREC (2016)
6. Chen, C.m.: Two types of code-switching in taiwan. In: 15th Sociolinguistics Symposium (2004)

7. Cheng, K.K.Y.: Code-switching for a purpose: Focus on pre-school malaysian children. Multilingua **22**(1), 59–78 (2003)
8. Cotterell, R., Renduchintala, A., Saphra, N., Callison-Burch, C.: An algerian arabic-french code-switched corpus. In: Workshop on Free/Open-Source Arabic Corpora and Corpora Processing Tools Workshop Programme. p. 34 (2014)
9. El Bolock, A.: Defining character computing from the perspective of computer science and psychology. In: Proceedings of the 17th International Conference on Mobile and Ubiquitous Multimedia. pp. 567–572. ACM (2018)
10. El Bolock, A.: What is character computing? In: Character Computing, pp. 1–16. Springer, Cham (2020)
11. El Bolock, A., Abdelrahman, Y., Abdennadher, S.: Character computing
12. El Bolock, A., Abdennadher, S., Herbert, C.: Applications of character computing from psychology to computer science. In: Character Computing, pp. 53–71. Springer (2020)
13. El Bolock, A., Salah, J., Abdelrahman, Y., Herbert, C., Abdennadher, S.: Character computing: Computer science meets psychology. In: 17th International Conference on Mobile and Ubiquitous Multimedia. pp. 557–562. ACM (2018)
14. Eldin, A.A.T.S.: Socio linguistic study of code switching of the arabic language speakers on social networking. International journal of English linguistics **4**(6), 78 (2014)
15. Gutierrez-Zotes, J., Bayon, C., Montserrat, C., Valero, J., Labad, A., Cloninger, C., Fernandez-Aranda, F.: Temperament and character inventory-revised (tci-r). standardization and normative data in a general population sample. Actas españolas de psiquiatría **32**(1), 8–15 (2004)
16. Hamed, I., Elmahdy, M., Abdennadher, S.: Collection and analysis of code-switch egyptian arabic-english speech corpus. In: LREC (2018)
17. Heredia, R.R., Altarriba, J.: Bilingual language mixing: Why do bilinguals code-switch? Current Directions in Psychological Science **10**(5), 164–168 (2001)
18. Holmes, J., Wilson, N.: An introduction to sociolinguistics. Routledge (2017)
19. Li, D.C.: Cantonese-english code-switching research in hong kong: A y2k review. World Englishes **19**(3), 305–322 (2000)
20. Lyu, D.C., Tan, T.P., Chng, E.S., Li, H.: An analysis of a mandarin-english code-switching speech corpus: Seame. Age **21**, 25–8 (2010)
21. McCrae, R.R., John, O.P.: An introduction to the five-factor model and its applications. Journal of personality **60**(2), 175–215 (1992)
22. Nerghes, A.: The impact of code-switching on persuasion: An elaboration likelihood perspective. Wageningen University (2011)
23. Poplack, S.: Sometimes i'll start a sentence in spanish y termino en espanol: toward a typology of code-switching1. Linguistics **18**(7-8), 581–618 (1980)
24. Poplack, S.: Code-switching (linguistic). In: Smelser, N.J., Baltes, B. (eds.) International Encyclopedia of the Social and Behavioral Sciences, pp. 2062–2065 (2001)
25. Rihane, W.M.: Why do people code-switch: A sociolinguistic approach
26. Ritchie, W.C., BHATIA, T.: 15 social and psychological factors in language mixing. The handbook of bilingualism and multilingualism p. 375 (2013)
27. Sayahi, L.: Code-switching and language change in tunisia. International Journal of the sociology of language **2011**(211), 113–133 (2011)
28. Tariq, A., Bilal, H., Abbas, N., Mahmood, A.: Functions of code-switching in bilingual classrooms. Research on Humanities and Social Sciences **3**(14), 29–34 (2013)
29. Velásquez, M.C.: Language and identity: Bilingual code-switching in Spanish-English interviews. Ph.D. thesis (2010)