A Comparative Study of Computer-mediated and Spoken Conversations from Pakistani and U.S. English using Multidimensional Analysis

Muhammad Shakir and Dagmar Deuber

Department of English, University of Münster
Johannisstr. 12 - 20, 48143 Münster, Germany
muhammad.shakir@uni-muenster.de, deuber@uni-muenster.de

Abstract

The present study compares four computer-mediated conversational registers (comments, Facebook (FB) groups, FB status updates and tweets), and spoken conversations from Pakistani and U.S. English using Biber's Multidimensional Analysis framework on three dimensions of variation, i.e. (i) Interactive versus Descriptive Explanatory Discourse, (ii) Expression of Stance, and (iii) Informational Focus versus 1st Person Narrative. Spoken conversations have a high score on dimension 2, while CM conversations show register and regional variation on dimension 1 and 3. FB groups are significantly different between both regional varieties, followed by FB status updates, comments and tweets. Pakistani FB groups discuss self-help related topics, and appear to be slightly interactive and highly informational, while the U.S. ones are interactive and narrative discussing community and political issues. Pakistani FB status updates and tweets use English mainly for informational purposes, while the U.S. counterparts have an interactive and personal orientation indicating a wider functional role of English.

Keywords: Register Variation, Multidimensional Analysis, World Englishes, Conversations

1. Introduction

Language users converse with each other to exchange news, views and ideas in an informal way (Oxford Online Dictionary, 2017). Traditionally conversations have been spoken only. With the advent of the internet, another medium has been added, i.e. the written medium. Spoken and newly emerging computer-mediated (CM) conversations are different in ways like turn-taking (Herring, 2011) or synchronicity (Bieswanger, 2016), but at the same time they are linguistically similar to each other (Jonsson, 2015). Though extensively studied, CM conversations need to be studied using a comparative and multi-dimensional approach (Herring, 2011) like Biber's (1988) Multidimensional Analysis (MDA) model, which combines an analysis of situational context with lexico-grammatical features, and interprets them functionally (Biber & Conrad, 2009). Present research paper aims to study emerging CM registers – i.e. comments, Facebook (FB) groups, FB status updates and tweets – in relation to spoken conversations. MDA studies on CM registers have, until now, largely focused either on U.S. English (Grieve et al., 2010) or native varieties of English (Biber and Egbert, 2016). Pakistani English is an outer circle variety in Kachru's (1992) three circle model, which is an important tool in the linguistic repertoire of Pakistani internet users, but not widely studied in relation to the internet. On the other hand, U.S. English, is an inner circle and globally dominant variety, which may be influencing varieties like Pakistani English due to contact and technological influence on the internet. Previous research (e.g. Hardy and Friginal, 2012) suggests that there might be differences between inner and outer circle varieties of English regarding the use of CM registers. Hence further aim is to combine the study of register and regional variation.

1.1 Previous Research

Spoken conversations are generally involved and interactive (Biber, 1988). However, later studies also show additional dimensions like narrativity, informational focus (Biber, 2004), and expression of stance (Biber, 2006). Various types of CM conversations have been studied using MDA. Collot and Belmore (1996) applied Biber's (1988) MDA to study bulletin boards – an ancestor of today's FB groups – and found them nearer to public interviews in spoken conversations. FB status updates and tweets have been said to be CM equivalents of spoken conversations but quite different (Sardinha, 2014), and to be highly informational and descriptive instead of being involved and interactive (Titak & Roberson, 2013). Similarly, comments have been found to be involved, personal and past oriented (ibid). Lastly, studies using MDA to find out regional variation, e.g. Xiao (2009) and Coats (2016), do not involve the comparison of CM and spoken conversations.

2. Material and Methods

Categories	Pakistani English		U.S. English	
	Words	Texts	Words	Texts
Comments	334,447	794	342,517	747
FB groups	163,940	502	163,158	426
FB S. U.	67,737	104	68,819	108
Tweets	58,771	115	62,086	103
Conv.	158,521	85	487,476	111

Table 1: Description of the corpus

Table 1 describes the data for both varieties. Four registers were selected for CM conversations as they were publicly accessible on the internet. Comments were collected from various blogs (single- and multi-writer blogs, newspaper blog posts, and technology blogs) using the website downloader software DarcyRipper and a custom software written in C#. The data for FB groups was manually copied and cleaned after identifying groups originating from Pakistan (mostly closed groups) and the U.S.A. (mostly open groups). Status updates were also manually collected by identifying user profiles from member lists of already scraped groups. Twitter profiles were identified from realtime tweets originating from Pakistan and the U.S.A. The tweets were downloaded using a custom software. The CM conversations data was reviewed and edited for spam, automatic messages, Roman Urdu code switching, and non-standard spellings to facilitate the tagger. However, the spoken data was not reviewed. The data for Pakistani English was extracted from an under-development corpus of Pakistani English (ICE-PK), and for the U.S. variety from the Corpus of Contemporary American English (COCA). The entire data was collected for the time period of 2009-15.

The data was then tagged using Biber Tagger (Biber, 1988; 2006). The tagger tags only approximately 140 specific lexico-grammatical features. Hardy and Friginal (2012) have reported up to 93% accuracy of Biber Tagger on blog posts. Though desirable, such a manual verification of tagging accuracy was beyond the scope of the present research. A new MDA was conducted following guidelines provided in Biber and Gray (2013) and Egbert and Staples (*forthcoming*). The statistical software package R was used to perform Exploratory Factor Analysis. 61 lexico-grammatical features were selected by studying previous research on conversational registers (Biber, 2004; Titak & Roberson, 2013; Biber et al., 1999). After conducting multiple factor analyses with factor solutions from 2-7, a 3-factor solution with Principal Axis Factoring as factor extraction method and Promax as rotation method was deemed fit to describe the data. The details and descriptive statistics are provided in table 2.

Factor	+/-	Linguistic Features with Loadings
1	+	present tense 0.70, 2nd person pronouns 0.49, contractions 0.38, activity verbs 0.34, models of prediction 0.29, models of possibility 0.27 (1st person pronouns 0.34)
	-	prepositions -0.40, attributive adjectives -0.38, nominalisations -.35 (word length -0.29)
2	+	*that* deletion 0.57, mental verbs 0.49, *that* clauses controlled by verbs 0.48, *that* clauses controlled by communication verbs 0.45, communication verbs 0.43, *that* clauses controlled by factive verbs 0.40, *that* clauses controlled by likelihood verbs 0.40 (communication verbs in other contexts 0.30)
	-	(common nouns -0.38)
3	+	word length 0.53, common nouns 0.51, communication verbs in other contexts 0.40, process nouns 0.34, abstract nouns 0.27 (communication verbs 0.38)
	-	1st person pronouns -0.38, adverbs of place -0.33, general adverbs -0.32, past tense -0.29 (contractions -0.37), (nominalisations -0.25)

Other Descriptive Statistics	
Total Variance Explained	22%
Variables in final FA	25
Cut-off	+/- 0.25
Kaiser-Meyer-Olkin (KMO) Measure of Sampling Adequacy	0.54 (Classification: Miserable)

Table 2: Results of Factor Analysis and other descriptive statistics

The total variance explained and KMO values of the factor analysis are quite low. As Egbert and Staples (*forthcoming*) analysed in their study, total variance explained values have been generally low for MDA studies. This is especially the case with internet-based registers. Similarly, they also reported KMO value less than .60 for one of their previous studies. A possible reason might be that the lexico-grammatical variables depend on other variables not included in present analysis. Heterogeneity of the data could be another possible reason.

Each factor in the solution has feature groups with positive and negative loadings, which are mutually less likely to co-occur (Biber, 1988). High factor loading indicates the feature is salient, and vice versa. The features within brackets overlapped with the ones in other factors, hence they were used in the interpretation of factors to dimensions, but not for dimension score calculation. The dimension scores were calculated for each text by summing z-scores of positive as well as negative features, and finally by subtracting negative total score from positive total score. The mean scores for each register category were also calculated to compare the registers on each dimension. Parametric (One-way) or non-parametric (Kruskal-Wallis) ANOVA and respective post-hoc tests were used to check if corresponding registers had significant differences between the regional varieties.

3. Results and Discussion

3.1 Dimension 1: 'Interactive versus Descriptive Explanatory Discourse'

Dimension 1 has eleven features with 7 on positive and 4 on negative side. The features on positive side belong to an interactive discourse (Grieve et al., 2010). The less important features, such as possibility and prediction modals combine with human subjects and dynamic verbs to denote to intrinsic meaning (Biber et al., 1999). The texts with a high positive score are from FB groups. They discuss about present and future events, and are highly interactive.

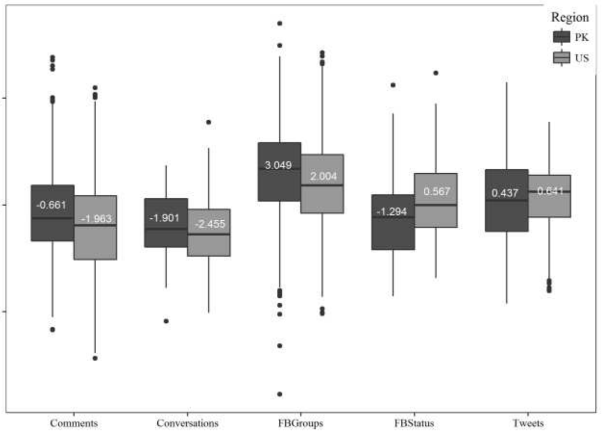

Figure 1: Conversational registers on Dimension 1: 'Interactive versus Descriptive Explanatory Discourse' (One-way ANOVA: F(9, 3085) = 70.23, p < .001; Post-hoc Tukey HSD significant groups between varieties: Comments, FB groups, FB status updates)

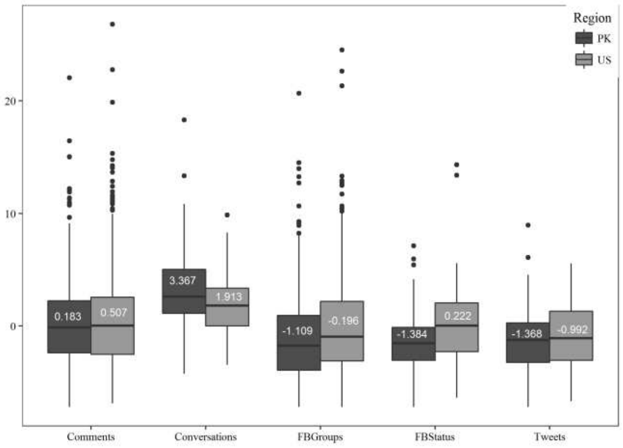

Figure 2: Conversational registers on Dimension 2: 'Expression of Stance' (Kruskal-Wallis ANOVA: H = 206.0, p = 0; Post-hoc Conover-Iman Test significant groups between varieties: FB groups, FB status updates)

The features on negative side are prepositions, attributive adjectives, and nominalisations. Attributive adjectives indicate the presence of descriptive discourses. Prepositional phrases are "the most common type of post-modifiers" (Biber et al, 1999, p. 631). A look at high scoring texts from comments and FB groups show that the texts are descriptive and explanatory in general. Thus, combining positive and negative features the dimension 1 can be interpreted as 'Interactive versus Descriptive Explanatory Discourse'.

Figure 1 shows comparison of register categories on dimension 1. Comments, FB groups and FB status updates in Pakistani English are significantly different from their counterparts in U.S. English, while tweets and conversations do not have significant differences. FB groups is by far the most interactive register, while the category of conversations has the highest inclination towards descriptive and explanatory side of the dimension.

3.2 Dimension 2: 'Expression of Stance'

Dimension 2 has eight linguistic features on positive side, and only one feature on negative side. The positive features include communication verbs like *ask, shout, tell* etc., which show the activity of communication. Mental verbs like *think, know, love, want* etc. are used for cognitive meaning as well as to express attitudes of the speakers (Biber et al., 1999). *That* complement clauses controlled by communication and likelihood verbs are used to convey stance (Biber, 2006), or the presence of reported speech or activities (Titak and Roberson, 2013). An examination of texts with high positive scores from comments, conversations and FB groups show that they contain the elements of opinion or stance. Considering only positive features of this dimension, it can be interpreted as 'Expression of Stance'.

Looking at the results in figure 2, expression of stance seems largely related to spoken conversations, which have the highest scores among all registers. Pakistani conversations have a higher score and a wider range as compared to the U.S. data, which is probably due to a wider variety of conversations (face-to-face, talk shows, interviews etc.). Comments do not have high mean scores, which indicates the possible presence of other stance marking devices like stance adverbs, nouns and adjectives as observed by Biber (2006, p. 92). FB groups, status updates and tweets are less stance oriented, though both FB related registers show significant differences between Pakistani and U.S. English.

3.3 Dimension 3: 'Informational Focus versus 1st Person Narrative'

Dimension 3 contains twelve linguistic features with six features on either side. The positive features include various kinds of nouns and word length, which generally have a positive correlation with each other, i.e. a higher frequency of nouns indicates lengthier words. The majority of texts with high positive score on this dimension are from Pakistani FB groups, which either contain job ads followed by infrequent formulaic comments, or discussions related to study that include abstract and process nouns.

Among features on negative side, 1st person pronoun and contractions normally occur in informal texts with a personal focus. Past tense verbs have been found relevant to narrative texts (e.g. Biber, 1988). The texts with high negative score are from U.S. FB groups, which generally talk about events with the mention of places in 1st person using past tense. Combining both interpretations, dimension 3 can be labelled as 'Informational Focus versus 1st Person Narrative'.

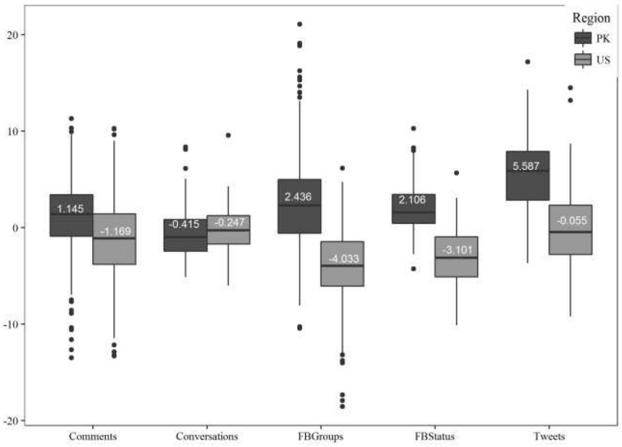

Figure 3: Conversational registers on Dimension 3: 'Informational Focus versus 1st Person Narrative' (Kruskal-Wallis ANOVA: H = 864.43, p = 0; Post-hoc Conover-Iman Test significant groups between varieties: Comments, FB groups, FB status updates, Tweets)

Figure 3 elaborates mean dimension scores of all registers on dimension 3. Though all CM conversational registers show significant differences between Pakistani and U.S. English, spoken conversations do not show much variation except a slight orientation towards the narrative side of the dimension. The most obvious differences are between FB groups, status updates and tweets, where all Pakistani registers have information focused orientation, while U.S. registers incline towards 1st person narration.

4. Conclusion

The differences between spoken and CM conversations mainly appear to be on dimension 1 and dimension 2. On dimension 1, spoken conversations incline towards descriptive and explanatory discourse. The reason for U.S. English seems to be the selection of spoken register for this category, i.e. broadcast discussions, which are different from spontaneous face-to-face conversations. For Pakistani English, though spoken conversations come from more than one registers, for example interviews, talk shows, and student face-to-face conversations, it appears that even face-to-face conversations are generally descriptive and explanatory instead of being involved and interactive. Dimension 2 'Expression of Stance' has been observed previously as well (Biber, 2004; 2006) for spoken conversations. The results apparently confirm that CM conversations are similar but quite different from spoken conversations (Titak & Roberson, 2013). Another possible reason could be the lesser representation of spoken registers in U.S. English and a smaller number of words in Pakistani English.

On the other hand, CM conversations show variation on all dimensions between registers as well as between regional varieties. Pakistani FB groups are generally related to study help, job, pet and game related talk, which makes them interactive as well as information oriented. However, U.S. FB groups are related to politics, community related issues, as well as pet and game related talk, so they are a little less interactive but highly inclined towards personal narration. Pakistani comments are slightly more interactive due to comments from "diary type blogs" (Grieve et al., 2010), while U.S. comments are descriptive in contrast due to an abundance of political "commentary type blogs" (ibid). FB status updates and tweets are highly informational in Pakistani English partially due to the use of local languages to talk about personal issues, while that is not the case with U.S. English. To conclude, Pakistani CM conversations differ from U.S. counterparts, though a more representative data of spoken conversations would help to better understand the relation between both types of conversations.

5. References

Biber, D. (1988). *Variation across speech and writing.* Cambridge University Press.

Biber, D. (2004). Conversation text types: A multi-dimensional analysis. In *Le poids des mots: Proc. of the 7th International Conference on the Statistical Analysis of Textual Data*. Louvain: Presses universitaires de Louvain, pp. 15—34.

Biber, D. (2006). *University language: A corpus-based study of spoken and written registers, Vol. 23.* John Benjamins Publishing.

Biber, D. & Conrad, S. (2009). *Register, genre, and style*. Cambridge University Press.

Biber, D. & Gray, B. (2013). Identifying multi-dimensional patterns of variation across registers. In M. Krug & J. Schlüter (Eds.), *Research Methods in Language Variation and Change*. Cambridge University Press, pp. 402—420.

Biber, D., Johansson, S., Leech, G., Conrad, S. & Finegan, E. (1999). *Longman Grammar of Spoken and Written English.* Longman Publications Group. (ISBN: 0-582-23725-4.)

Bieswanger, M. (2016). Electronically-mediated Englishes: Synchronicity revisited. In L. Squires (Ed.), *English in Computer-Mediated Communication: Variation, Representation, and Change Vol. 93*. Walter de Gruyter GmbH & Co KG, pp. 281—300.

Coats, S. (2016). Grammatical feature frequencies of English on Twitter in Finland. In L. Squires (Ed.), *English in Computer-Mediated Communication: Variation, Representation, and Change Vol. 93*. Walter de Gruyter GmbH & Co KG, pp 179—209.

Collot, M. & Belmore, N. (1996). Electronic Language: A New Variety of English. In S. C. Herring (Ed.), *Computer-mediated communication: Linguistic, social, and cross-cultural perspectives Vol. 39*. John Benjamins Publishing, pp. 13—28.

Conversation. (n.d.). In *Oxford Living Dictionaries*. Retrieved May 15, 2017, from https://en.oxforddictionaries.com/definition/conversation.

Grieve, J., Biber, D., Friginal, E. & Nekrasova, T. (2010). Variation among blogs: A multi-dimensional analysis. In Alexander M., & S. Sharoff (Eds.), *Genres on the Web Computational Models and Empirical Studies*. Springer, pp. 303—322.

Herring, S. C. (2011). Computer-mediated conversation Part II: Introduction and overview. *Language@ internet* 8(2), pp. 1—12. Retrieved from http://www.languageatinternet.org/articles/2011/Herring..

Egbert, J. & Staples, S. (forthcoming). Doing multidimensional analysis in SPSS, SAS and R. In T. B. Sardinha & M. V. Pinto (Eds.), *Multidimensional Analysis*. London: Bloomsbury.

Jonsson, E. (2015). *Conversational Writing-A Multidimensional Study of Synchronous and Supersynchronous Computer-Mediated Communication.* PETER LANG LTD International Academic Publishers.

Kachru, B. B. (1992). *The other tongue: English across cultures.* University of Illinois Press.

Sardinha, T. B. (2014). 25 years later Comparing Internet and pre-Internet registers. In T. B. Sardinha & M. V. Pinto (Eds.),, *Multi-Dimensional Analysis, 25 years on A tribute to Douglas Biber Vol. 60*. John Benjamins Publishing Company, pp. 81—105.

Titak, A. & Roberson, A. (2013). Dimensions of web registers: an exploratory multi-dimensional comparison. *Corpora* 8(2), pp. 235—260.

Xiao, R. (2009). Multidimensional analysis and the study of world Englishes. *World Englishes* 28(4), pp. 421—450.