Structural linguistic characteristics of podcasts as an emerging register of computer-mediated communication

Aminat Babayode[1], Laurens Bosman[1], Nicole Chan[1], Katharina Ehret[1,2], Ivan Fong[1], Noelle Harris[1], Alissa Hewton[1], Danica Reid[1], Maite Taboada[1], Rebekah Wong[1]

[1]Department of Linguistics, Simon Fraser University, Canada
[2]Department of English, University of Freiburg, Germany
Corresponding authors: kehret@sfu.ca, mtaboada@sfu.ca

Abstract

Podcasts, a relatively recent audio medium, have risen in popularity since their initial appearance in the mid-2000s. Yet, little is known about their structural linguistic characteristics and their relation to other registers. Addressing this gap in the literature, we apply Biber-style multidimensional analysis (MDA) to a representative sample of Spotify podcast transcripts and compare their structural linguistic characteristics to those of selected computer-mediated registers (e.g., informational blog, interview) as well as traditional spoken registers (e.g., broadcast, conversation). Our results reveal that, while podcasts share some linguistic characteristics with traditional spoken registers such as broadcast discussion and unscripted speech, they are unlike any of the analysed registers. In fact, they exhibit unique structural characteristics combining features of involved spoken language with some features typical of informational production and narration. In short, we show that podcasts are a newly emerging register of computer-mediated communication.

Keywords: computer-mediated communication, register analysis, corpus linguistics, multidimensional analysis, podcasts

1. Podcasts as a new medium

Podcasts are a new audio-based medium, similar to radio, television, and other traditional media facilitating the sharing and broadcasting of content to large audiences (Levinson, 2013). Originally, podcasts were intended to convey information and act as a source of entertainment (Nurekeshova, 2016). Due to their relevance to diverse contexts, podcasts are a versatile form of media, with podcasts that are similar to traditional interview shows, news and politics, audiobooks, music, games, plays, and educational shows. Their broad appeal, however, has resulted in an emerging set of practices that may differ from traditional radio (Berry, 2016). Despite their popularity, little research on the characteristic linguistic features of podcasts exists. Addressing this gap, we apply Multidimensional Analysis (MDA) (Biber, 1988) to provide insight into how linguistic features associated with the conversational style of podcasts differentiate them from other types of broadcasts and other emerging registers of computer-mediated communication.

2. Register variation and MDA

Biber and Conrad (2001) define register as the result of linguistic variation in the lexical and grammatical choices that language users make as appropriate to the context of usage. The tool of choice for analysing register variation is Multidimensional Analysis. MDA is a multivariate statistical technique based on the frequency and co-occurrence of lexico-grammatical features, and examines how the co-occurrence patterns correlate with particular registers (Biber, 1988; Biber and Egbert, 2018). Through dimensionality reduction, MDA allows us to abstractly interpret linguistic features as representing the underlying communicative functions of the texts analysed.

3. Podcast transcripts as corpus

Podcasts are typically available as audio files, but also in transcript format. The data used in this study is a subset of the English part of the Spotify Podcasts Dataset (Clifton et al., 2020). We classified the podcasts by topic as well as length and selected the top 20 categories by number of podcasts. These include topics such as Arts, Business, Comedy, History, Religion & Spirituality, Science, Sports, or True Crime. We then divided sorted each topic into 4 bins by length (up to 15 minutes in length; 15-30 minutes; 30-60 minutes; >60 minutes), from which we sampled 10% from each bin across the top 20 topics. Our final corpus counts 9,789 podcast transcripts and 64,239,291 words.

MDA involves exploring the features of the register in question in comparison to other registers to situate the register of interest in a space of linguistic variation. For this comparison, we draw on different corpora. First, we use selected registers of the Corpus of Online Registers of English (CORE) (Biber and Egbert, 2018)), to compare podcasts to other registers of computer-mediated communication (e.g., interactive discussion, informational blog). Second, we use the British National Corpus (BNC) (Aston and Burnard, 2020)) as a source for traditional spoken registers (e.g., conversation, broadcast discussion), the Santa Barbara Corpus of Spoken American English (SBCSAE) (Du Bois et al., 2000)) for conversation in a North American context, and the English Pear Stories[1] as a source of oral narratives.

In total, we analyse 9 different traditional registers and 10 registers of computer-mediated communication totalling over 27 million words, comparing them to the 64 million words in the podcasts (see Appendix, Table 1).

Short abstract for presentation at the *International Conference on CMC and Social Media Corpora for the Humanities*. Mannheim, Germany. September 2023. https://www.uni-mannheim.de/cmc-corpora2023/

[1] http://www.pearstories.org/english/english.htm

4. Podcasts as an emerging register

4.1. Podcasts and computer-mediated registers

In the MDA of podcasts and computer-mediated registers in CORE two well-defined dimensions emerge; the third dimension was extracted for statistical reasons but is not linguistically interpretable (see Appendix, Table 2). The first dimension, "Involved vs. informational discourse" has two poles. On the positive pole cluster features typical of spontaneous spoken and involved language such as present tense verbs, contractions, and private verbs. The negative pole is defined by only a handful of features, all of which indicate an informational style: nouns, prepositions and perfect aspect. Dimension 2 comprises only positive features, namely, nominalisations, average word length, and predicative adjectives. These features are typical of abstract-informational language and can together with secondary features such as THAT-relative clauses and complements be interpreted as representing "Abstract-informational elaboration".

Looking at the distribution of registers on these two dimensions, we find that all the written registers and interactive discussion are positioned on the negative pole of Dimension 1, i.e., they are representative of informational discourse. On the positive, involved pole, we find the spoken registers podcasts, formal speech, spoken, and interview. As a matter of fact, podcasts emerge as the most involved register in this dataset. On Dimension 2, podcasts are located somewhere in between, along with interview, while formal speech and informational blog are most representative of abstract-informational elaboration. Thus, podcasts clearly emerge as a spoken register and one unlike all the other computer-mediated registers (Appendix, Figure 1). They are strongly characterised by features of spontaneous spoken and involved language and, to some extent, features of abstract-informational elaboration.

4.2. Podcasts and traditional spoken registers

The MDA comparison of podcasts and traditional spoken registers comprises three variational dimensions (see Appendix, Table 3). The first dimension, "Involved vs. informational production" is defined by features typical of spontaneous spoken and involved language such as contractions, emphatics, and first personal pronouns on the positive pole. On the negative pole, it is defined by the co-occurrence of average word length, nouns, attributive adjectives, and other features indicative of information-focused production. The second dimension which we label "Narrative" is largely defined by positive features typically associated with narration: past tense, third person pronouns and perfect aspect. Dimension 3 "Abstract elaboration" consists only of a positive pole which comprises indicators of elaboration and abstract description such as THAT-verb complements, THAT-relatives, and predicative adjectives.

The overall distribution of registers (Appendix, Figure 2) confirms this interpretation of the dimensions, for instance, broadcast news is highly informational while conversation is involved; interviews and oral narrative load high on the narrative dimension and broadcast documentary is information-elaborate. Where, then, are podcasts positioned? Interestingly, none of the three dimensions represents our podcast data very well and it is located in the middle. In terms of other registers it resembles oral narratives, interviews, (unscripted) speech and broadcast discussion across the three dimensions. Hence, podcasts do share some features with these registers but are also clearly unlike any of them. Rather, they uniquely combine features of narration, spontaneous speech and informational production.

5. Conclusion

This paper presented a multidimensional analysis of podcasts as an emerging register of computer-mediated communication. Comparing podcasts to a set of written and spoken computer-mediated registers from the Corpus of Online Registers of English and well-known corpora of broadcasts, conversations, and narratives, we show that podcasts are firmly a spoken register, yet, unlike all the other computer-mediated registers. Our analysis of podcasts and traditional spoken registers confirms this finding: podcasts are clearly a newly emerging register. Precisely, podcasts exhibit some similarities with a range of spoken registers, i.e. interviews, (unscripted) speech, oral narratives and broadcast discussion. Hence, they are characterised by a unique set of features and combine features of on-line spontaneous production, narration, and informational production. This characterisation dovetails with the intended purpose of podcasts as a source of both information and entertainment (Nurekeshova, 2016). It is this versatility of podcasts that makes them unique but probably also makes them a register with a comparatively large degree of internal variability (like the registers broadcast and letters). The natural next step, then, is to explore the extent of register-internal variability and further detail the lexico-grammatical features of the emerging podcast (sub)register(s). Last but not least, despite the fact that our data samples English-language podcasts only, our findings constitute a first step towards understanding and describing podcasts in general.

6. Acknowledgements

We are grateful for helpful comments and suggestions by two anonymous reviewers.

Bibliographical references

Aston, G. and Burnard, L. (2020). *The BNC handbook: Exploring the British National Corpus with SARA*. Edinburgh University Press.

Berry, R. (2016). Podcasting: Considering the evolution of the medium and its association with the word 'radio'. *Radio Journal: International Studies in Broadcast & Audio Media*, 14(1):7–22, April.

Biber, D. and Conrad, S. (2001). Register variation: A corpus approach. In Deborah Schiffrin, et al., editors, *The Handbook of Discourse Analysis*, pages 175–196. Blackwell.

Biber, D. and Egbert, J. (2018). *Register Variation Online*. Cambridge University Press.

Biber, D. (1988). *Variation across Speech and Writing*. Cambridge University Press.

Clifton, A., Reddy, S., Yu, Y., Pappu, A., Rezapour, R., Bonab, H., Eskevich, M., Jones, G. J., Karlgren, J., Carterette, B., and Jones, R. (2020). 100,000 Podcasts: A Spoken English Document Corpus. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5903–5917.

Du Bois, J. W., Chafe, W. L., Meyer, C., Thompson, S. A., and Martey, N. (2000). Santa barbara corpus of spoken american english. *CD-ROM. Philadelphia: Linguistic Data Consortium*.

Levinson, P. (2013). *New New Media*. Pearson, 2nd edition.

Nurekeshova, G. R. (2016). Podcasting as a technical way of interactive communication of XXI century. *European Journal of Natural History*, pages 112–116.

Appendix

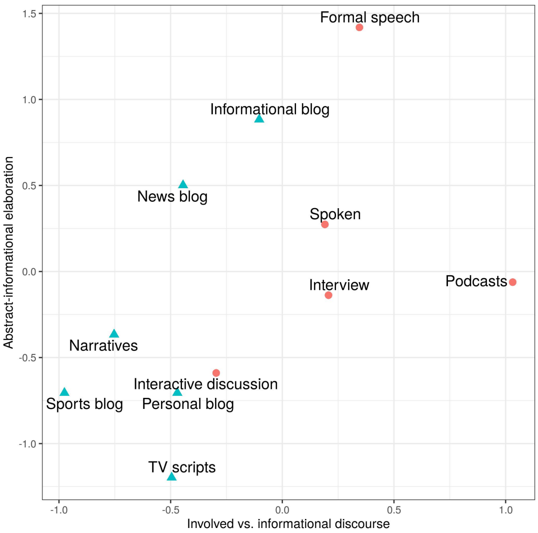

Figure 1: Podcasts compared to other registers of computer-mediated communication in the Corpus of Online Registers of English. Positive values on Dimension 1 indicate involved discourse; negative values on Dimension 1 indicate informational discourse. Red dots index spoken, green triangles index written registers.

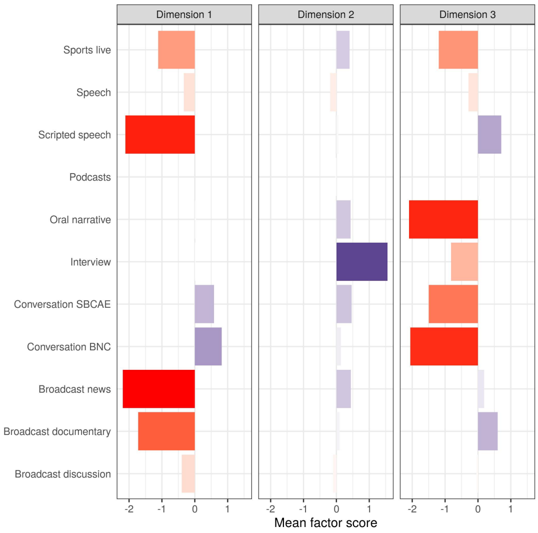

Figure 2: Distribution of podcasts and traditional spoken registers on the three dimensions. Colour intensity indicates strength of mean factor scores. Red bars indicate negative values; blue bars indicate positive values.

Register	Mode	Corpus	Words
Broadcast discussion	spoken	BNC	666,098
Broadcast documentary	spoken	BNC	37,496
Broadcast news	spoken	BNC	225,024
Conversation	spoken	BNC	3,836,745
Interview	spoken	BNC	111,155
Scripted speech	spoken	BNC	164,244
Unscripted speech	spoken	BNC	410,690
Sportslive	spoken	BNC	29,957
Formal speech	spoken	CORE	80,109
Informational blog	written	CORE	2,141,271
Interactive discussion	spoken	CORE	3,099,725
Interview	spoken	CORE	451,593
Narrative	written	CORE	424,614
News report/blog	written	CORE	9,806,239
Personal blog	written	CORE	3,264,463
Spoken	spoken	CORE	224,703
Sports report	written	CORE	2,729,925
TV scripts	written	CORE	32,502
Conversation	spoken	SBCAE	209,308
Oral narratives	spoken	Pear	16,149
TOTAL			27,962,010

Table 1: Registers by corpus, mode (written vs. spoken) and word count.

Dimension 1: Involved vs. informational discourse	
Present tense verbs	0.857
Contractions	0.761
Demonstrative pronouns	0.752
Private verbs	0.749
Causal subordinators	0.663
Emphatics	0.638
BE as main verb	0.607
Demonstratives	0.6
Second person pronouns	0.588
Analytic negation	0.573
Hedges	0.558
Adverbs	0.554
First person pronouns	0.52
Pronoun IT	0.533
THAT deletion	0.446
Pro-verb DO	0.42
Predicative adjectives	0.413
WH-clauses	0.392
Conditional subordinators	0.308
—	—
Nouns	-0.831
Prepositions	-0.637
Average word length	-0.391
Present participle clauses	-0.334
Dimension 2: Abstract-informational elaboration	
Nominalizations	0.783
Average word length	0.711
Attributive adjectives	0.381
Phrasal coordination	0.316

Table 2: Dimensions and features with significant loadings $\geq |0.3|$ for the podcast and CORE data. Positive loadings indicate co-occurrence of the features; negative loadings indicate complementary distribution. Crossloading features with the same polarity and uninterpretable dimensions are excluded.

Dimension 1: Involved vs. informational production	
Contractions	0.846
First person pronouns	0.699
Analytic negation	0.661
Private verbs	0.625
Present tense verbs	0.615
THAT deletion	0.595
Pronoun IT	0.592
Demonstrative pronouns	0.579
Emphatics	0.547
Causal subordinators	0.544
BE as main verb	0.54
Pro verb DO	0.462
WH-clauses	0.412
Hedges	0.4399
Adverbs	0.387
Predicative adjectives	0.309
—	—
Average word length	−0.935
Nouns	−0.826
Prepositions	−0.779
Attributive adjectives	−0.705
Nominalizations	−0.655
Phrasal coordination	−0.612
Present participle WHIZ deletion	−0.467
Past participle WHIZ deletion	−0.464
BY-passives	−0.424
Passives	−0.362
Conjunctions	−0.359
Gerunds	−0.342
Dimension 2: Narrative	
Past tense verbs	0.956
Third person pronouns	0.567
Perfect aspect	0.406
Dimension 3: Abstract elaboration	
THAT verb complements	0.487
Nominalisations	0.472
THAT-relatives (obj.)	0.457
Demonstrative pronouns	0.447
Average word length	0.37
Split auxiliaries	0.325
THAT-relatives (subj.)	0.322
Predicative adjectives	0.314
TO infinitives	0.301

Table 3: Dimensions and features with significant loadings $\geq |0.3|$ for the podcast and traditional spoken data. Positive loadings indicate co-occurrence of the features; negative loadings indicate complementary distribution. Cross-loading features with the same polarity and uninterpretable dimensions are excluded.