# Structural linguistic characteristics of podcasts as an emerging register of CMC

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

Department of Linguistics, Simon Fraser University, Canada ✳ Department of English, University of Freiburg, Germany ✳ kehret@sfu.ca, mtaboada@sfu.ca

## 1. Why study podcasts?

- Podcasts are a new audio-based medium
  - They facilitate the sharing and broadcasting of content to large audiences
  - They serve as both source of information and entertainment
  - They are marked by usage practices different from traditional radio
- Little is known about their structural linguistic characteristics
  - Which linguistic features are used in podcasts?
  - How do their linguistic characteristics differ from other registers?
  - Are they a newly emerging register of CMC?

## 2. Register variation

Registers result from linguistic variation in the **lexical and grammatical choices** that language users make in different **contexts of usage** (Biber & Conrad 2001).

➪ Explore this linguistic variation in podcasts by comparing them to other registers and situate them in a **space of linguistic variation**

## 3. Multi-dimensional analysis

The tool to study register variation is multi-dimensional analysis (MDA).

➪ Reduce a large number of linguistic features to a small number of dimensions

1. Count the frequency of linguistic features in texts
2. Analyse their co-occurrence patterns and correlate them to different registers
3. Interpret linguistic features in terms of their communicative functions in texts

## 4. Podcast transcripts as corpus

Our data samples 64 million words of **podcast transcripts** in English (Clifton et al. 2020) and 27 million words across 9 different **traditional spoken** and 10 **computer-mediated registers** of English. The podcasts cover topics like Arts, Business, Comedy, History, Science, or Sports.

| | |
|---|---|
| Podcasts | Spotify Podcasts Dataset |
| Traditional registers | British National Corpus, Santa Barbara Corpus of Spoken American English, The Pear Stories |
| CMC registers | Corpus of Online Registers of English |

## 5. Linguistic characteristics of podcasts

Podcasts emerge as a firmly **spoken register**. They combine features of **involved** and **on-line spontaneous** discourse with some features of **narration** and **informational language**.

➪ Reflects their versatile use as medium of information and entertainment

**Involved features**

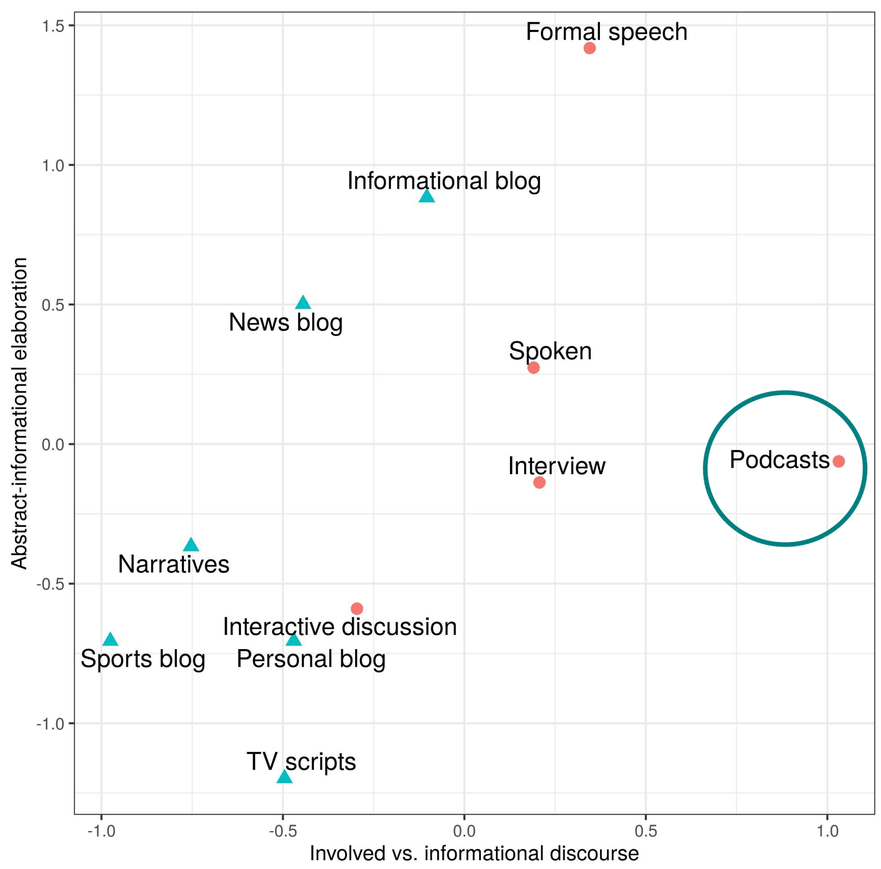
- Contractions, demonstratives, first and second person pronouns, present tense verbs, emphatics like *for sure*, private verbs like *feel*, *think*

**Narrative features**

- Past tense verbs, third person pronouns, and perfect aspect

**Informational features**

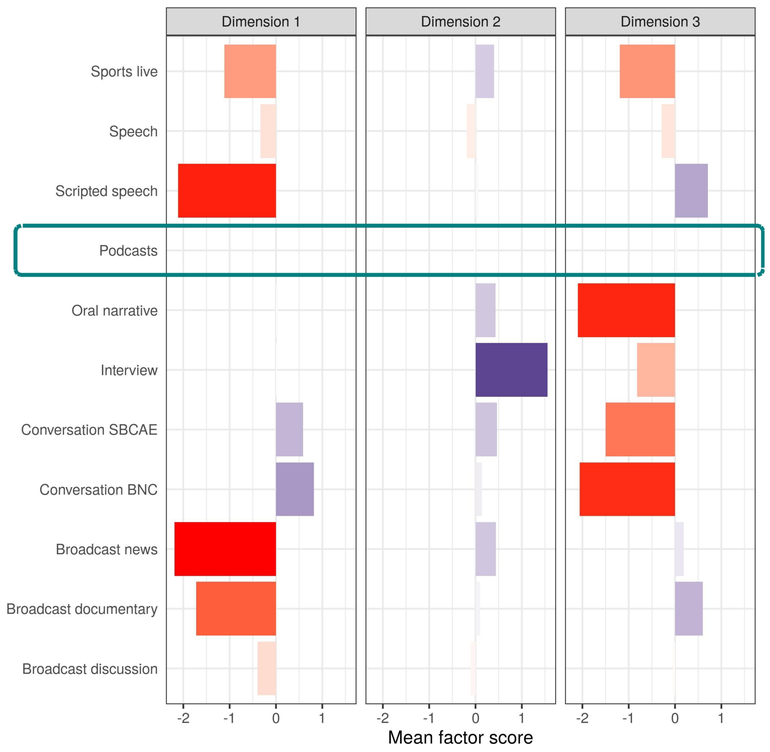
- Average word length, nouns, nominalisations (nouns ending, e.g., in *-ity* or *-tion*), attributive adjectives, passives, or conjuncts




Podcasts and registers of computer-mediated communication in CORE. Red dots index spoken, green triangles index written registers.

## 6. Podcasts are unique

- Podcasts are unlike any of the analysed spoken and written CMC registers
- If compared to other CMC registers, they come closest to interview
- Podcasts are also different from traditional spoken registers across all three dimensions
- On individual dimensions they share some features with the traditional spoken registers oral narratives, interviews and broadcast discussion
- Dimension 1: Involved vs. informational. Dimension 2: Narrative. Dimension 3: Abstract elaboration.




Podcasts and traditional registers of English. Colour intensity indicates strength of mean factor scores. Red bars indicate negative values; blue bars indicate positive values.

## 7. Examples

How are linguistic features used to create the conversational style of podcasts?

Involved features in **red**; narrative in **blue**; informational in **green**

> **That's** how **I** try to live now. Like **that's** how **I've kind of** grown up already like **I** grew vegetables and people **think that's** harder than it is [...]

> This conflict **led** to the **creation** of the 17th parallel [...].

> **but** the plague in the 6th century **evaporated**, **you know**, somewhere between a third or half of the Middle East population **was wiped out**. The Persian Empire **basically** went bankrupt **because** of the plague [...]

## 8. Conclusions & further information

Podcasts do not align well with any other register.

➪ Emerging register of CMC
➪ Some amount of internal variability due to its versatile purposes

Ongoing work

- Exploring the extent of register-internal variability in podcasts
- Describing the lexico-grammatical features of podcast subregisters

Paper, with references: