Methodological issues in designing and evaluating a corpus to identify YouTube videos containing informal English speech.

Christopher R. Cooper

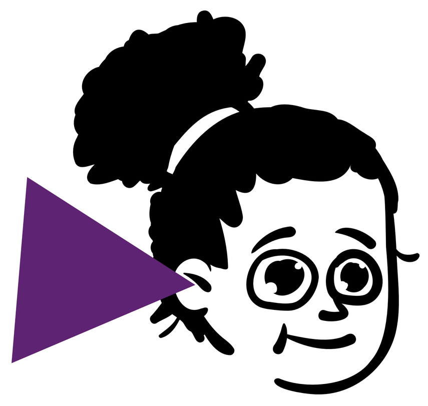
RIKKYO UNIVERSITY
Center for Foreign Language Education and Research

Background

Large, representative sample of language (Ellis & Wulff, 2020)

Track pragmatic situations, Co-occurring words ➡ chunks (Beckner et al., 2009)

Conversation most basic form of human communication (Biber et al., 2021)



Viewing (+ captions) = Comprehension (Durbahn et al., 2020, Gass et al., 2019), vocabulary (Peters & Webb, 2018) & multiword expression acquisition (Majuddin et al., 2021), culture-specific schemata (Gilmore, 2007)

YouTube = 2 billion monthly users

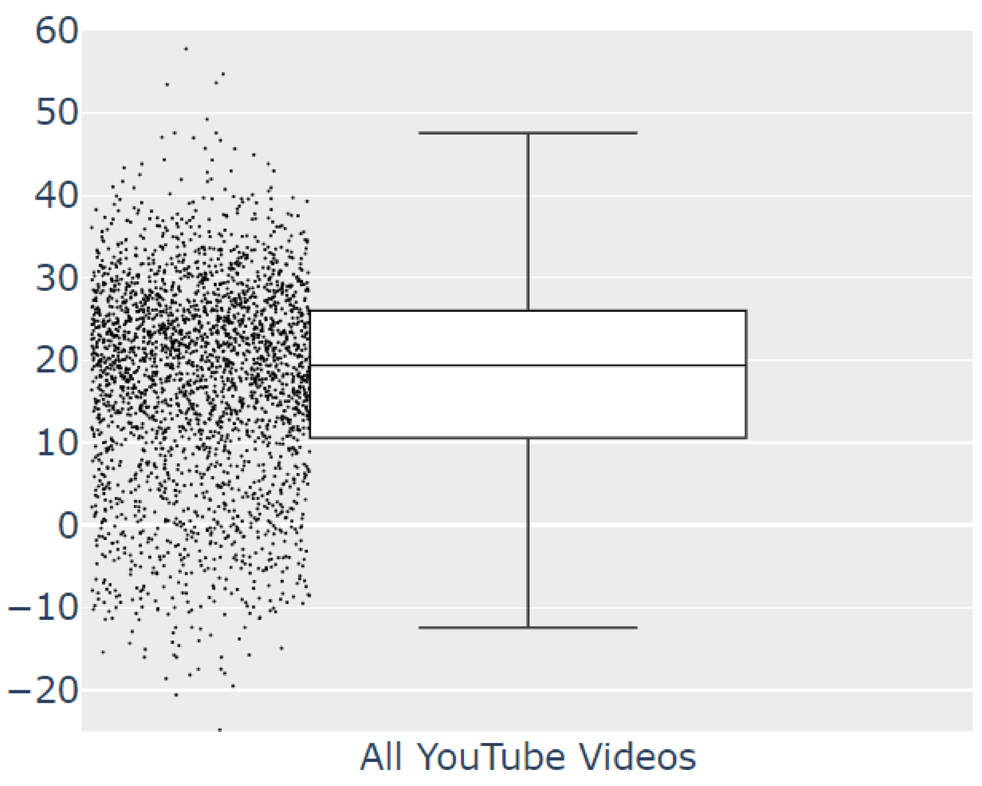
MD Analysis (Biber, 1988) distinguishes between registers. Sitcom (Quaglio, 2009), reality shows, movies (Berber Sardinha and Veirano Pinto, 2019), video game interactive speech (Dixon, 2022) close to conversation

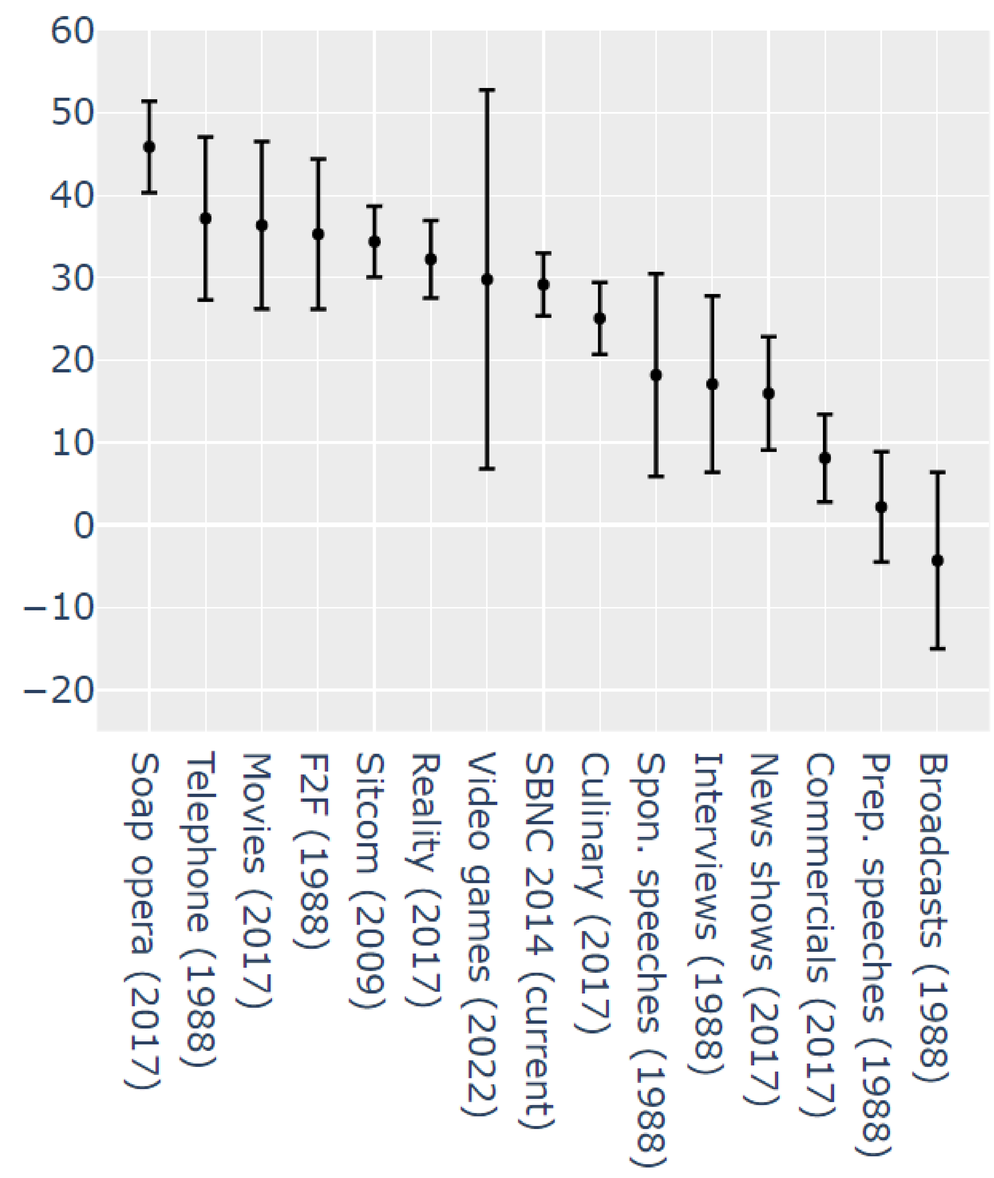
Corpus	Texts	Tokens	Tokens per text *M*	Tokens per text *SD*
YouTube Corpus	2,602	4,351,386	1,683	924
Spoken BNC 2014 Sample (Love et al., 2017)	200	1,837,064	9,185	6,912

MAT tagger (Nini, 2019) Data collection, processing, analysis

Results from the pilot study

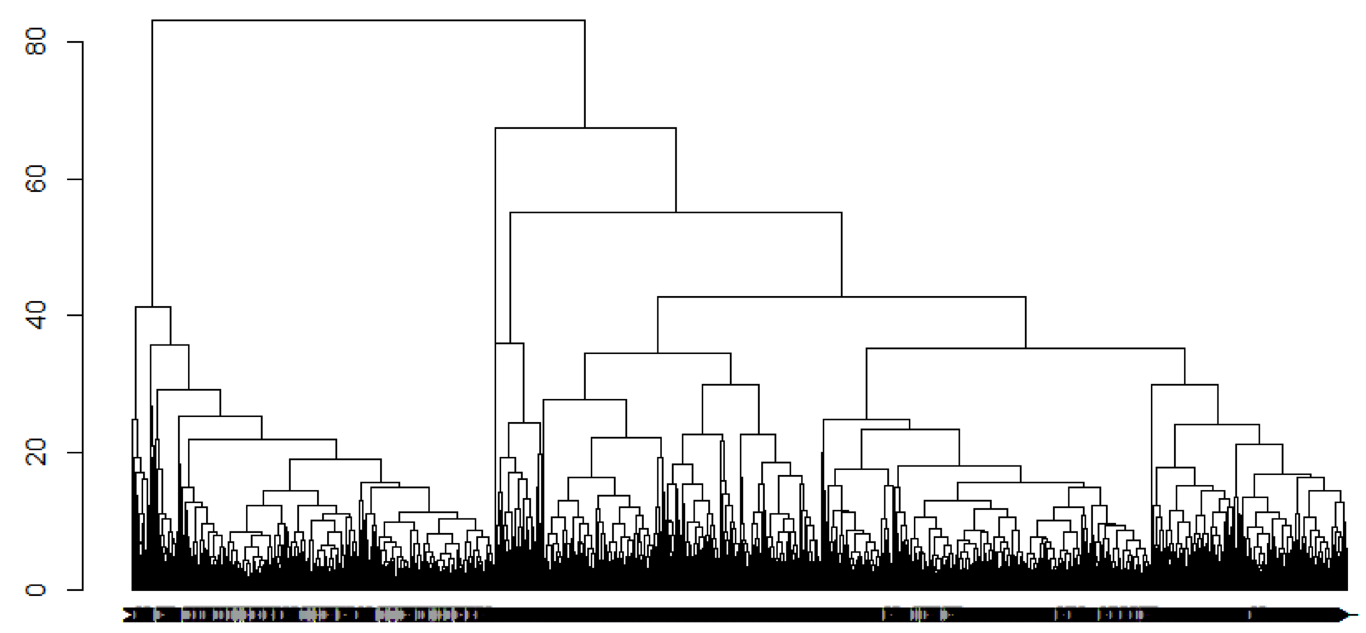
YouTube videos compared with other spoken registers on Biber's (1988) Dimension 1



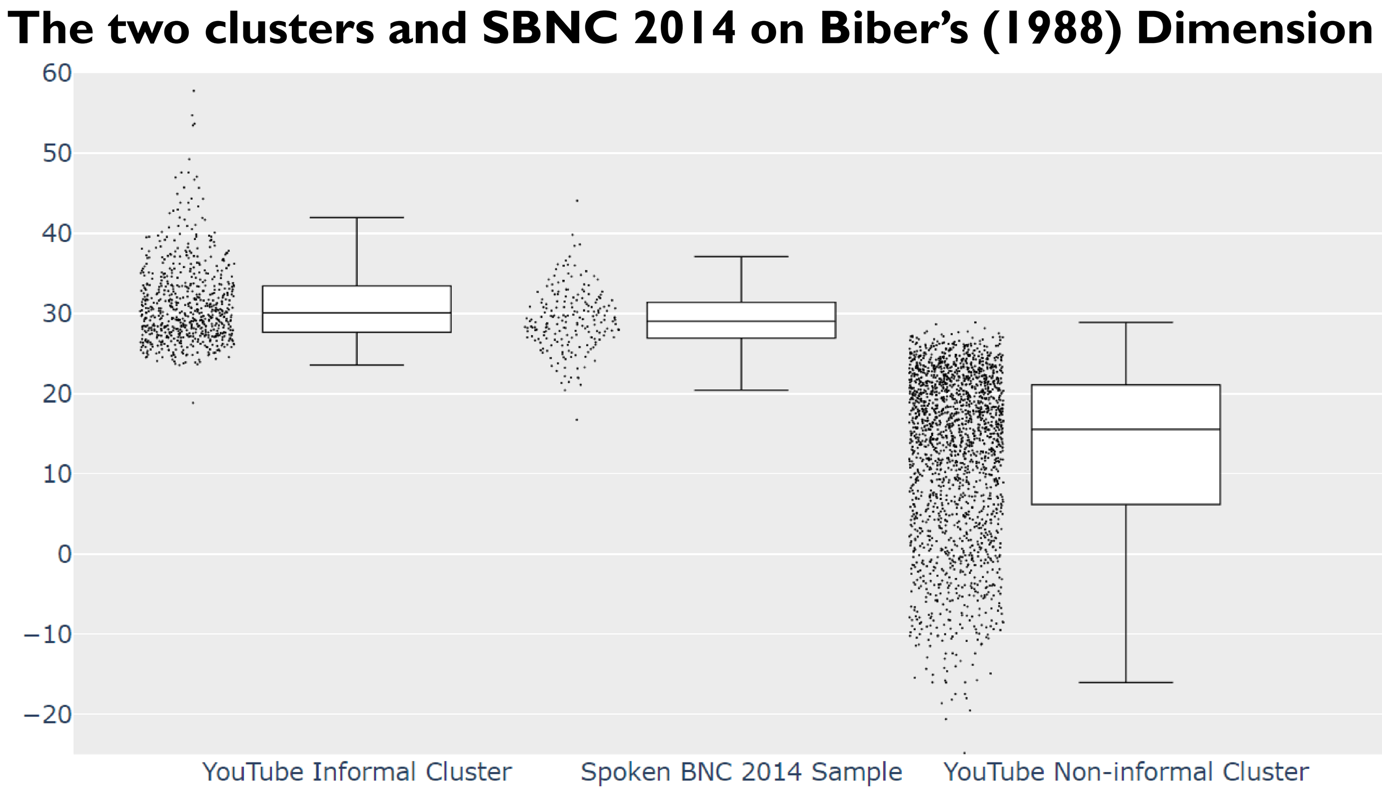


- 666 YouTube videos clustered with 171 Spoken BNC 2014 texts
- Videos in both clusters seemed to have similar topics (but further analysis needed)
- Distinguishing features include 1st/2nd person pronouns, contractions, private & present tense verbs

Cluster analysis to identify informal videos



The two clusters and SBNC 2014 on Biber's (1988) Dimension 1



(Cooper, 2023)

Methodological issues for future research

Corpus size (Egbert et al., 2022) For each linguistic feature

$$n = \frac{S^2}{\left(\frac{.5 * CI\ Range}{t}\right)^2}$$

Standard deviation

Confidence interval range 95% (+ / - 5% of mean score) =10% of mean score

Desired t-value 1.96 for 30+ sample (Biber, 1993)

2,602 → **21,834+**

- Conservative estimate based on least frequent feature
- Past participle clauses (e.g. Built in a single week, the house would stand for fifty years)
- Not common in conversation

Corpus evaluation

Get to know the corpus with keyword analysis (Kilgariff, 2012)

Informal cluster as target
Non-informal cluster as reference

Qualitatively sort words into categories, some examples from pilot study:

Informal	Non-informal
Informal (oh, stuff)	Sport (yard, touchdown)
Greetings (hey, bye)	Numbers (fourth, third)
Pronouns (me, my, y')	Function (its, by, from)
Private V (feel, guess)	Formal DM (however,
People (guys, bro, girl)	despite)

Identify 'topics' in the corpus with topic modelling (Murakami et al., 2017)

Several methods trialled on the main corpus (collected after the pilot study) using Python (BERTopic & Top2Vec)

Top2Vec (Word2Vec + Doc2Vec) seems to be more interpretable

- Can include 50 words in topic
- Meaningful topics
- E.g. self-improvement, songs, European football, various specific video games, Christianity, cars, anime, junk food, pranks, Star Wars…

Corpus compilation

Search terms

(pilot) BNC top 200 includes: *government, system, house, life, local, man, Mr…*

➡ Could influence the content

Stop words

Function words & highly frequent content words, little semantic weight (Jurasky & Martin, 2023)

e.g. *after too by, because should has*

vlog + 3 stop words used in YouTube vlog corpus compilation (Egbert et al., 2022)

Optimal minimum text length

2,000 words+ ?

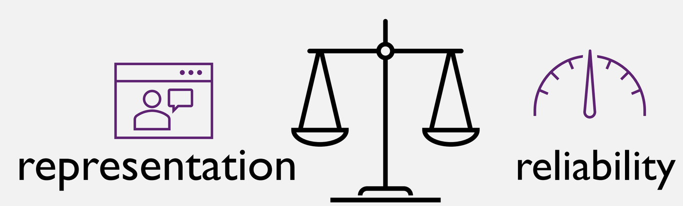
(Biber, 1990, Thompson et al., 2017)

Too strict for YouTube?

- Many videos < 2,000 words

MD Analysis of conversation (Biber, 2004) = 200+ words

'because of the difficulties in obtaining reliable rates of occurrence for linguistic features in shorter texts' (p. 18)



200 words +

CEFR Level (work in progress)

Machine Learning

Linguistic features TAALES, TAALED, TAACO, TAASC (Kyle & Crossley)

- Random Forests
- Ordinal Logistic Regression
- Support Vector Machines

So far low accuracy on CEFR labelled listening corpus ≈ 60%

↑ C2 C1 B2 B1 A2 A1 ↑

LLMs

BERT v. accurate (97%) at classifying learner writing with 50k-100k texts in training data (Schmalz & Brutti, 2021)

➡ fine-tune for listening?

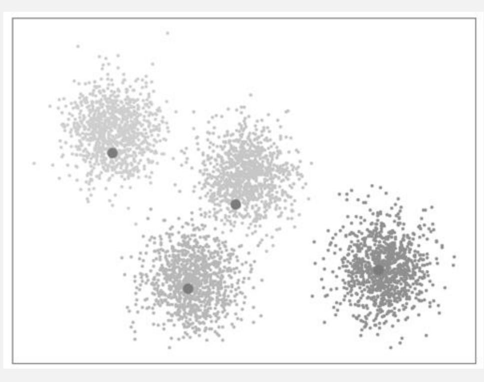
CEFR descriptors seem to be too difficult for LLMs to classify accurately

➡ LLM prompts + training data?

Cluster Analysis Method

K-means?

- K chosen by researcher
- Hierarchical levels not analysed in this research



https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_plusplus.html

Poster / References

@coopersensei
cooper@rikkyo.ac.jp