Christopher Robert Cooper

Rikkyo University, Tokyo, Japan

cooper@rikkyo.ac.jp

A corpus of short YouTube news videos to inform course design and materials development in an EFL university setting in Japan

Bio data

Chris Cooper currently works as an Adjunct Lecturer in English at the Center for Foreign Language Education and Research at Rikkyo University in Tokyo, Japan. He is also a PhD student at Tokyo University of Foreign Studies. He has been teaching English as a foreign language in Japan since 2010 and his current research interests include corpus linguistics, natural language processing and using YouTube videos for language learning.

Abstract

The aim of the current study was to inform the course design of an elective English news listening course at a private university in Japan. YouTube channels from twelve countries across Asia, Africa, Europe and North America were manually selected by the researcher as a source of in class materials and for course participants to view outside the classroom. To create materials for language-focused learning and assess the vocabulary demands of the videos, transcripts from the channels were extracted using the YouTube Data API and Python, and a corpus of 8,286 video transcripts uploaded in 2021 was randomly sampled to represent the channels. The transcripts were cleaned, and frequency lists were created for adjectives, nouns, and verbs at CEFR B1 level and above. In addition, proper noun and multi-word unit frequency lists were created. An online concordancer was created using the open-source tool *ShinyConc*, so the learners could investigate the usage of the words in the frequency lists by themselves. A Python script was written to assess the lexical coverage of the videos using the CEFR-J wordlist, and the results suggested that learners may need to be at the CEFR B2 level or above to comfortably comprehend short news YouTube videos. Suggestions for future research are made in the paper, and Python code and supplementary data are available at the author's GitHub page (https://github.com/cooperchris17/yt_short_news).

Conference paper

Introduction

For listening courses, Nation and Yamamoto (2012) suggested adapting Nation's (2007) Four Strands theory in the following way:

- 50% of the time allocated to meaning-focused input (including watching videos, listening to stories or taking part in discussions)
- 25% of the time on fluency-development activities
- 25% on language-focused activities

A corpus representing the texts used in a listening course could be a useful tool to evaluate if materials are suitable for meaning-focused input and create materials for the fluency and language-focused strands.

Corpus-based materials

According to Chambers (2010), data-driven learning (DDL) involves learners interacting with concordance lines prepared by the teacher or by using concordance software directly to study the patterns of language and work out how words and phrases are used by themselves. One of the pioneers of this approach, Tim Johns, believed that the computer should be an informant, with the L2 user asking questions, noticing patterns and making their own generalisations (Johns, 1991). He also pointed out that the most important DDL tool is the concordancer. In the intervening years, DDL has become an increasingly popular approach. In Boulton and Cobb's (2017) meta-analysis of DDL research, the data from control/experimental group comparisons (d = 0.95) and pre/posttest designs (d = 1.50) indicated that DDL has a large effect on learning. In the EFL context of Japan, Mizumoto and Chujo's (2015) meta-analysis also showed positive effects. The studies were grouped by the gains measured in the original study, and it was shown that studies that used a vocabulary test had a large effect size, papers that tested by category, for example parts of speech or basic grammar, or at the phrase level, including TOEIC type items, had a medium effect size. Studies that used a proficiency test, specifically the TOEIC Bridge, had a small effect size.

To do DDL, a corpus and corpus-tools are needed. According to Anthony (2020), programming is useful to understand the limitations of corpus-based tools, and advantageous if researchers want to create their own corpus and deal with the inevitable noisy data. A further advantage of conducting corpus-based research using a programming language, such as Python or R is the replicability of the research methods. On the subject of quantitative SLA research, Gass et al. (2021) call for materials and data to be made available on repositories for transparency and reproducibility, in the spirit of open science.

Lexical coverage

In reading research, Hu and Nation (2000) estimated that most readers would need to know 98% of the vocabulary in a text to comfortably comprehend it. In listening, Van Zeeland and Schmitt (2013) suggested a lower threshold of 95% in most circumstances, or 90% for some L2 users. As a caveat, they noted that 98% is probably more appropriate if high comprehension is necessary. Research on lexical coverage and the comprehension of videos is still limited, but some research has suggested that 90% coverage may be enough to understand a TV program, such as a documentary, without assistance (Durbahn et al., 2020). If the goal is purely to understand the main points of a video, which also includes visual information, then knowing somewhere between 90-95% of the vocabulary could be an appropriate target.

Another question to consider is what word list to use to assess the vocabulary of texts. There has been some debate recently (e.g., McLean, 2018; Stoeckel et al., 2020; Webb, 2021) in the vocabulary research field about what word counting unit to use: types, lemmas, flemmas, or word families. According to Gablasova and Brezina (2021), there is no real lemma debate and 'the advantage of using lemma as a unit is that it is more precise and requires fewer assumptions about the (morphological and semantic) knowledge on the part of learners than other units' (p. 959). The CEFR-J wordlist (Tono, 2020) is lemma-based and was initially based on a textbook corpus of major textbooks used in compulsory education in China, Korea and Taiwan. Then the wordlists were compared with the English Vocabulary Profile (https://www.englishprofile.org/wordlists), and extra words were added. One benefit of the CEFR-J wordlist over other wordlists that are typically split into 1,000-word bands is the interpretability of the CEFR level, as it can

be matched to the proficiency of learners, who are often put into classes by proficiency level.

Video in language learning

According to Mayer et al. (2014) research, audio lectures were more comprehensible for ESL university learners in the U.S. when they were accompanied by video supporting the meaning. Vanderplank (2016) has argued the case that the use of L2 captions generally makes TV and movies more comprehensible for L2 users than viewing without captions. In addition, Baraowska (2020) showed that both L1 and L2 subtitles reduced cognitive load and increased comprehension when watching a 12-minute TV drama clip. In a longitudinal study, Muñoz et al.'s (2021) results suggested that viewing with L1 subtitles resulted in higher gains in understanding the meaning of vocabulary, but L2 captions led to gains in recalling the written form. Some researchers, such as Majuddin et al. (2021) have taken a more specific approach, with their investigation of multiword expression acquisition showing that two viewings was beneficial compared to one, and viewing with captions resulted in higher gains than no captions.

The advantage of using YouTube videos in the classroom is their flexibility. Not only are they easily accessed on any device, but many videos have the option to choose L1 or L2 captions, transcripts are available, and there is a large selection of short videos, increasing the chance of repeated viewings. In addition a great deal of videos feature L2 users, providing a model of English as a Lingua Franca for viewers.

Research Questions:

1. What materials can be produced from a corpus of YouTube videos to promote language focused learning?
2. At what CEFR level do learners have sufficient vocabulary knowledge to reach the 90-95% lexical coverage threshold?

Methodology

Corpus design, compilation, and sampling

The corpus was designed to represent short YouTube news videos for EFL learners at the intermediate B1 level and above. It was designed to be used in an elective course at a private university in Tokyo, Japan. The pedagogical goals of the course are to develop the participants' ability to comprehend and discuss domestic and international English language news broadcasts and learn about topical issues.

YouTube channels were manually selected by the author with the following criteria. The channels should include English language news videos, they should be of less than 4 minutes in length, include closed captions and transcripts, and several countries should be represented, with only one channel chosen for each country. When the channels had been selected, metadata was extracted for 500 videos uploaded in 2021 from each channel using the YouTube Data API (2022). Then the youtube-transcript-api (Depoix, 2021) was used to extract auto-generated transcripts from each of the channels. It was necessary to specify auto-generated transcripts to avoid downloading transcripts for videos with no English sound. At this point, channels that had no or few transcripts available were excluded from selection and a list of 12 channels was chosen to collect a larger dataset. In the second round of data collection, the metadata for all short videos uploaded in 2021 was obtained for each channel by making a separate request to the API for each calendar month of the year for each channel. This metadata was used to download a plain text file version of all available transcripts for these videos. The specific YouTube channels and number of transcripts obtained are summarised in Table 1.

Table 1. *Data collection summary*

Channel	Country	2021 Short Videos	Corpus Sample
ABC News	U.S.A.	1,990	720
Al Jazeera English	Qatar	2,507	720
Arirang News	South Korea	2,842	720
BBC News	U.K.	982	720
CBC News	Canada	667	667
CGTN	China	2,024	720
CNA	Singapore	2,515	720
DW News	Germany	419	419
Nippon TV News 24	Japan	876	720
TVC News	Nigeria	2,855	720
WION	India	2,465	720
i24NEWS English	Israel	1,004	720
	Total	21,146	8286

To more evenly represent each channel and reduce the size of the dataset, a random sample of each channel was selected using the Pandas (McKinney et al., 2010) library in Python. As shown in Table 1, for each channel, 720 videos were randomly selected. This meant that most channels were represented by the same number of videos. As only 667 transcripts were available for CBC News and 419 for DW News, all available transcripts were used for those channels. Due to the sampling, the corpus size was reduced from around 7.2 million tokens to around 2.9 million tokens.

Corpus cleaning

A sample of transcripts were read and reviewed and features that were not representative of speech were removed from the plain text transcripts using Python. The transcripts were all lower case with no sentence punctuation. It was found that the only occurrences of periods after a word were those occurring after numbers, these were deleted from the text files. The following three phrases that indicate sounds other than speech; *[Music]*, *[Laughter], [Applause]* were deleted. It was decided that transcripts containing less than 50 tokens should be deleted as manual checking of those videos indicated that the majority were largely videos that were mainly visuals with subtitles or were advertisements for the news channels. Some words are automatically censored by YouTube and are displayed as *[__]* in the transcripts. Censored words only appeared in ten videos, upon manual checking, any words that were mis-transcribed were edited in the text files. Whether the words should be censored is controversial because the audio is not censored, so hard of hearing viewers cannot access the same information. However, all other censored words were not amended in alignment with YouTube's censorship policy.

The final point that was noticed during the cleaning process was a range of different words co-occurring with *-19* that clearly represented *covid-19*. As this was an important word in the news in 2021, collocations of *19* and *-19* were searched for and a total of 204 patterns were replaced with *covid-19*. Some words were clearly non-words, such as *kovit-19*, in other cases, the word was mis-transcribed as an actual word, such as *coffee 19*. After further manual checking, 69 words were replaced with *covid*, which is also regularly used as a standalone noun without the 19, and 46 other cases were also amended, such as *yukovit 19* to *new covid-19*.

Frequency list construction

To answer research question one, Part-of-speech(POS)-tagged frequency lists were constructed to allow participants in the course the opportunity to learn vocabulary relevant to the genre of news videos at their level. Due to the issue of tagging proper

nouns caused by the lower-case nature of the transcripts, proper nouns were investigated first. After attempting to tag the texts using the 'off-the-shelf' *NLTK* tagger, which uses the Penn Treebank tagset (https://www.nltk.org/api/nltk.tag.html; Bird et al., 2009), other methods were investigated, using *spaCy* (https://spacy.io/) and the *roberta-large-ner-model*, which is a named-entity recognition model available at Hugging Face (https://huggingface.co/Jean-Baptiste/roberta-large-ner-english). The *roberta-large-ner-model* is described as working well with lower case entities and this seemed to be the case when testing several models from the Hugging Face website (https://huggingface.co/) on a small number of texts.

Table 2. *Summary of the most frequent proper nouns using three tagging methods*

NLTK		spaCy		Hugging Face	
word	frequency	word	frequency	word	frequency
xi	141	china	3414	china	2763
xinjiang	129	u.s.	3306	u.s.	2712
south	44	israel	2321	covid-19	2152
october	37	news	2168	israel	1968
joe	32	covid-19	2151	taliban	1223
taiwan	27	president	1927	japan	1131
zealand	26	united	1793	united states	1112
november	23	abc	1682	india	1105
kamala	22	korea	1644	chinese	1077
khan	17	south	1614	afghanistan	981

As can be seen in Table 2, the *NLTK* tagger clearly did not tag proper nouns correctly and there is a difference between the frequencies tagged by *spaCy* and the *roberta-large-ner-model*. This is because *spaCy* tags the individual words in multi-word items, such as *United States*, as individual words unless the model is retrained. However, many entities identified by the *Hugging Face* model did contain several words. To use the most frequent proper noun as an example, the only two words tagged by *spaCy* containing *china* were *china* and *chinatown*. With the Hugging Face model, more than 200 unique entities were identified containing the word *china*. Some of these should probably not be tagged as unique entities, such as *china china* and *china japan.* However, the aim of the proper nouns materials was for learners to compare frequent proper nouns between countries and have a long list of proper nouns that could be used by the instructor to prepare listening activities to identify lesser-known proper nouns. Therefore, Hugging Face was used to identify frequent entities per channel, as entities such as *united states* seemed more logical to present to learners than *united* as a standalone proper noun.

Due to the proper noun tagging issue, spaCy was used to prepare frequency lists of adjectives, verbs and nouns. Knowledge of the usage of A1 and A2 level words is crucial, due to the large amount of any text covered by these words. However, for texts that have not been specifically created for language learners, B1 and B2 level words are likely to be essential for comprehension. This being the case and due to the target level of the learners (B1+), any words in the CEFR-J wordlist at the A1 or A2 level were deleted from the frequency lists. Horizontal bar graphs displaying normalised frequencies (per million words) and the percentage of documents containing the word were produced for the top 100 most frequent words at the B1 level and above for each part of speech.

Finally multi-word item lists were produced using *AntGram* (Anthony, 2021), which is software designed to produce lists of frequently occurring words, or *n*-grams. The *n* in *n*-gram represents the number of words, for example, *a lot of* is a 3-gram and *for the first time* is a 4-gram. The following parameters were used in the software:

- 3-grams to 5-grams

- Minimum frequency of 20 and minimum document frequency of 20
- Numbers replaced by #
- No open slots
- Top 1000 sorted by frequency

Multi-word item lists are usually manually edited by multiple researchers with set criteria (e.g. Martinez & Schmitt, 2012). In the present study, multi-word items were deleted by the author from the list if:

- They were used in only one channel (e.g., 'the ABC News')
- N-grams that were actually 2 words (e.g., 'the country's')
- Parts (e.g., 'the same time' and 'at the same' were deleted, 'at the same time' was not)
- all 3-grams that included apostrophes (e.g., 'we're seeing', 'we've seen')
- Any n-gram containing 2 numbers and one word (e.g., the # #)
- N-grams with 'the [noun] and' (e.g., 'the pandemic and', 'the world and')
- Any n-grams that were extracted because there was a reoccurring segment in many videos, for example most ABC News videos ended like this:

hi everyone george stephanopoulos here thanks for checking out the abc news youtube channel if you'd like to get more videos show highlights and watch live event coverage click on the right over here to subscribe to our channel and don't forget to download the abc news app for breaking news alerts thanks for watching

The list was edited by the sole-researcher, meaning there was an element of subjectivity. Also, in contrast to the POS-tagged frequency lists, the multi-word unit list was not filtered by CEFR level.

Online concordancer for data-driven learning

To allow the learners to explore the language in the frequency lists and other words and phrases that they found difficult or interesting, a concordancer was prepared. The open-source tool *ShinyConc* (http://shinyconc.de/) was chosen because it is fairly easy to set up and allows a corpus to be uploaded, which can be deployed online for free using https://www.shinyapps.io/. After it is deployed, the concordancer can be used in any web browser, including on mobile devices. More sophisticated tools may be available, but one benefit of *ShinyConc* is that learners do not need to download computer software and upload the corpus by themselves like they would with tools like *AntConc* (Anthony, 2022) or pay for a subscription as they might need to with tools like *Sketch Engine* (Kilgarriff et al., 2014; http://www.sketchengine.eu/). In addition, the link to the YouTube video can be added next to the concordance line in *ShinyConc*, so users can view the video containing the concordance line. However, it is not a hyperlink, meaning that it must be copied and pasted into a new browser window. When the concordancer was deployed online, it was necessary to reduce the size to 5,809 files, due to a file limit for Shiny Apps. This means that learners were exploring a sample of the corpus, and not the full 8,286 files.

Lexical coverage

The decision was made to use the CEFR-J wordlist to assess the vocabulary load of the videos. This was mainly because of the interpretability of the results, as the materials described in the current study were designed for learners at CEFR level B1 and above, this could be used as a benchmark to judge whether the materials were appropriate and how much support should be provided. The New Word Level Checker (NWLC) (Mizumoto, 2021) is an excellent resource for checking the lexical coverage of single texts for a number of wordlists including CEFR-J. However, as the corpus in the current study contained 8,286 texts, a Python script was written using spaCy to assess the vocabulary

demands of all the texts in batches. A list of tuples was produced from the CEFR-J wordlist, including each lemma, its part of speech, and CEFR level. The texts were tagged and lemmatised using spaCy and matched with the list of tuples to assign a CEFR level to each word in each text. Following how texts are profiled in NWLC, proper nouns and numbers were counted as known words first, and the cumulative lexical coverage was calculated at each CEFR level, to give an indication of the lexical difficulty of the texts, lemmas that were not in any of the lists were added to an *others* list. Extra lemmas were added to the list of tuples to match how spaCy tags some words. This was done by checking how words were tagged by New Word Level Checker and in consultation with the English Profile (https://www.englishprofile.org/). A full list of the additions with justifications, along with a Jupyter notebook containing the code to calculate the lexical coverage for multiple texts is available on an online repository (see below). The performance of the script written for this project seems to be close to NWLC. However, one area that needs improvement is how proper nouns are tagged. NWLC seems to be more accurate at tagging proper nouns.

Additional materials and data

The main steps of the methodology have been described here. However, more detailed information including Python code, Jupyter Notebooks, and additional results are available on the following GitHub repository: https://github.com/cooperchris17/yt_short_news.

Results and discussion

Frequency Lists

An example of the part of speech frequency list graphs, and multi-word units list created as classroom materials can be seen in Figure 1 and Table 3. Some of the words in the example in Figure 1 would be classed as A1 words in a different part-of-speech. For example, the noun *test* is an A1 level word, but the verb *test* is a B2 level word. The aim of the lists is for learners to search for words or phrases they are interested in using the concordancer and notice their own patterns of how the word is used in context. As the corpus is genre specific, this may help them understand the language in other news videos that they watch.

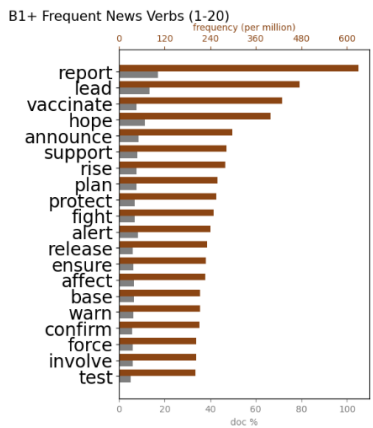

Figure 1. *Example of part-of-speech wordlist*

Table 3. *Top 15 frequent multi-word units (not filtered by CEFR level)*

Original Rank	Multi-word unit	Frequency	Document frequency
2	a lot of	1589	1077
3	more than #	1551	1207
4	one of the	1458	1204
9	going to be	820	620
10	the end of	771	669
11	as well as	733	634
12	# percent of	725	576
14	be able to	708	601
177	this is the	654	586
178	some of the	652	561
179	# year old	643	510
181	in the country	618	519
202	the number of	581	463
203	the first time	567	510
209	re going to	536	433

For proper nouns, it might be noticed by learners that most countries use proper nouns related to their country or neighbouring countries—for example, *Canada, Ontario* and *Vancouver* in the Canadian news channel shown in Figure 2. Many of the lists also included the major economic countries, the U.S. and China. Most of the frequent proper nouns would probably be known by learners, but some may not be, such as *RCMP* (Royal Canadian Mounted Police). If a learner was interested in news from a particular country, it could be worth becoming familiar with proper nouns such as this.

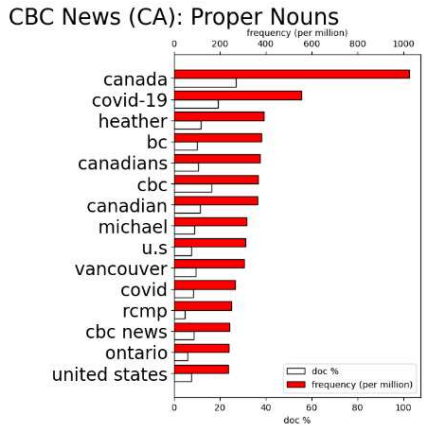

Figure 2. *Example of proper noun frequency list*

Online concordancer

A screenshot of the concordancer created using *ShinyConc* can be seen in Figure 3. The concordancer (https://coopersensei.shinyapps.io/yt_news_shinyconc/) is very functional and its greatest benefit is that users can get started with it immediately on any device without the need to download software or upload a corpus. The concordance lines can be ordered by the word to the right of the node word, however, this is not the case for words to the left of the node word. If the word 'left' is clicked, the concordance lines are ordered by the first word in that column, which limits pattern searching to a certain extent. The display of metadata and YouTube links is also very useful. However, as the purpose is to improve listening ability, it would be better if a tool was designed to link to an embedded YouTube video at the point of the concordance line. This could probably be done using JSON files extracted by the *youtube-transcript-api*, which include timestamps for each line. At least two online tools are already available that search YouTube transcripts and display embedded videos. However, either the register of the videos is unclear (https://youglish.com/) or they are fixed to the genre of TED Talks (https://yohasebe.com/tcse/; Hasebe, 2015). A tool loaded with a range of specified genres, or a tool where users could upload their own transcripts would be useful additions to what is already available.

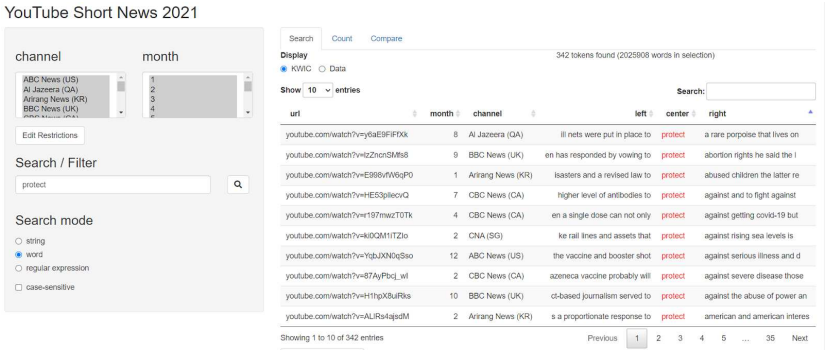

Figure 3. *Screenshot of ShinyConc*

Lexical Coverage

The lexical coverage of the videos including proper nouns and numbers at the CEFR A2, B1 and B2 levels are shown in Figures 6, 7 and 8. The line in the centre of each boxplot represents the median lexical coverage score for the videos in the specified channel. The boxes represent 50% of the data points between the first and third quartiles, the dots to the left of each boxplot show all of the data points, so the dispersion of the results can be clearly seen. At the A2 level, the boxes containing the data points 25% above and 25% below the median fall between around 75% and 85% lexical coverage. At the B1 level these figures rise to just below 85% and around 92%, and at the B2 level between 89% and 95%. Based on these results, it seems that L2 users' level should be at least B2 before they can freely and comfortably watch short news YouTube videos. However, as news videos are often supported by visuals that are likely to aid comprehension and often have the option to display captions, they could still be suitable for B1 level L2 users.

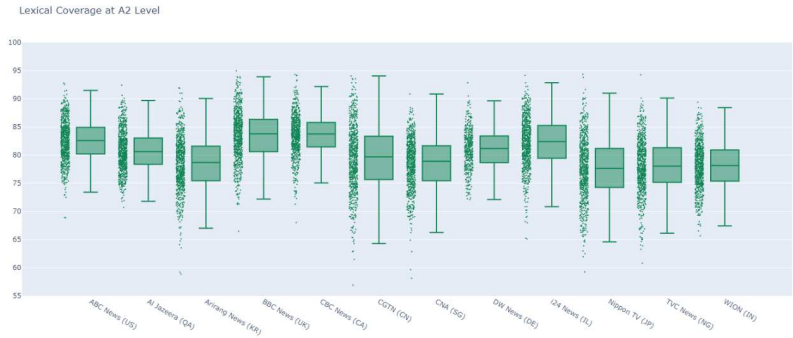

Figure 6. *Boxplots showing lexical coverage at A2 level*

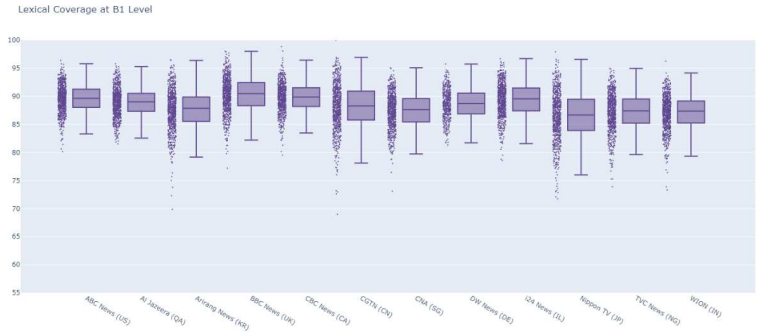

Figure 7. *Boxplots showing lexical coverage at B1 level*

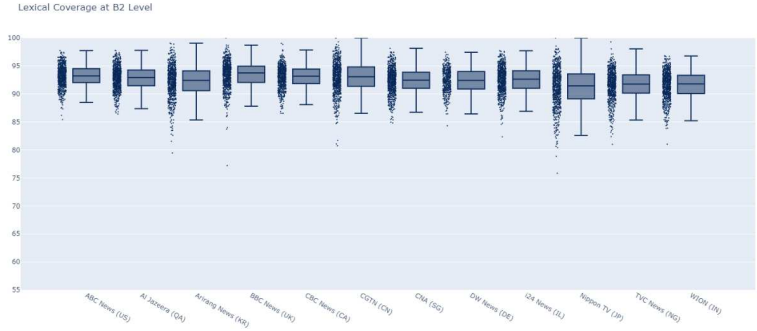

Figure 8. *Boxplots showing lexical coverage at B2 level*

If we compare the lexical coverage of the results in this study with the CEFR guidelines related to watching TV shown in Table 4, we can see that the CEFR descriptors suggest that learners can deal with news to a certain extent from the A2 level. It could be the case that setting reasonable learner expectations when listening and teaching listening strategies (e.g., Vandergrift et al., 2006) is equally as important as checking vocabulary. In addition, many other factors have been shown to affect listening comprehension, such

as redundancy, schema, concreteness and orality (Bloomfield et al., 2010). For example, the channel with seemingly the lowest lexical coverage scores, Nippon TV, may be more comprehensible to L2 users from Japan, because they may be familiar with much of the content.

Table 4. *TV news-related CEFR can-do descriptors*

CEFR Level	Can-do descriptor
A2	Can identify the main point of TV news items reporting events, accidents, etc. where the visuals support the commentary.
B1	Can understand a large part of many TV programmes on topics of personal interest such as interviews, short lectures and news reports when the delivery is relatively slow and clear.
B2	Can understand most TV news and current affairs programmes.

Council of Europe (2020, p. 53)

Whether proper nouns and numbers should be counted as known words is also a point of discussion for listening texts. In written texts, proper nouns are capitalised, which may be easier for L2 users to recognise during reading. This is not the case with listening texts and classifying proper nouns as known words has been called into question by some researchers. Kobeleva (2012), who compared groups listening to texts with known and unknown proper nouns, concluded that proper nouns increase the difficulty of texts, particularly if the text contains more than 4-5% proper nouns. This seems to be true for many of the videos in the current study (see Figure 9). The teaching of strategies for recognising proper nouns in listening texts could be an activity worth spending time on in the classroom and a potential topic for future research in this area.

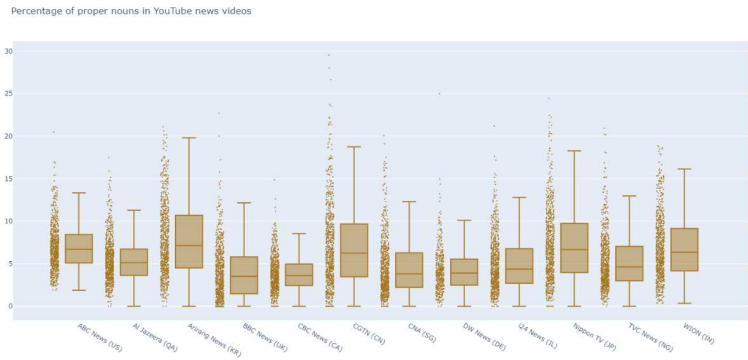

Figure 9. *Percentage of proper nouns in the corpus by channel*

A final point worth mentioning related to lexical coverage is the words that were not tagged as proper nouns, numbers, or A1-B2 words. A total of 5,912 lemma types were in this category, so clearly it would not be advisable to draw attention to many of these words in a classroom situation. However, some of the frequent words that may be considered specialized vocabulary based on current trends and usage could be introduced to learners, such as Covid-19 related words (e.g., *variant*, *vaccinate*, *vaccination* and *booster*), along with register-specific words, such as interjections (e.g., *uh*, *well*, *um*, and *like*). Identifying irregularly high frequency vocabulary is a benefit of reviewing the lexical

coverage of a corpus of texts that are specifically related to the target genre of class content, as it is not possible to assign a CEFR level to all vocabulary, especially when the high frequency may be related to a specific time-period or genre.

Conclusion and future suggestions

It has been shown in the current study that it is possible to extract transcripts from YouTube that are highly related to classroom content and use the transcripts to create materials that can be used in the classroom. There were various obstacles when dealing with the YouTube data, such as the tagging of proper nouns and transcript cleaning. In addition, a script was written in Python to calculate the lexical coverage of a large number of texts using the lemma-based CEFR-J wordlist. For future projects, the lexical coverage calculation could be improved, specifically the way that proper nouns are tagged, or the inclusion of C1 and C2 level words to give a more accurate lexical profile. Triangulating the results with user ratings would also be beneficial to investigate how accurate the CEFR-J wordlist is at predicting the difficulty level of videos. It could be that L2 users at a lower CEFR level than the lexical coverage score could also comprehend news videos due to the visual support. In addition, a concordancer could be developed linking concordance lines directly to their timestamp in YouTube videos. These are points that the researcher would like to investigate in the future, and any interested parties are welcome to view the code and datasets on GitHub (https://github.com/cooperchris17/yt_short_news), and make comments or get in contact with any suggestions.

References

Anthony, L. (2020). Programming for corpus linguistics. In M. Pacquot & S. T. Gries (Eds.), *A practical handbook of corpus linguistics* (pp. 181-207). Springer.

Anthony, L. (2021). AntGram (Version 1.3.0) [Computer Software]. Tokyo, Japan: Waseda University. Available from https://www.laurenceanthony.net/software

Anthony, L. (2022). AntConc (Version 4.0.6) [Computer Software]. Tokyo, Japan: Waseda University. Available from https://www.laurenceanthony.net/software

Baranowska, K. (2020). Learning most with least effort: subtitles and cognitive load. *ELT Journal, 74*(2), 105-115. https://doi.org/10.1093/elt/ccz060

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python: Analyzing text with the natural language toolkit*. O'Reilly Media, Inc. Retrieved from https://www.nltk.org/book/

Bloomfield, A., Wayland, S. C., Rhoades, E., Blodgett, A., Linck, J., & Ross, S. (2010). *What makes listening difficult? Factors affecting second language listening comprehension.* University of Maryland. Retrieved from https://apps.dtic.mil/sti/pdfs/ADA550176.pdf

Boulton, A., & Cobb, T. (2017). Corpus use in language learning: A meta-analysis. *Language learning, 67*(2), 348-393. https://doi.org/10.1111/lang.12224

Chambers, A. (2010). What is data-driven learning? In A. O'Keefe, & M. McCarthy (Eds.) *The routledge handbook of corpus linguistics*. (pp.345-358). Routledge.

Council of Europe. (2020). *Common European framework of reference for ;anguages: Learning, teaching, assessment – Companion volume*. Council of Europe Publishing. Retrieved from www.coe.int/lang-cefr

Depoix (2021). youtube-transcript-api 0.4.2. Retrieved from https://pypi.org/project/youtube-transcript-api/

Durbahn, M., Rodgers, M., & Peters, E. (2020). The relationship between vocabulary and viewing comprehension. *System, 88*(102166), 1-13. https://doi.org/10.1016/j.system.2019.102166

Gablasova, D., & Brezina, V. (2021). Words that matter in L2 research and pedagogy: A corpus-linguistics perspective. *Studies in Second Language Acquisition, 43*(5), 958-961. https://doi.org/10.1017/S027226312100070X
Gass, S., Loewen, S., & Plonsky, L. (2021). Coming of age: The past, present, and future of quantitative SLA research. *Language Teaching, 54*(2), 245-258. https://doi.org/10.1017/S0261444819000430
Hasebe, Y. (2015). Design and implementation of an online corpus of presentation transcripts of TED talks. *Procedia-Social and Behavioral Sciences, 198*, 174-182. https://doi.org/10.1016/j.sbspro.2015.07.434
Hu, M. and P. Nation. 2000. Unknown vocabulary density and reading comprehension. *Reading in a Foreign Language, 13*(1), 403–430. Retrieved from https://scholarspace.manoa.hawaii.edu/bitstream/10125/66973/13_1_10125_66973_rfl131hsuehchao.pdf
Johns, T. (1991). Should you be persuaded: Two samples of data-driven learning materials. *English Language Research Journal, 4*, 1-16. Retrieved from https://lexically.net/wordsmith/corpus_linguistics_links/Tim%20Johns%20and%20DDL.pdf
Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., & Suchomel, V. (2014). The Sketch Engine: Ten years on. *Lexicography, 1*(1), 7-36. Retrieved from https://www.sketchengine.eu/wp-content/uploads/The_Sketch_Engine_2014.pdf
Kobeleva, P. P. (2012). Second language listening and unfamiliar proper names: Comprehension barrier?. *RELC journal, 43*(1), 83-98. https://doi.org/10.1177%2F0033688212440637
Majuddin, E., Siyanova-Chanturia, A., & Boers, F. (2021). Incidental acquisition of multiword expressions through audiovisual materials: The role of repetition and typographic enhancement. *Studies in Second Language Acquisition, 43*(5), 985-1008. https://doi.org/10.1017/S0272263121000036
Martinez, R., & Schmitt, N. (2012). A phrasal expressions list. *Applied Linguistics, 33*(3), 299-320. https://doi.org/10.1093/applin/ams010
Mayer, R. E., Lee, H., & Peebles, A. (2014). Multimedia learning in a second language: A cognitive load perspective. *Applied Cognitive Psychology, 28*(5), 653-660. https://doi.org/10.1002/acp.3050
McKinney, W. (2010). Data structures for statistical computing in python. *Proceedings of the 9th Python in Science Conference, 445*, 56–61. Retrieved from http://conference.scipy.org/proceedings/scipy2010/pdfs/mckinney.pdf
McLean, S. (2018). Evidence for the adoption of the flemma as an appropriate word counting unit. *Applied Linguistics, 39*(6), 823–845. https://doi.org/10.1093/applin/amw050
Mizumoto, A. (2021). New word level checker [Web application]. https://nwlc.pythonanywhere.com/
Mizumoto, A., & Chujo, K. (2015). A meta-analysis of data-driven learning approach in the Japanese EFL classroom. *English Corpus Studies 22*, 1-18.
Muñoz, C., Pujadas, G., & Pattemore, A. (2021). Audio-visual input for learning L2 vocabulary and grammatical constructions. *Second Language Research*. Advance online publication. https://doi.org/10.1177%2F02676583211015797
Nation, P. (2007). The four strands. *International Journal of Innovation in Language Learning and Teaching, 1*(1), 2-13. https://doi.org/10.2167/illt039.0
Nation, P., & Yamamoto, A. (2012). Applying the four strands. *International Journal of Innovation in English Language Teaching and Research, 1*(2), 167-181. Retrieved from https://openaccess.wgtn.ac.nz/articles/journal_contribution/Applying_the_four_strands/12552020/1/files/23373002.pdf
Stoeckel, T., Ishii, T., & Bennett, P. (2020). Is the lemma more appropriate than the flemma as a word counting unit?. *Applied Linguistics, 41*(4), 601-606. https://doi.org/10.1093/applin/amy059

Tono, Y. (2020). The CEFR-J wordlist version 1.6. Compiled by Yukio Tono, Tokyo University of Foreign Studies. Retrieved from http://www.cefr-j.org/download.html

van Zeeland, H., & Schmitt, N. (2013). Lexical coverage in L1 and L2 listening comprehension: The same or different from reading comprehension?. *Applied Linguistics, 34*(4), 457-479. https://doi.org/10.1093/applin/ams074

Vandergrift, L., Goh, C. C., Mareschal, C. J., & Tafaghodtari, M. H. (2006). The metacognitive awareness listening questionnaire: Development and validation. *Language learning, 56*(3), 431-462. https://doi.org/10.1111/j.1467-9922.2006.00373.x

Vanderplank, R. (2016). *Captioned media in foreign language learning and teaching: Subtitles for the deaf and hard-of-hearing as tools for language learning.* Palgrave Macmillan.

Webb, S. (2021). The lemma dilemma: How should words be operationalized in research and pedagogy? *Studies in Second Language Acquisition, 43*(5), 941-949. https://doi.org/10.1017/S0272263121000784

YouTube Data API (2022) *YouTube data API*. Retrieved from https://developers.google.com/youtube/v3