Beyond Film Subtitles: Is YouTube the Best Approximation of Spoken Vocabulary?

Adam Nohejl **Frederikus Hudi** **Eunike Andriani Kardinata**
Shintaro Ozaki **Maria Angelica Riera Machin** **Hongyu Sun**
Justin Vasselli **Taro Watanabe**
Nara Institute of Science and Technology
{nohejl.adam.mt3, frederikus.hudi.fe7, eunike.kardinata.ef9}@is.naist.jp
{ozaki.shintaro.ou6, riera_machin.maria.rn9, sun.hongyu.sg6}@naist.ac.jp
{vasselli.justin_ray.vk4, taro}@is.naist.jp

Abstract

Word frequency is a key variable in psycholinguistics, useful for modeling human familiarity with words even in the era of large language models (LLMs). Frequency in film subtitles has proved to be a particularly good approximation of everyday language exposure. For many languages, however, film subtitles are not easily available, or are overwhelmingly translated from English. We demonstrate that frequencies extracted from carefully processed YouTube subtitles provide an approximation comparable to, and often better than, the best currently available resources. Moreover, they are available for languages for which a high-quality subtitle or speech corpus does not exist. We use YouTube subtitles to construct frequency norms for five diverse languages, Chinese, English, Indonesian, Japanese, and Spanish, and evaluate their correlation with lexical decision time, word familiarity, and lexical complexity. In addition to being strongly correlated with two psycholinguistic variables, a simple linear regression on the new frequencies achieves a new high score on a lexical complexity prediction task in English and Japanese, surpassing both models trained on film subtitle frequencies and the LLM GPT-4. Our code, the frequency lists, fastText word embeddings, and statistical language models are freely available online.[1]

1 Introduction

Word frequency is crucial for psycholinguistic research, as well as for assistive or educational applications involving production or comprehension of words.

Psycholinguistic analyses of the relative strength of variables affecting lexical processing, such as word frequency and age of acquisition (e.g. Garlock et al., 2001), hinge on accurate data for these variables. As word frequency and age of acquisition are correlated with each other, having less representative frequency data can easily change the result of such an analysis.

Traditionally, written corpora have been used for estimates of word frequency, with Kučera and Francis (1967) frequency norms long dominating psycholinguistic research of English. While the size of written language corpora has grown over time, speech corpora are still costly to develop and comparably limited in extent. When user-generated text became available on a large scale, it was possible to approximate everyday language exposure by collecting English text from Usenet newsgroups (Burgess and Livesay, 1998), and later French film and TV subtitles (New et al., 2007). Subtitle-based norms for US English, SUBTLEX-US, (Brysbaert and New, 2009) were found more predictive of lexical decision times (LDT) than frequencies based on traditional written corpora or the Usenet-based corpus.

These pioneering studies on subtitle corpora spurred the creation of film and TV subtitle-based frequency norms, dubbed SUBTLEX, for other languages such as Spanish (Cuetos et al., 2011), or British English (van Heuven et al., 2014). SUBTLEX frequencies for two Asian languages, Chinese (Cai and Brysbaert, 2010) and Vietnamese (Pham et al., 2019) were compiled as well. Most of the research, however, has remained focused on languages spoken in WEIRD[2] countries.

Subtitle frequencies are currently being used in a variety of practical tasks which need to model familiarity with words, such as lexical simplification or readability assessment. Despite their practical utility, film subtitle corpora are far from perfect approximations of spoken language. A large part of the non-English SUBTLEX corpora comes from translations of English-language movies. For instance, SUBTLEX-ESP (Cuetos et al., 2011) consists of less than 3% original Spanish subtitles, while more than 92% are translations from English. In Vietnamese, Pham et al. (2019) did not find subtitles more predictive of LDT than a written cor-

[1] https://github.com/naist-nlp/tubelex

[2] Western, Educated, Industrial, Rich, and Democratic (WEIRD), an acronym coined by Henrich (2020).

Proceedings of the 31st International Conference on Computational Linguistics, pages 9566–9585
January 19–24, 2025. ©2025 Association for Computational Linguistics

pus, citing translation artifacts and cultural differences from predominantly American material as the likely causes. Moreover, the content presented in film dialogue is a very specific subset of spoken language. The speech is almost exclusively scripted and skewed to particular topics and vocabulary (Paetzold and Specia, 2016).

In this work, we build a corpus of untranslated YouTube video subtitles and evaluate the correlation of its frequencies with LDT, word familiarity, and lexical complexity, comparing them with frequencies based on available subtitle and speech corpora. We purposely target two languages spoken in WEIRD countries, English and Spanish, with a wealth of previous research to compare with, as well as three languages with diverse characteristics and amounts of resources available, Chinese, Japanese, and Indonesian.

As full corpus data cannot be published due to copyright, we release two basic language models based on the TUBELEX corpus for each language in addition to the frequency lists: a statistical language model (Heafield et al., 2013), which provides smoothed frequencies of word 1-grams to 5-grams, and fastText word embeddings (Bojanowski et al., 2017) to enable modeling of semantic similarity or analogy, as well as representation of words in downstream application. FastText extends the Word2vec model (Mikolov et al., 2013). Preprocessing details and hyperparameters are provided in Appendix A, model sizes in Appendix B, and evaluation of the embeddings in Appendix C.

2 Related Work

2.1 Subtitle Corpora

New et al. (2007) collected French film subtitles from the web to create a subtitle corpus. A similar procedure was then used for SUBTLEX-US (Brysbaert and New, 2009), and other SUBTLEX corpora, in some cases adding duplicate removal, e.g. for SUBTLEX-ESP (Cuetos et al., 2011), or various forms of cleaning. While most of the film subtitle corpora are collected from the web (often the OpenSubtitles website[3]), the British SUBTLEX-UK (van Heuven et al., 2014) acquired television subtitles from the BBC broadcasts.

Francom et al. (2014) used film metadata to build a relatively small corpus of untranslated Spanish subtitles, ACTIV-ES, and released lists of its n-grams. Paetzold and Specia (2016) restricted movies and series to particular genres to build the SubIMDB corpus.

All of these corpora aim to approximate spoken language and most of them were evaluated against psycholinguistic data. Other subtitle corpora were built for different purposes:

OpenSubtitles2016 (Lison and Tiedemann, 2016) and its updated version OpenSubtitles2018 (Lison et al., 2018) are large-scale collections of parallel film and TV subtitles downloaded from the OpenSubtitles website. In addition to parallel text aligned via subtitle timing, word frequencies for individual languages were released as well. For some languages, such as Indonesian, the OpenSubtitles corpus is the only subtitle corpus available.

Takamichi et al. (2021) downloaded audio and subtitles from YouTube to create JTubeSpeech, a Japanese corpus for speech recognition and speaker verification. The corpus or derived data was not published, and the corpus was evaluated only on these two tasks.

2.2 Evaluation Methods and Applications

New et al. (2007) evaluated a French subtitle corpus using correlation with LDT to demonstrate that it reflects language exposure better than written corpora. The same approach was subsequently adopted by others for different languages. Paetzold and Specia (2016) additionally evaluated the English SubIMDB on four other psycholinguistic ratings including word familiarity.

Van Paridon and Thompson (2021) used the data from several OpenSubtitles2018 languages to train word embeddings and evaluated them on word analogy and psycholinguistic ratings. The study excluded Chinese, Japanese, and other languages that do not separate words with spaces.

Shardlow (2013) demonstrated that frequency in SUBTLEX-US outperforms frequency in written corpora in ranking lexical simplifications for native speakers. Subtitle frequencies have been widely applied to lexical simplification in various languages for native and non-native speakers, where suitable SUBTLEX corpora were available (e.g. Štajner et al., 2022). Meanwhile, lexical complexity modeling for other languages, such as Japanese (Nishihara and Kajiwara, 2020) or Indonesian (Wibowo et al., 2019), has had to rely on web-scraped corpora instead.

Subtitle frequencies have also been used in a number of other tasks broadly connected to text comprehension, assistive technologies, and language learning, e.g. text readability assessment in English (Chen and Meurers, 2016) and Italian (Okinina et al., 2020), modeling of the ortho-

[3] http://www.opensubtitles.org/

graphic neighborhood effect in English and Dutch (Tulkens et al., 2020), a cross-linguistic study of the mental lexicon in English, German, and Chinese (Tjuka, 2020), construction of a vocabulary list for Finnish language learners (Robertson et al., 2022), or evaluating and improving performance of LLMs in colloquial English (Sun et al., 2024).

3 Corpus Construction

We build the corpus using several stages of processing. Table 1 shows statistic of the process.

3.1 Subtitle Scraping

As there is no public index of YouTube videos, we use YouTube's search function to search for all Wikipedia article titles in a given language to discover videos, following Takamichi et al. (2021).

To avoid translated or machine-generated subtitles, we restrict videos to those with both audio and manual subtitles explicitly labeled as the target language. For Chinese videos, we did not find enough videos with labeled audio language, so we also accept videos with unlabeled audio. The resulting numbers of videos are listed as Found in Table 1. We sample 120,000 videos for each language, for which we download subtitles.

3.2 Cleaning and Duplicate Removal

We identify the language of each subtitle line using the compressed fastText language identification model[4] (Joulin et al., 2016a,b). We discard files containing less than 95% of the target language. From the remaining files, we remove both lines that do not contain any valid characters for the target language (e.g. Latin alphabet for English), and lines that are identified as a different language. Lastly, we discard any files less than three lines long. The resulting numbers of files (i.e. files not discarded during cleaning out of the sample of size 120,000) are listed as Cleaned in Table 1.

We consider files duplicate if the cosine similarity between their 1-gram TF-IDF vectors is 0.95 or higher. We remove duplicate files heuristically to achieve a state without any duplicate pair. The final numbers of unique files and tokens in them are listed as Unique and Tokens in Table 1.

3.3 Subtitle Processing

We parse the WebVTT[5] subtitle files, and remove formatting and repetition caused by subtitle scrolling. We preserve words censored by YouTube (replaced with "[__]")[6] and audio descriptions in brackets (e.g. English "[ominous music]", Japanese "【エンジン音】") as special tokens.

[4]https://fasttext.cc/docs/en/language-identification.html

[5]https://www.w3.org/TR/webvtt1/

Language	Videos			Tokens‡
	Found	Cleaned†	Unique†	
Chinese	5,848,257	10,172	10,146	17,865,686
English	4,748,327	105,976	105,752	170,750,870
Indonesian	5,265,240	34,818	34,684	34,903,381
Japanese	4,970,247	101,664	100,754	163,439,781
Spanish	3,840,068	107,166	106,676	169,188,689

Table 1: Corpus construction statistics. †Out of 120,000 downloaded subtitle files. ‡In default tokenization.

3.4 Masking Personal Information

We also use special tokens to replace sequences of digits (after tokenization) and anonymize email addresses, web addresses including those without an explicit protocol (e.g. `x.com/username`), and apparent social network handles starting with `@`. Our approach to anonymization is informed by the analysis by Subramani et al. (2023) and extends the approach of Soldaini et al. (2024) by masking web addresses and social network handles.

3.5 Tokenization and Frequency Lists

We provide frequency lists in multiple variants:

default English, Indonesian, and Spanish segmented using Stanza (Qi et al., 2020) `tokenize,mwt` pipeline; Japanese segmented using MeCab (Kudo et al., 2004) with UniDic 2.1.2 (Den et al., 2007, distributed as unidic-lite[7]); Chinese segmented using the jieba[8] segmenter 0.42.1.

base Base form of Japanese tokens, preserving original spelling, obtained from MeCab/UniDic (書字形基本形).

lemma English, Indonesian and Spanish lemmatized using Stanza `tokenize,mwt,lemma` pipeline; Japanese lemmatized using MeCab/Unidic (語彙素), i.e. words in the orthographically normalized base form.

regex English, Indonesian, and Spanish orthographic words, matching a Python regular expression for sequences of characters belonging to the `\w` (word) class, but not to the `\d` (digit) class.

All tokens are lower-cased and normalized to Unicode NFKC (Whistler, 2023). For each word,

[6]https://support.google.com/youtube/answer/6373554?hl=en

[7]https://pypi.org/project/unidic-lite/

[8]https://github.com/fxsjy/jieba

Status	Evaluation Result	Chinese	English	Indonesian	Japanese	Spanish
OK	Subtitles match speech in the target language	65%	91%	84%	84%	89%
Problematic	Subtitles match song in the target language	—	2%	1%	1%	2%
	Subtitles provide audio description (e.g. *phone rings*)	—	1%	1%	0%	0%
	Audio is neither speech or song	—	2%	0%	0%	2%
	Audio is synthesized speech	34%	3%	13%	14%	5%
	Audio language differs from the target language	1%	1%	1%	1%	2%
	Subtitle language differs from the target language	—	—	—	—	—

Table 2: Human evaluation of a sample size 300 for each language, consisting of 100 videos with 3 cues per video. Each of the 300 subtitle cues per language was assigned to exactly one of the evaluation results.

the frequency lists provide: **`count`** – number of occurrences, **`videos`** – number of videos containing the word, **`channels`** – number of channels the word occurrs in, **`count:`**C – number of occurrences of the word in the YouTube video category C.

4 Human Evaluation

We performed human evaluation to verify how representative the corpus is of the target languages, and spoken language in particular. We took a sample of size 300 for each language, consisting of 100 videos with 3 random subtitle time stamps for each. The videos were selected using stratified sampling by category and duration[9] for each language. Each sampled timestamp was examined and labeled by a CEFR C2-level non-native speaker for English and by native speakers for the other languages. We examined each example by playing the video and comparing the subtitle cue (the subtitle text displayed at the given timestamp) with the corresponding audio, extending the examined video segment as deemed necessary by the evaluator to categorize the example.

The evaluation results are shown in Table 2. To be evaluated as "OK", the subtitles must match human speech in audio and both must be in the target language. Otherwise, the instance is assigned to one of the "Problematic" evaluation results. While in principle problems not listed in the table may occur (e.g. the language matches, but the content of audio and subtitles differs) or the listed problems could co-occur, it did not happen in our sample.

Most importantly, among the 1,500 subtitle cues, we have not found a single one whose language would differ from the target language. We have, however, observed different dialects or varieties of each language, as well as apparent non-native speech, sometimes co-occurring in the same video. In the case of Chinese, 2 of the 100 videos were in Cantonese, with traditional Chinese subtitles, while the majority was in Mandarin Chinese. To better understand the composition of the Chinese subtitles, we also analyzed the script used in the whole corpus, and found that 66% videos use simplified Chinese, 33% videos use traditional Chinese, and 1% mix both.[10]

Most proportions of potentially problematic phenomena were relatively low (up to 2%), with the exception of synthesized speech, which ranged from 3% for English to 34% for Chinese. Synthesized speech with subtitles, or subtitles provided for scenes without speech, could effectively be written language, rather than spoken. We further discuss the implications in Section 6.2.

5 Extrinsic Evaluation

We evaluate multiple corpora on three tasks, LDT, word familiarity, and lexical complexity, comparing them with TUBELEX in default, base, lemma, and regex variants, described in Section 3.5.

Note that the datasets available for different languages generally have different characteristics (such as part of speech or word frequency distribution), so while our experiments allow comparison of different corpora for a particular task and language, they do not allow comparison across languages. For instance, we should not surmise that TUBELEX provides better data for Spanish than for English (or that lexical decision time is generally more strongly correlated with frequency in Spanish than in English) only because TUBELEX achieves (or because all corpora achieve) a higher correlation for the task in Spanish than English.

For each evaluated corpus, we report correlation measured by Pearson's correlation coefficient (PCC), and the statistical significance of its difference from the correlation with $\text{TUBELEX}_{\text{default}}$ on three levels: *** ($p < 0.001$), ** ($p < 0.01$), and * ($p < 0.05$). We compute the p-values using

[9]We divided the videos in three similarly large duration classes: [0, 3 min), [3 min, 10 min), and [10 min, ∞).

[10]We used the Hanzi Identifier package (`https://github.com/tsroten/hanzidentifier`). and considered subtitles mixed if they contained the non-majority script variant on at least 2 lines and at least 5% of lines.

Steiger's (1980) test for dependent correlations and consider $p \geq 0.05$ not statistically significant.

To demonstrate the practical usefulness of the TUBELEX frequencies, we also predict lexical complexity based on them, and compare our results with the top submissions of the BEA 2024 Multilingual Lexical Simplification Pipeline Shared Task (Shardlow et al., 2024).

5.1 Evaluated Corpora and Resources

We evaluate traditional speech corpora, subtitle corpora, and three additional resources:

Speech corpora: **BNC-Spoken**, the spoken subset of the British National Corpus (BNC Consortium, 2007); **CREA-Spoken**, the spoken subset of Corpus de Referencia del Español Actual (Real Academia Española, 2004); **CSJ**, the Corpus of Spontaneous Japanese (NINJAL, 2016); **HKUST/MTS** (Liu et al., 2006), a Mandarin telephone speech corpus. We could not find a large enough Indonesian speech corpus.

Subtitle corpora: **ACTIV-ES**; **EsPal** (Duchon et al., 2013); **LaboroTV1+2**, the combination of the two releases of LaboroTVSpeech (Ando and Fujihara, 2021); **OpenSubtitles**, the 2018 version; **SubIMDB**; **SUBTLEX** (US, CH, ESP); **SUBTLEX-UK**.

Other resources: **GINI**, a Twitter-based metric of words' dispersion in frequency of use by different people (Murayama et al., 2018; also see Appendix D); **Wikipedia**; **wordfreq**, a Python library (Speer, 2022) pooling frequency from multiple corpora. Wordfreq combines Wikipedia, Twitter and a subtitle corpus for each of the evaluated languages, as well as 4 more sources for English, Chinese, and Spanish, and 2 more for Japanese. The subtitle data used by wordfreq is OpenSubtitles2018, SUBTLEX-US and SUBTLEX-UK for English, SUBTLEX-CH for Chinese.

For each corpus, we provide technical details in Appendix E, and token and type counts in Appendix F.

5.2 Computing Frequency

To deal with words missing in a corpus, we use the formula with Laplace smoothing recommended by Brysbaert and Diependaele (2013) to compute frequency of a token w:

$$f(w) = \frac{\mathrm{count}(w) + 1}{\#\mathrm{tokens} + \#\mathrm{types}}, \quad (1)$$

where $\mathrm{count}(w)$ is the number of occurrences of the word w, #tokens is the total number of tokens in the corpus, and #types is the number of types in the corpus. As a result, even words missing in the corpus are assigned a non-zero frequency. In all experiments we use the logarithm of frequency. Appendix E provides details about specific corpora.

	Corpus	Chinese	English	Spanish
speech	BNC-Spoken	—	-0.548^{***}	—
	CREA-Spoken	—	—	-0.645^{***}
	HKUST/MTS	-0.465^{***}	—	—
film/TV subtitles	ACTIV-ES	—	—	-0.600^{***}
	EsPal	—	—	-0.807^{***}
	OpenSubtitles	-0.568	$\mathbf{-0.647}^{***}$	-0.811
	SubIMDB	—	-0.646^{***}	—
	SUBTLEX	$\mathbf{-0.587}^{**}$	-0.633^{**}	-0.763^{***}
	SUBTLEX-UK	—	-0.625	—
other	GINI	—	-0.420^{***}	—
	Wikipedia	-0.424^{***}	-0.540^{***}	-0.705^{***}
	wordfreq	-0.423^{***}	-0.632^{**}	-0.801^{***}
our	$\mathrm{TUBELEX}_{\mathrm{default}}$	-0.575	-0.627	-0.811
	$\mathrm{TUBELEX}_{\mathrm{regex}}$	—	-0.627	$\mathbf{-0.811}$
	$\mathrm{TUBELEX}_{\mathrm{lemma}}$	—	-0.624^{*}	-0.808^{***}

Table 3: LDT correlation. Strongest (lowest) correlations for each language are in bold.

5.3 Lexical Decision Time

Lexical decision is one of the basic psycholinguistic tasks, where subjects decide whether a sequence of characters is a valid word or not. The reaction time for each word is its lexical decision time (LDT).

We measure correlation (PCC) with mean LDT from three studies: the English Lexicon Project (Balota et al., 2007), restricted to lower-case words following the approach of Brysbaert and New (2009); the MELD-SCH database (Tsang et al., 2018) of simplified Chinese words; and SPALEX (Aguasvivas et al., 2018) for Spanish. For English and Chinese, we use the published mean LDT. SPALEX only provides raw participant data, which we process by removing times out of the range [200 ms, 2000 ms], as outlined by Aguasvivas et al. (2018), and computing the means.

The results in Table 3 show that in each of the three languages OpenSubtitles and TUBELEX are among the top similarly performing corpora. While OpenSubtitles achieve a stronger correlation for English, $\mathrm{TUBELEX}_{\mathrm{default}}$ achieves a stronger correlation for Chinese.

Other corpora either performed comparably, but only covered a single language (SubIMDB, EsPal), or underperformed noticeably in at least one language (SUBTLEX in Spanish, wordfreq in Chinese, Wikipedia in all languages, and GINI and ACTIV-ES in the one language they cover). Fur-

	Corpus	Chinese	English	Indonesian	Japanese	Spanish
speech	BNC-Spoken	—	0.741***	—	—	—
	CREA-Spoken	—	—	—	—	0.535
	CSJ	—	—	—	0.523***	—
	HKUST/MTS	0.414***	—	—	—	—
film/TV subtitles	ACTIV-ES	—	—	—	—	0.526
	EsPal	—	—	—	—	0.428***
	LaboroTV1+2	—	—	—	0.565***	—
	OpenSubtitles	0.444***	0.776	0.582***	0.314***	**0.553**
	SubIMDB	—	**0.781**	—	—	—
	SUBTLEX	0.505	0.773	—	—	0.538
	SUBTLEX-UK	—	0.779	—	—	—
other	GINI	—	0.664***	—	0.633***	—
	Wikipedia	0.334***	0.661***	0.455***	0.466***	0.329***
	wordfreq	0.242***	0.771**	**0.632**	0.522***	0.495***
our	TUBELEX$_{\text{default}}$	**0.506**	0.777	0.625	0.624	0.547
	TUBELEX$_{\text{regex}}$	—	0.777	0.617**	—	0.545
	TUBELEX$_{\text{base}}$	—	—	—	**0.641*****	—
	TUBELEX$_{\text{lemma}}$	—	0.774	0.618	0.637***	0.551

Table 4: Word familiarity correlation. Strongest (highest) correlations for each language are in bold.

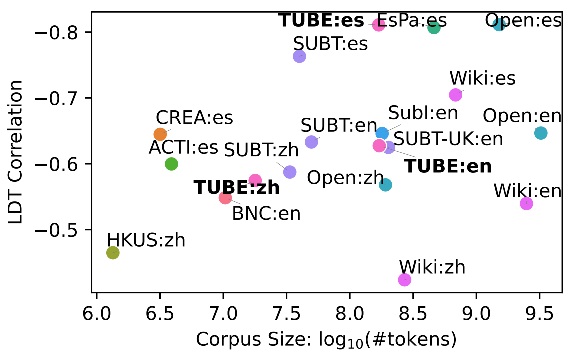

Figure 1: LDT correlation and corpus size. Labeled "*corpus abbr.*:*lang. code*", "TUBE" is TUBELEX$_{\text{default}}$.

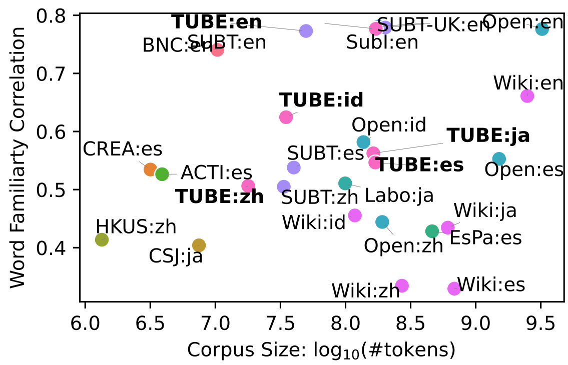

Figure 2: Word familiarity correlation and corpus size. Labeled "*corpus abbr.*:*lang. code*", "TUBE" is TUBELEX$_{\text{default}}$, not showing outlier "Open:ja".

thermore, Figure 1 shows that TUBELEX and SUBTLEX perform remarkably well relative to their size.

5.4 Word Familiarity

Word familiarity is a subjective rating of exposure to a given word. Among the subjective variables measured for words in psycholinguistics, it is typically the one most strongly correlated with frequency, and norms for it are available for a wide array of languages.

We measure correlation (PCC) with mean word familiarity from five databases: Chinese familiarity ratings (Su et al., 2023), English MRC lexical database (Coltheart, 1981; Coltheart and Wilson, 1987), Indonesian lexical norms (Sianipar et al., 2016), Japanese word familiarity ratings (Asahara, 2019)[11], and Spanish lexical norms (Guasch et al., 2016). Evaluation on three alternative, smaller databases for English, Spanish, and Japanese can be found in Appendix G.

As shown in Table 4, in Japanese, TUBELEX$_{\text{base}}$'s correlation is the strongest one, and in all other languages TUBELEX$_{\text{default}}$'s correlation is either the strongest one or not significantly weaker. Correlations without any significant difference from TUBELEX$_{\text{default}}$ are achieved by SUBTLEX in Chinese, by all subtitle corpora in English, by wordfreq in Indonesian, and by all subtitle corpora except EsPal, and by CREA-Spoken in Spanish. In Japanese, GINI achieves a remarkably strong correlation, significantly stronger than TUBELEX$_{\text{default}}$ but still significantly weaker than TUBELEX$_{\text{lemma}}$ (**) and TUBELEX$_{\text{base}}$ (***).[12]

[11] We use the published ratings for reception estimated using a Bayesian linear mixed model.

[12] We computed the levels of significance separately, as Table 4 compares all corpora only against TUBELEX$_{\text{default}}$.

	Corpus	English	Japanese	Spanish
speech	BNC-Spoken	−0.695***	—	—
	CREA-Spoken	—	—	−0.508***
	CSJ	—	−0.563***	—
film/TV subtitles	ACTIV-ES	—	—	−0.516***
	EsPal	—	—	−0.627
	LaboroTV1+2	—	−0.610**	—
	OpenSubtitles	−0.721***	−0.191***	−0.628
	SubIMDB	−0.717***	—	—
	SUBTLEX	−0.696***	—	−0.618
	SUBTLEX-UK	−0.724**	—	—
other	GINI	−0.349***	−0.379***	—
	Wikipedia	−0.651***	−0.487***	−0.454***
	wordfreq	−0.761	−0.605**	−0.559***
our	$\text{TUBELEX}_{\text{default}}$	**−0.762**	**−0.661**	−0.604
	$\text{TUBELEX}_{\text{regex}}$	−0.761**	—	−0.588*
	$\text{TUBELEX}_{\text{base}}$	—	−0.658	—
	$\text{TUBELEX}_{\text{lemma}}$	−0.749	−0.622**	**−0.650****

Table 5: Lexical complexity correlation. Strongest (lowest) correlations for each language are in bold.

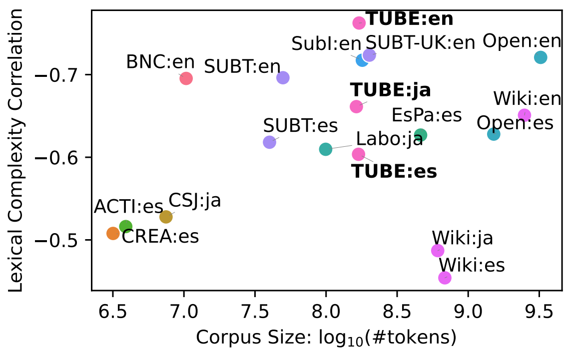

Figure 3: Lexical complexity correlation and corpus size. Labeled "*corpus abbr.*:*lang. code*", "TUBE" is $\text{TUBELEX}_{\text{default}}$, not showing outlier "Open:ja".

No corpus achieves results comparable to TUBELEX results across all five languages, but SUBTLEX corpora do not differ significantly on the three languages where they are available. Similarly to LDT, Figure 2 shows that TUBELEX and SUBTLEX perform remarkably well relative to their size, roughly one order of magnitude smaller than the OpenSubtitles corpora.

5.5 Lexical Complexity

Lexical complexity is a subjective rating of word comprehension difficulty in a sentence context. Its prediction can be used for the practical NLP task of lexical simplification. The MultiLS dataset (Shardlow et al., 2024), which we use for evaluation, was annotated by non-native speakers (Japanese) or a mix of natives and non-natives (English and Spanish), whereas the two previously evaluated psycholinguistic tasks only use data collected from natives. No lexical complexity dataset is available for Chinese or Indonesian.

	Corpus / ST System	English	Japanese	Spanish
speech	BNC-Spoken	0.475	—	—
	CREA-Spoken	—	—	0.186
	CSJ	—	0.306	—
film/TV subtitles	ACTIV-ES	—	—	−0.253
	EsPal	—	—	0.170
	LaboroTV1+2	—	0.349	—
	OpenSubtitles	0.445	0.019	0.332
	SubIMDB	0.377	—	—
	SUBTLEX	0.394	—	−0.254
	SUBTLEX-UK	0.513	—	—
other	GINI	0.041	0.105	—
	Wikipedia	0.365	0.231	−0.255
	wordfreq	**0.578**	0.364	0.268
our	$\text{TUBELEX}_{\text{default}}$	0.553	0.405	0.328
	$\text{TUBELEX}_{\text{regex}}$	0.552	—	0.308
	$\text{TUBELEX}_{\text{base}}$	—	**0.424**	—
	$\text{TUBELEX}_{\text{lemma}}$	0.561	0.234	0.299
top ST	Archaelogy (ID=2)	0.439	−0.098	0.230
	GMU (ID=1)	0.525	−0.039	−0.073
	TMU-HIT (ID=2)	0.515	0.413	**0.494**

Table 6: Coefficient of determination R^2 achieved in lexical complexity prediction, compared with top shared task systems (top ST), citing their results from Shardlow et al. (2024). Best (highest) results for each language are in bold.

As shown in Table 5, the strongest correlation in English and Japanese is achieved by $\text{TUBELEX}_{\text{default}}$, and in Spanish by $\text{TUBELEX}_{\text{lemma}}$. Correlations without any significant difference from $\text{TUBELEX}_{\text{default}}$ are achieved by wordfreq in English, and all subtitle corpora except ACTIV-ES for Spanish. Similarly to the previous tasks, Figure 3 shows that TUBELEX and SUBTLEX perform remarkably well relative to their size.

As the dataset was used for evaluation of lexical complexity prediction in a shared task (Shardlow et al., 2024), we also compare predictions based on TUBELEX with top shared task participants. To do so, we fit a linear regression model using a single variable, log-frequency or GINI values to the shared task's trial data (30 instances for each language), and clip the predicted values to the range [0, 1]. We compare our results with the shared task submissions that achieved the highest coefficient of determination R^2 (and also the highest correlation) in individual languages (TMU-HIT, Enomoto et al., 2024; and GMU, Goswami et al., 2024), and on average across all shared task languages (Archaeology, Cristea and Nisioi, 2024).

As shown in Table 6, the best results are achieved by $\text{TUBELEX}_{\text{lemma}}$, closely followed by $\text{TUBELEX}_{\text{default}}$, in English; $\text{TUBELEX}_{\text{base}}$, closely followed by TMU-HIT and $\text{TUBELEX}_{\text{default}}$, in Japanese; and by TMU-HIT in

Spanish, where it outperformed others by a large margin. As we are using cited R^2 values achieved by task participants, we cannot evaluate statistical significance in this case.

A simple linear regression using TUBELEX frequencies has therefore outperformed the top shared task submissions in English and Japanese, namely gradient boosting using multiple features (Archaelogy), ensemble of finetuned BERT models (GMU), and GPT-4 few-shot chain-of-thought prompting (TMU-HIT). It also outperformed the first two of them in Spanish.

It should be noted that, in this case, we are evaluating the prediction of lexical complexity using R^2, not a mere correlation with lexical complexity.[13]

If we looked only at correlation, thus ignoring misprediction of mean and variance, TMU-HIT's predictions would be more strongly correlated than TUBELEX's in the three languages, and those of the other two systems in English (complete results provided in Appendix I). This may indicate limitations of LLM prompting as a regression method.

6 Discussion

6.1 Tokenization and Lemmatization

The differences between TUBELEX variants in the evaluation were generally small, but a few observations can be made about each language:

For English, the default variant performs the best across the tasks, but the simpler regex tokenization is never significantly worse. Both always outperform lemmatization.

In the evaluation of Indonesian, limited to familiarity, default tokenization performed the best.

For Japanese, the base form performs the best for familiarity, and default tokenization for non-native lexical complexity. Both always outperform the orthographically normalized lemma, which show the importance of the exact written form in Japanese.

For Spanish, lemma performed the best for both familiarity and lexical complexity. Regex and default tokenization performed well in LDT only because the data is already limited to uninflected words. For an inflected language such as Spanish, lemmatization has a clear benefit.

6.2 Spoken vs. Written Language

There is a continuum of what we might call spoken and written language in simple terms. During human evaluation we have, for instance, observed that speakers in some videos are reading aloud. On one hand, this would make the subtitles representative of written language, not spoken language. On the other hand, the texts being read cover diverse registers (e.g. the Bible, professionally announced news, or a pre-written speech), and the speakers often shift between reading and commenting. Singing, recitation, scripted acting, and speech rehearsed to various degrees, all appear on YouTube and fall on this written-spoken continuum. Perhaps surprisingly, the same issues apply to corpora of "spontaneous" speech, as they often collect speeches that are prepared (e.g. whole CSJ, news in CREA-Spoken).

In human evaluation (Section 3.5) we have labeled two categories in TUBELEX that we think require attention: songs, which could be overrepresented on YouTube compared to everyday exposure, and synthesized speech, which is effectively written language in disguise.

Overall, we believe that the diverse content found on YouTube contributes to the representativeness of the whole spectrum of spoken language at the small expense of including some amount of written language or songs. This contrasts with most speech corpora, as well as and film subtitle corpora. Speech corpora typically restrict the type or topics of speech they contain by design.

For instance, CSJ is limited to prepared monologue in common Japanese (Maekawa et al., 2000), omitting any dialogue and dialect, and HKUST/MTS (Liu et al., 2006) is limited to dialogue about 40 specified topics. Film and TV subtitle corpora, on the other hand, consist predominantly of scripted dialogue.

6.3 Corpus Size

Previous studies found that in addition to depending on corpus content, correlation with LDT or familiarity grows approximately logarithmically with corpus size (Tanaka-Ishii and Terada, 2011; Paetzold and Specia, 2016). For corpora over 10^7 tokens, such growth reflects better frequency estimates for low-frequency words, and is measurable and statistically significant only if such words are sufficiently represented in the evaluation dataset.

As we built TUBELEX using a fixed size (120,000) sample of videos for each language, the final corpus size depends on the number of valid videos after cleaning (see Table 1). The Chinese (18M tokens) and Indonesian (38M tokens) corpora are substantially smaller than the others (163M to 171M tokens). We therefore expect that

[13] While R^2 can also be defined as the square of PCC, we use the definition consistent with the shared task evaluation (Shardlow et al., 2024), and implemented in scikit-learn (`https://scikit-learn.org/`) as `r2_score`. With this definition, R^2 is a linear function of the mean squared error, therefore penalizing misprediction of mean and variance.

improvements in correlation could be made by collecting larger corpora, particularly for these two languages, although the effect could be difficult to assess on available data. Increased corpus size would also likely benefit language models (see Appendix C).

Sizes of all corpora and datasets used in this study can be found in Appendix F and in Appendix H, respectively.

6.4 Beyond Frequency

Our goal was to evaluate the TUBELEX corpus as an approximation of spoken vocabulary. For a comprehensive comparison with other corpora, we limited the evaluation to word frequency – the only statistic available for many of the compared corpora. Similarly, most of the data we used for evaluation (LDT, familiarity and complexity datasets) is limited to or focused on single-word items. There are several ways, TUBELEX could be used or extended for other purposes:

Language modeling. More complex language models would allow more advanced applications of the corpus data such as modeling human surprisal (e.g. Wilcox et al., 2023). We used the current corpus data to train two basic language models: an n-gram model and word embeddings, which achieved a mixed performance in our evaluation (Appendix C). Collecting larger data may improve the performance of the embeddings, and it would also be essential for training more complex language models (masked language models or generative language models). We expect that, similarly to the results achieved by the current embeddings, models trained exclusively on subtitles would lack in some areas, and consequently that training on mixed data would be suitable for a wide range of applications.

Measures of dispersion and contextual diversity. Single-word corpus statistics are not limited to frequency. We have compared TUBELEX frequency to GINI, a measure of dispersion based on Twitter data, but we leave evaluation of various dispersion metrics computed from TUBELEX for future research. Computation of dispersion metrics generally requires a corpus divided into suitable units such as documents, which TUBELEX provides (videos and channels). Another metric that uses such units is contextual diversity (the number of units in which a word occurs), proposed by Adelman et al. (2006) as an alternative to frequency for psycholinguistic modeling. TUBELEX word lists readily provide contextual diversity as numbers of videos and channels for each word.

Combining sources. Combination of frequencies from multiple corpora often achieve more robust results than using a single corpus. While TUBELEX often outperformed wordfreq, which is a resource combining multiple corpus frequencies, combining TUBELEX frequencies with other sources (e.g. film subtitles) could also result in a more robust performance, especially for languages underrepresented on YouTube.

7 Conclusion

We built a YouTube subtitle corpus of untranslated Chinese, English, Indonesian, Japanese, and Spanish. The frequencies showed consistently strong correlation with LDT, word familiarity, and lexical complexity across the languages. In a comparison with film and TV subtitle corpora, speech corpora, and other common frequency resources, only the SUBTLEX corpora were comparable in correlation strength and consistence. TUBELEX, however, covers Japanese and Indonesian, for which a SUBTLEX corpus is not available. TUBELEX also excelled in the practical task of lexical complexity prediction, where a linear regression based on its frequencies not only outperformed all subtitle and speech corpora but also all submissions in a recent shared task on English and Japanese and all but one on Spanish.

TUBELEX data can be easily used in applications that require spoken vocabulary frequencies (examples given in Section 2.2), which have typically relied on film subtitles but lacked a suitable resource for many languages. Our method can be used to create TUBELEX corpora for additional languages. We see extension beyond using single-word frequencies (outlined in Section 6.4) as a promising direction for future research.

Limitations

We focused on evaluation of our corpus in terms of unigram frequencies as an approximation of language exposure, evaluating them using psycholinguistic data and lexical complexity.

While we also released the higher n-grams based on our corpus, we did not evaluate them. We provided only limited evaluation of the word embeddings trained on the corpus (in Appendix C). While our embeddings outperformed those based on Wikipedia in the word similarity task, they achieved lower scores than the much larger OpenSubtitles corpus. We assume this to be an effect of modest corpus size, and expect this to affect the *n*-gram model as well. In this work, we artificially limited the subtitles collected from YouTube to a quantity suitable for modeling unigram frequency. In future work, we plan to explore training more complex models from more extensive YouTube data or joint training from subtitle and written data.

While TUBELEX exceeds traditional speech corpora in size and outperforms them in our evaluation, it has serious limitations for linguistic research. Compared to most specialized speech corpora, it lacks information about types of speech or demographic composition, and suffers from varying quality of transcription. TUBELEX is not limited to any particular language standard and mixes both different language varieties and registers.

We have only collected and evaluated data for five languages. By intentionally selecting a diverse set of languages, however, we demonstrated that our approach is widely applicable to languages with a large enough presence on YouTube. We made our complete source code available for others to reuse and extend.

Estimating how much subtitle data for a particular language is available on YouTube requires nontrivial effort. For a reliable estimate, it is necessary to identify videos by keywords search, scraping the metadata for each video, and evaluating at least a sample of the subtitles using automated language identification (see Section 3). We did this for two additional languages: For Czech, which has much fewer speakers than any of the current five languages (9.6 million L1 speakers; Eberhard et al., 2024), we found enough data to build a corpus comparable to the current five TUBELEX language. We could not, however, find enough data for Pashto with 44 million L1 speakers (ibid.). We hypothesize that this reflects not only the number of speakers and the popularity of YouTube among them, but also the relative prestige of Pashto in the two multi-lingual countries where it is mainly spoken, Afghanistan and Pakistan. Similar challenges may also affect other low-resource languages.

Ethical Considerations

The content of YouTube subtitles is copyrighted, which precludes us from distributing the full text of the corpus. In terms of copyright, it is no different from film subtitles, but since YouTube consists of user-generated content, we also had to consider the privacy of the video authors.

We only downloaded subtitles for videos that could be found using the YouTube website search function. The search function is restricted to public videos, excluding any unlisted or private videos.[14]

None of the data that we have released contains identification of individual videos, channels, or video uploaders. The statistical language models we have released contain sequences of at most five consecutive words. The released data does not contain longer excerpts from the original subtitles.

We anonymized multiple kinds of potentially sensitive information before we derived any frequency lists or models from it. In particular, we masked email addresses, HTTP(S) URLs, apparent web URLs without an explicit protocol (e.g. `x.com/username`), apparent social network handles starting with `@`, and all sequences of digits, which are the primary constituent of phone numbers, IP addresses, account numbers, and other personally identifying information.

As we have accessed YouTube without sign-in, our corpus does not contain any subtitles for age-restricted videos, which YouTube defines as not appropriate for viewers under 18.[15] Note that while YouTube's age restriction also applies to "excessive profanity", some subtitles in our corpus still contain vulgar or otherwise inappropriate language, which we did not attempt to remove.

References

James S. Adelman, Gordon D.A. Brown, and José F. Quesada. 2006. Contextual Diversity, Not Word Frequency, Determines Word-Naming and Lexical Decision Times. *Psychological Science*, 17(9):814–823.

Jose Armando Aguasvivas, Manuel Carreiras, Marc Brysbaert, Paweł Mandera, Emmanuel Keuleers, and Jon Andoni Duñabeitia. 2018. SPALEX: A Spanish Lexical Decision Database From a Massive Online Data Collection. *Frontiers in Psychology*, 9.

María Angeles Alonso, Angel Fernandez, and Emiliano Díez. 2011. Oral frequency norms for 67,979 Spanish words. *Behavior Research Methods*, 43(2):449–458.

Shigeaki Amano and Tadahisa Kondo. 1999. *NTT Database Series: Lexical Properties of Japanese [NTT dētabēsu shirīzu nihongo no goi tokusei] (in Japanese)*, volume 1–6. Sanseidō.

Shintaro Ando and Hiromasa Fujihara. 2021. Construction of a Large-Scale Japanese ASR Corpus on TV Recordings. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6948–6952.

[14] `https://support.google.com/youtube/answer/157177?hl=en`

[15] `https://support.google.com/youtube/answer/2802167?hl=en`

Masayuki Asahara. 2019. Word familiarity rate estimation using a Bayesian linear mixed model. In *Proceedings of the First Workshop on Aggregating and Analysing Crowdsourced Annotations for NLP*, pages 6–14, Hong Kong. Association for Computational Linguistics.

David A. Balota, Melvin J. Yap, Michael J. Cortese, Keith A. Hutchison, Brett Kessler, Bjorn Loftis, James H. Neely, Douglas L. Nelson, Greg B. Simpson, and Rebecca Treiman. 2007. The English Lexicon Project. *Behavior Research Methods*, 39(3):445–459.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Marc Brysbaert and Kevin Diependaele. 2013. Dealing with zero word frequencies: A review of the existing rules of thumb and a suggestion for an evidence-based choice. *Behavior Research Methods*, 45(2):422–430.

Marc Brysbaert and Boris New. 2009. Moving beyond Kucera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, 41(4):977–990.

Curt Burgess and Kay Livesay. 1998. The effect of corpus size in predicting reaction time in a basic word recognition task: Moving on from Kučera and Francis. *Behavior Research Methods, Instruments & Computers*, 30(2):272–277.

Qing Cai and Marc Brysbaert. 2010. SUBTLEX-CH: Chinese Word and Character Frequencies Based on Film Subtitles. *PLoS ONE*, 5(6):e10729.

Xiaobin Chen and Detmar Meurers. 2016. Characterizing text difficulty with word frequencies. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 84–94, San Diego, CA. Association for Computational Linguistics.

M. (Max) Coltheart and Michael John Wilson. 1987. MRC Psycholinguistic Database Machine Usable Dictionary : Expanded Shorter Oxford English Dictionary entries / Max Coltheart and Michael Wilson. *Oxford Text Archive*.

Max Coltheart. 1981. The MRC psycholinguistic database. *The Quarterly Journal of Experimental Psychology A: Human Experimental Psychology*, 33A(4):497–505.

BNC Consortium. 2007. British National Corpus, XML edition.

Petru Cristea and Sergiu Nisioi. 2024. Archaeology at mlsp 2024: Machine translation for lexical complexity prediction and lexical simplification. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 610–617, Mexico City, Mexico. Association for Computational Linguistics.

Fernando Cuetos, Maria Glez-Nosti, Analía Barbón, and Marc Brysbaert. 2011. SUBTLEX-ESP: Spanish word frequencies based on film subtitles. *Psicológica*, 32(2):133–143.

Yasuharu Den, Toshinobu Ogiso, Hideki Ogura, Atsushi Yamada, Nobuaki Menematsu, Kiyotaka Uchimoto, and Hanae Koiso. 2007. The development of an electronic dictionary for morphological analysis and its application to Japanese corpus linguistics [Kōpasu Nihongogaku no tame no gengo shigen : Keitaiso kaiseki yō denshika jisho no kaihatsu to sono ōyō] (in Japanese). *Japanese Linguistics [Nihongo kagaku]*, 22(5):101–123.

Andrew Duchon, Manuel Perea, Nuria Sebastián-Gallés, Antonia Martí, and Manuel Carreiras. 2013. EsPal: One-stop shopping for Spanish word properties. *Behavior Research Methods*, 45(4):1246–1258.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig, editors. 2024. *Ethnologue: Languages of the World (Online Version)*, 27th edition edition. Ethnologue Series. SIL International, Dallas, Texas.

Taisei Enomoto, Hwichan Kim, Tosho Hirasawa, Yoshinari Nagai, Ayako Sato, Kyotaro Nakajima, and Mamoru Komachi. 2024. TMU-HIT at MLSP 2024: How well can GPT-4 tackle multilingual lexical simplification? In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 590–598, Mexico City, Mexico. Association for Computational Linguistics.

Real Academia Española. 2004. Corpus de Referencia del Español Actual.

Jerid Francom, Mans Hulden, and Adam Ussishkin. 2014. ACTIV-ES: a comparable, cross-dialect corpus of 'everyday' Spanish from Argentina, Mexico, and Spain. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1733–1737, Reykjavik, Iceland. European Language Resources Association (ELRA).

Victoria M. Garlock, Amanda C. Walley, and Jamie L. Metsala. 2001. Age-of-Acquisition, Word Frequency, and Neighborhood Density Effects on Spoken Word Recognition by Children and Adults. *Journal of Memory and Language*, 45(3):468–492.

Dhiman Goswami, Kai North, and Marcos Zampieri. 2024. GMU at MLSP 2024: Multilingual lexical simplification with transformer models. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 627–634, Mexico City, Mexico. Association for Computational Linguistics.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of* *the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Marc Guasch, Pilar Ferré, and Isabel Fraga. 2016. Spanish norms for affective and lexico-semantic variables for 1,400 words. *Behavior Research Methods*, 48(4):1358–1369.

Kenneth Heafield, Ivan Pouzyrevsky, Jonathan H. Clark, and Philipp Koehn. 2013. Scalable modified Kneser-Ney language model estimation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 690–696, Sofia, Bulgaria. Association for Computational Linguistics.

Joseph Henrich. 2020. *The WEIRDest People in the World: How the West Became Psychologically Peculiar and Particularly Prosperous*. Farrar, Straus and Giroux.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016a. FastText.zip: Compressing text classification models. *ArXiv preprint*, arXiv:1612.03651v1 [cs].

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016b. Bag of Tricks for Efficient Text Classification. *ArXiv preprint*, arXiv:1607.01759v3 [cs].

Henry Kučera and Winthrop Nelson Francis. 1967. *Computational Analysis of Present-day American English*. Brown University Press.

Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. 2004. Applying conditional random fields to Japanese morphological analysis. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 230–237, Barcelona, Spain. Association for Computational Linguistics.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929, Portorož, Slovenia. European Language Resources Association (ELRA).

Pierre Lison, Jörg Tiedemann, and Milen Kouylekov. 2018. OpenSubtitles2018: Statistical rescoring of sentence alignments in large, noisy parallel corpora. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Yi Liu, Pascale Fung, Yongsheng Yang, Christopher Cieri, Shudong Huang, and David Graff. 2006. HKUST/MTS: A Very Large Scale Mandarin Telephone Speech Corpus. In *Chinese Spoken Language Processing*, pages 724–735, Berlin, Heidelberg. Springer.

Kikuo Maekawa, Hanae Koiso, Sadaoki Furui, and Hitoshi Isahara. 2000. Spontaneous speech corpus of Japanese. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00)*, Athens, Greece. European Language Resources Association (ELRA).

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. *ArXiv preprint*, arXiv:1301.3781v3 [cs].

F. Javier Moreno-Martínez, Pedro R. Montoro, and Inmaculada C. Rodríguez-Rojo. 2014. Spanish norms for age of acquisition, concept familiarity, lexical frequency, manipulability, typicality, and other variables for 820 words from 14 living/nonliving concepts. *Behavior Research Methods*, 46(4):1088–1097.

Taichi Murayama, Shoko Wakamiya, and Eiji Aramaki. 2018. WORD GINI: A proposal and application of an index to capture word usage bias [WORD GINI: Go no shiyō no katayori wo tsukamaeru shihyō no teian to sono ōyō] (in Japanese). *The 24th Annual Conference of the Association for Natural Language Processing [Gengoshori gakkai dai 24 kai nenji taikai]*, pages 698–701.

Boris New, Marc Brysbaert, Jean Veronis, and Christophe Pallier. 2007. The use of film subtitles to estimate word frequencies. *Applied Psycholinguistics*, 28(4):661–677.

NINJAL (National Institute for Japanese Language and Linguistics [Kokuritsu Kokugo Kenkyūjo]). 2016. Construction of the Corpus of Spontaneous Japanese [Nihongo hanashikotoba kōpasu no kōchikuhō] (in Japanese).

NINJAL (National Institute for Japanese Language and Linguistics [Kokuritsu Kokugo Kenkyūjo]). 2018. The Wordlist for the Corpus of Spontaneous Japanese' (Version 201803) [Nihongo hanashikotoba kōpasu goihyō (Version 201803)] (in Japanese).

Daiki Nishihara and Tomoyuki Kajiwara. 2020. Word complexity estimation for Japanese lexical simplification. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3114–3120, Marseille, France. European Language Resources Association.

Nadezda Okinina, Jennifer-Carmen Frey, and Zarah Weiss. 2020. CTAP for Italian: Integrating components for the analysis of Italian into a multilingual linguistic complexity analysis tool. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 7123–7131, Marseille, France. European Language Resources Association.

Gustavo Paetzold and Lucia Specia. 2016. Collecting and exploring everyday language for predicting psycholinguistic properties of words. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages

1669–1679, Osaka, Japan. The COLING 2016 Organizing Committee.

Hien Pham, Benjamin V. Tucker, and R. Harald Baayen. 2019. Constructing two Vietnamese corpora and building a lexical database. *Language Resources and Evaluation*, 53(3):465–498.

Ayu Purwarianti, Alvin Andhika, Alfan Farizki Wicaksono, Irfan Afif, and Filman Ferdian. 2016. InaNLP: Indonesia natural language processing toolkit, case study: Complaint tweet classification. In *2016 International Conference On Advanced Informatics: Concepts, Theory And Application (ICAICTA)*, pages 1–5.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

Frankie Robertson, Li-Hsin Chang, and Sini Söyrinki. 2022. TallVocabL2Fi: A tall dataset of 15 Finnish L2 learners' vocabulary. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6377–6386, Marseille, France. European Language Resources Association.

Nipun Sadvilkar and Mark Neumann. 2020. PySBD: Pragmatic Sentence Boundary Disambiguation. *ArXiv preprint*, arXiv:2010.09657v1 [cs].

Graham G. Scott, Anne Keitel, Marc Becirspahic, Bo Yao, and Sara C. Sereno. 2019. The Glasgow Norms: Ratings of 5,500 words on nine scales. *Behavior Research Methods*, 51(3):1258–1270.

Matthew Shardlow. 2013. A comparison of techniques to automatically identify complex words. In *51st Annual Meeting of the Association for Computational Linguistics Proceedings of the Student Research Workshop*, pages 103–109, Sofia, Bulgaria. Association for Computational Linguistics.

Matthew Shardlow, Fernando Alva-Manchego, Riza Batista-Navarro, Stefan Bott, Saul Calderon Ramirez, Rémi Cardon, Thomas François, Akio Hayakawa, Andrea Horbach, Anna Hülsing, Yusuke Ide, Joseph Marvin Imperial, Adam Nohejl, Kai North, Laura Occhipinti, Nelson Peréz Rojas, Nishat Raihan, Tharindu Ranasinghe, Martin Solis Salazar, Sanja Štajner, Marcos Zampieri, and Horacio Saggion. 2024. The BEA 2024 shared task on the multilingual lexical simplification pipeline. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 571–589, Mexico City, Mexico. Association for Computational Linguistics.

Agnes Sianipar, Pieter van Groenestijn, and Ton Dijkstra. 2016. Affective Meaning, Concreteness, and Subjective Frequency Norms for Indonesian Words. *Frontiers in Psychology*, 7.

Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, Valentin Hofmann, Ananya Harsh Jha, Sachin Kumar, Li Lucy, Xinxi Lyu, Nathan Lambert, Ian Magnusson, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew E. Peters, Abhilasha Ravichander, Kyle Richardson, Zejiang Shen, Emma Strubell, Nishant Subramani, Oyvind Tafjord, Pete Walsh, Luke Zettlemoyer, Noah A. Smith, Hannaneh Hajishirzi, Iz Beltagy, Dirk Groeneveld, Jesse Dodge, and Kyle Lo. 2024. Dolma: An Open Corpus of Three Trillion Tokens for Language Model Pretraining Research. *ArXiv preprint*, arXiv:2402.00159v2 [cs].

Robyn Speer. 2022. Rspeer/wordfreq: V3.0. Zenodo.

Sanja Štajner, Daniel Ferrés, Matthew Shardlow, Kai North, Marcos Zampieri, and Horacio Saggion. 2022. Lexical simplification benchmarks for English, Portuguese, and Spanish. *Frontiers in Artificial Intelligence*, 5.

James H. Steiger. 1980. Tests for comparing elements of a correlation matrix. *Psychological Bulletin*, 87(2):245–251.

Yongqiang Su, Yixun Li, and Hong Li. 2023. Familiarity ratings for 24,325 simplified Chinese words. *Behavior Research Methods*, 55(3):1496–1509.

Nishant Subramani, Sasha Luccioni, Jesse Dodge, and Margaret Mitchell. 2023. Detecting personal information in training corpora: an analysis. In *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023)*, pages 208–220, Toronto, Canada. Association for Computational Linguistics.

Zhewei Sun, Qian Hu, Rahul Gupta, Richard Zemel, and Yang Xu. 2024. Toward informal language processing: Knowledge of slang in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1683–1701, Mexico City, Mexico. Association for Computational Linguistics.

Shinnosuke Takamichi, Ludwig Kürzinger, Takaaki Saeki, Sayaka Shiota, and Shinji Watanabe. 2021. JTubeSpeech: Corpus of Japanese speech collected from YouTube for speech recognition and speaker verification. *ArXiv preprint*, arXiv:2112.09323v1 [cs.CL].

Kumiko Tanaka-Ishii and Hiroshi Terada. 2011. Word Familiarity and Frequency. *Studia Linguistica*, arXiv:1806.03431 [cs]:96–116.

Annika Tjuka. 2020. General patterns and language variation: Word frequencies across English, German, and Chinese. In *Proceedings of the Workshop on the Cognitive Aspects of the Lexicon*, pages 23–32, Online. Association for Computational Linguistics.

Yiu-Kei Tsang, Jian Huang, Ming Lui, Mingfeng Xue, Yin-Wah Fiona Chan, Suiping Wang, and Hsuan-Chih Chen. 2018. MELD-SCH: A megastudy of lexical decision in simplified Chinese. *Behavior Research Methods*, 50(5):1763–1777.

Stéphan Tulkens, Dominiek Sandra, and Walter Daelemans. 2020. Orthographic codes and the neighborhood effect: Lessons from information theory. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 172–181, Marseille, France. European Language Resources Association.

Walter J. B. van Heuven, Pawel Mandera, Emmanuel Keuleers, and Marc Brysbaert. 2014. SUBTLEX-UK: A new and improved word frequency database for British English. *The Quarterly Journal of Experimental Psychology*, 67(6):1176–1190.

Jeroen van Paridon and Bill Thompson. 2021. Subs2vec: Word embeddings from subtitles in 55 languages. *Behavior Research Methods*, 53(2):629–655.

Ivan Vulić, Simon Baker, Edoardo Maria Ponti, Ulla Petti, Ira Leviant, Kelly Wing, Olga Majewska, Eden Bar, Matt Malone, Thierry Poibeau, Roi Reichart, and Anna Korhonen. 2021. Multi-SimLex: A Large-Scale Evaluation of Multilingual and Crosslingual Lexical Semantic Similarity. *Computational Linguistics*, 46(4):847–897.

Ken Whistler. 2023. UAX #15: Unicode Normalization Forms. Technical report, Unicode Consortium.

Muhammad Satrio Wibowo, Ade Romadhony, and Siti Sa'adah. 2019. Lexical and Syntactic Simplification for Indonesian Text. In *2019 International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*, pages 64–68.

Ethan G. Wilcox, Tiago Pimentel, Clara Meister, Ryan Cotterell, and Roger P. Levy. 2023. Testing the predictions of surprisal theory in 11 languages. *Transactions of the Association for Computational Linguistics*, 11:1451–1470.

A Preprocessing and Hyperparameters for Word Embeddings and Statistical Language Models

	Preprocessing
Sentence Splitting	On subtitle cue boundaries, and rule-based using PySBD 0.3.4 (Sadvilkar and Neumann, 2020), with rules added for Indonesian based on InaNLP (Purwarianti et al., 2016).
Tokenization	TUBELEX regex tokenization for English, Japanese, and Spanish, and default tokenization for Japanese and Spanish (details in Section 3.5).
Normalization	Lower case, Unicode NFKC (Whistler, 2023).
	Hyperparameters
Word Embeddings	300-dimensional fastText CBOW model with position weights, 10 negative samples, 10 epochs, character 5-grams, other: default (Grave et al., 2018). – software: `https://github.com/facebookresearch/fastText` – CLI: `fasttext cbow -dim 300 -neg 10 -epoch 10 -minn 5 -maxn 5`
Statistical Model	Modified Kneser-Ney language model of order 5 (Heafield et al., 2013). – software: `https://kheafield.com/code/kenlm/` – CLI: `lmplz -o 5`

Table 7: Preprocessing and hyperparameters used to train word embeddings and statistical language models on TUBELEX. We used the same preprocessing for both.

B Sizes of Word Embeddings and Statistical Language Models

Language	Statistical Language Model n-Grams					FastText Vocabulary
	$n = 1$	$n = 2$	$n = 3$	$n = 4$	$n = 5$	
Chinese	432,670	5,760,278	12,642,939	15,320,475	15,264,065	114,237
English	420,583	12,798,615	53,560,054	99,046,960	126,309,472	131,757
Indonesian	300,647	6,746,497	20,766,383	29,052,968	31,555,431	81,801
Japanese	405,676	10,898,295	45,793,067	85,956,149	113,644,170	145,429
Spanish	613,056	15,482,447	62,465,043	113,520,573	139,641,687	197,107

Table 8: Numbers of n-grams of the statistical language models (KenLM) and vocabulary size of the word embeddings (fastText) trained on TUBELEX. (Minimum frequency for the fastText model is 5.)

C Evaluation of Word Embeddings

In Tables 9 and 10, we compare the performance of TUBELEX embeddings in word analogy and word similarity with previously published embeddings trained on Wikipedia (Grave et al., 2018), and OpenSubtitles2018 (van Paridon and Thompson, 2021).

The TUBELEX embeddings were trained using the same hyperparameters (see Appendix A) as the Wikipedia embeddings by Grave et al. (2018), while van Paridon and Thompson (2021) used a different setup. All compared embeddings are fastText, but we do not use character n-grams to embed out-of-vocabulary words. For fairness, we always evaluate the whole dataset including out-of-vocabulary words. We only evaluate on English and Spanish, for which comparable evaluation data and pre-trained OpenSubtitles2018 embeddings are available.

In word analogy, TUBELEX embeddings underperform the other embeddings. In word similarity, they outperform Wikipedia, but slightly underperform OpenSubtitles. The overall performance is very close to the OpenSubtitles embeddings, and we hypothesize that the gap between the two is caused by the TUBELEX corpus being an order of magnitude smaller than OpenSubtitles. While we have observed that TUBELEX's size does not affect the quality of unigram frequencies, the word embeddings would likely benefit from a larger corpus.

Language	Embeddings	Sem.: Geography	Sem.: Family	Semantic	Syntactic	Total
English	Wikipedia	0.775	0.822	0.778	0.721	0.747
	OpenSubtitles	0.144	0.852	0.184	0.757	0.497
	TUBELEX	0.142	0.626	0.170	0.628	0.420
Spanish	Wikipedia	0.466	0.863	0.484	0.572	0.524
	OpenSubtitles	0.087	0.892	0.125	0.516	0.301
	TUBELEX	0.064	0.839	0.101	0.501	0.281

Table 9: Accuracy in word analogy evaluated on English data (Mikolov et al., 2013) and Spanish data derived from it (`https://crscardellino.net/SBWCE/`). We list separately accuracy in Geography and Family subcategories of the Semantic category.

Language	Embeddings	Pearson's r	Spearman's ρ
English	Wikipedia	0.379	0.434
	OpenSubtitles	0.468	0.532
	TUBELEX	0.385	0.457
Spanish	Wikipedia	0.342	0.387
	OpenSubtitles	0.445	0.475
	TUBELEX	0.415	0.450

Table 10: Correlation in word similarity evaluated on parallel English and Spanish data from Multi-SimLex (Vulić et al., 2021).

D GINI Metric

The GINI metric was proposed by Murayama et al. (2018) for simplification and readability assessment. It is inspired by the Gini index of income inequality, and measures words' dispersion in frequency of use by different people. Similarly to TUBELEX, it uses user-generated data (from Twitter), and its pre-computed values are available for English and Japanese. As the original study is available only in Japanese, we summarize its computation for the reader's convenience:

1. Construct a matrix of numbers of word occurrences shaped $(m, n) = (\#\text{users}, \#\text{words})$.
2. Normalize twice: first making each row sum to 1, then making each column sum to 1.
3. For each column (word) $\mathbf{x}$, compute: $Gini(\mathbf{x}) = \frac{1}{2\mu_{\mathbf{x}} n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} |x_i - x_j|$
4. To compute the final values of the metric, apply: $-\log(1 - Gini(\mathbf{x}))$.

E Details of the Evaluated Corpora

Table 11 shows details of the evaluated corpora. We compute frequency with Laplace smoothing (see Section 5.2) from them with the following exception: for ACTIV-ES and wordfreq, which do not provide token counts, we instead assign the corpus minimum frequency to missing words. We also directly use GINI values, analogously assigning the corpus maximum value to missing words, as high GINI values indicate high dispersion. The words featured in our experiments may be tokenized as multiple tokens. We assign them the minimum of the frequencies of the individual tokens, and maximum of the GINI values. In all experiments, we use the logarithm of frequency. For GINI, we use the additive inverse ($\log(1-Gini(\mathbf{x}))$) of the original metric's values described in Appendix D, for the purpose of comparison with log-frequencies used for the other corpora.

	Corpus	Details	Source
speech	**BNC-Spoken**	We construct a frequency list for the spoken subset of BNC by computing a difference of the "all" and "written" unlemmatized BNC frequency lists compiled by Adam Kilgarif.	– https://www.kilgarriff.co.uk/bnc-readme.html
	CREA-Spoken	We use the frequency list of the spoken subset of CREA.	Alonso et al. (2011)
	CSJ	We use the published CSJ frequency list lemmatized using MeCab/Unidic.	NINJAL (2018)
	HKUST/MTS	We construct a frequency list from the corpus transcripts using the jieba tokenization.	– https://catalog.ldc.upenn.edu/LDC2005T32
film/TV subtitles	**ACTIV-ES**	We use the published 1-gram frequency list, version 0.2.	– https://github.com/francojc/activ-es
	EsPal	We use the public web form, to retrieve frequencies of all tokens for our experiments.	– https://www.bcbl.eu/databases/espal/wordidx.php
	LaboroTV1+2	LaboroTVSpeech (2020) and LaboroTVSpeech2 (2024): We combine pre-tokenized training and development data of the two releases of to generate a single frequency list.	– https://laboro.ai/activity/column/engineer/eg-laboro-tv-corpus-jp/ – https://laboro.ai/activity/column/engineer/laborotvspeech2/
	OpenSubtitles	We use the published frequency lists from the updated 2018 version of the collection. For Chinese we use the list identified as "China mainland", which mostly uses simplified Chinese characters. (In the 2018 version, Chinese subtitles are divided into "China mainland" and "Taiwan". Details of the division are not documented. The OpenSubtitles website itself divides Chinese into simplified, traditional, and Cantonese.)	– https://opus.nlpl.eu/OpenSubtitles/&/v2018/OpenSubtitles
	SubIMDB	We generate a frequency list from the full SubIMDB corpus, which comes in a pre-tokenized form. No frequency list was published for the corpus.	– https://zenodo.org/records/2552407
	SUBTLEX	We use the published SUBTLEX raw frequency counts for English (US), Spanish (ESP), and Chinese (CH).	– https://www.ugent.be/pp/experimentele-psychologie/en/research/documents/subtlexus/subtlexus2.zip (US) – http://www.ugent.be/pp/experimentele-psychologie/en/research/documents/subtlexch/subtlexchwf.zip (CH) – https://web.archive.org/web/20220702151524/http://crr.ugent.be/papers/SUBTLEX-ESP.zip (ESP)
	SUBTLEX-UK	For English, we use raw frequency counts from SUBTLEX-UK as well.	– https://www.psychology.nottingham.ac.uk/subtlex-uk/SUBTLEX-UK.txt.zip
other	**GINI**	We use the published WORD GINI lists for English and Japanese. (Details about the GINI metric in Appendix D.)	– https://sociocom.naist.jp/word-gini-en/
	Wikipedia	We use frequency lists based on cleaned up Wikipedia text tokenized using a regular expression.	– https://github.com/adno/wikipedia-word-frequency-clean
	wordfreq	We use the default (large) lists available from the Python library.	– https://pypi.org/project/wordfreq/

Table 11: Detailed information and sources for the corpora used for evaluation. Source is the publication, if it contains the frequency lists as supplementary material, or an URL from which the data (corpus or frequency list) is available. The corpora are introduced and cited in Section 2 and Section 5.1.

F Statistics of the Evaluated Corpora

Corpus		Chinese	English	Indonesian	Japanese	Spanish
BNC-Spoken	tokens	—	10,365,473	—	—	—
	types	—	669,417	—	—	—
CREA-Spoken	tokens	—	—	—	—	3,171,903
	types	—	—	—	—	67,979
CSJ	tokens	—	—	—	7,479,773	—
	types	—	—	—	40,630	—
HKUST/MTS	tokens	1,342,379	—	—	—	—
	types	42,247	—	—	—	—
ACTIV-ES	tokens	—	—	—	—	3,897,234
	types	—	—	—	—	80,787
EsPal	tokens	—	—	—	—	462,611,693
	types	—	—	—	—	35,257
LaboroTV1+2	tokens	—	—	—	99,367,439	—
	types	—	—	—	218,762	—
OpenSubtitles	tokens	191,379,324	3,235,391,790	137,231,876	23,665,222	1,512,443,143
	types	1,009,838	2,290,458	456,125	58,856	1,629,907
SubIMDB	tokens	—	179,967,485	—	—	—
	types	—	899,603	—	—	—
SUBTLEX	tokens	33,546,516	49,719,560	—	—	40,017,237
	types	99,121	74,286	—	—	94,261
SUBTLEX-UK	tokens	—	201,706,753	—	—	—
	types	—	160,022	—	—	—
GINI	tokens	—	—	—	—	—
	types	—	324,713	—	208,275	—
Wikipedia	tokens	271,230,431	2,489,387,103	117,956,650	610,467,200	685,158,870
	types	1,403,791	2,161,820	373,461	522,210	986,947
wordfreq	tokens	—	—	—	—	—
	types	334,609	321,180	31,188	214,960	342,072
$\text{TUBELEX}_{\text{default}}$	tokens	17,865,686	170,750,870	34,903,381	163,439,781	169,188,689
	types	432,532	467,296	307,633	409,503	632,112
$\text{TUBELEX}_{\text{regex}}$	tokens	—	170,816,384	34,293,878	—	166,423,254
	types	—	420,718	300,870	—	613,181
$\text{TUBELEX}_{\text{base}}$	tokens	—	—	—	163,439,781	—
	types	—	—	—	378,276	—
$\text{TUBELEX}_{\text{lemma}}$	tokens	—	170,764,637	34,904,605	163,462,537	169,188,635
	types	—	433,545	266,827	329,303	527,060

Table 12: Numbers of tokens and types in the corpora evaluated in Section 5.1. Number of types is always based on the actual frequency lists we use (see Appendix E), after lowercasing and combining equivalent words (a few corpora list separately words differing only in case or POS). Number of tokens are either sums of individual token counts or explicit total token counts if available. GINI and wordfreq data do not report numbers of tokens (only index values and relative frequencies, respectively). Wordfreq also removes types with frequency less than 10^{-8}.

G Evaluation on Alternative Word Familiarity Norms

	Corpus	English (Glasgow)	Japanese (Amano+Kondo)	Spanish (Moreno-Martínez)
speech	BNC-Spoken	0.658*	—	—
	CREA-Spoken	—	—	0.510***
	CSJ	—	0.441***	—
film/TV subtitles	ACTIV-ES	—	—	0.495***
	EsPal	—	—	0.557**
	LaboroTV1+2	—	0.536***	—
	OpenSubtitles	0.650	0.354***	**0.612**
	SubIMDB	**0.675*****	—	—
	SUBTLEX	0.642	—	0.585
	SUBTLEX-UK	0.674***	—	—
other	GINI	0.482***	**0.572*****	—
	Wikipedia	0.446***	0.423***	0.430***
	wordfreq	0.638**	0.510***	0.557***
our	TUBELEX$_{default}$	0.646	0.544	0.610
	TUBELEX$_{regex}$	0.646	—	0.610
	TUBELEX$_{base}$	—	0.564***	—
	TUBELEX$_{lemma}$	0.639*	0.538***	0.609

Table 13: Word familiarity (alternative norms) correlation (PCC). Strongest (highest) correlations for each language are in bold. Glasgow norms (Scott et al., 2019) for English, norms by Moreno-Martínez et al. (2014) for Spanish, and written word familiarity ratings by Amano and Kondo (1999) for Japanese. All three databases are smaller than the ones presented in Section 5.4.

H Evaluation Dataset Sizes

Task	Chinese	English	Indonesian	Japanese	Spanish
Lexical Decision Time	12,576	38,130	—	—	45,190
Lexical Complexity	—	570	—	570	593
Word Familiarity	24,325	4,923	1,490	81,271	1,400
Word Familiarity (Alternative)	—	4,682	—	76,883	820

Table 14: Numbers of instances in the datasets used for evaluation. The individual datasets are introduced in Section 5.3 for lexical decision time, Section 5.5 for lexical complexity, Section 5.4 for word familiarity, and Appendix G for word familiarity – alternative datasets.

I Correlation with Lexical Complexity Predictions

	Corpus / ST System	English	Japanese	Spanish
speech	BNC-Spoken	0.701	—	—
	CREA-Spoken	—	—	0.508
	CSJ	—	0.565	—
film/TV subtitles	ACTIV-ES	—	—	−0.516
	EsPal	—	—	0.627
	LaboroTV1+2	—	0.610	—
	OpenSubtitles	0.721	0.191	0.628
	SubIMDB	0.717	—	—
	SUBTLEX	0.696	—	−0.618
	SUBTLEX-UK	0.726	—	—
other	GINI	0.349	0.379	—
	Wikipedia	0.651	0.487	−0.454
	wordfreq	0.763	0.605	0.559
our	$\text{TUBELEX}_{\text{default}}$	0.766	0.661	0.604
	$\text{TUBELEX}_{\text{regex}}$	0.764	—	0.588
	$\text{TUBELEX}_{\text{base}}$	—	0.663	—
	$\text{TUBELEX}_{\text{lemma}}$	0.758	0.622	0.650
top ST	Archaelogy (ID=2)	0.790	0.485	0.230
	GMU (ID=1)	**0.850**	0.035	−0.073
	TMU-HIT (ID=2)	0.820	**0.733**	**0.762**

Table 15: Correlation (PCC) with lexical complexity predictions, discounting misprediction of mean and variance. Best (highest) results for each language are in bold. Values for top shared task submissions (top ST) are cited from Shardlow et al. (2024). See the corresponding R^2 results, which measure the goodness of fit, in Table 6, and the discussion at the end of Section 5.5.