Collecting and Exploring Everyday Language for Predicting Psycholinguistic Properties of Words

Gustavo Henrique Paetzold and **Lucia Specia**
Department of Computer Science
University of Sheffield, UK
{g.h.paetzold,l.specia}@sheffield.ac.uk

Abstract

Exploring language usage through frequency analysis in large corpora is a defining feature in most recent work in corpus and computational linguistics. From a psycholinguistic perspective, however, the corpora used in these contributions are often not representative of language usage: they are either domain-specific, limited in size, or extracted from unreliable sources. In an effort to address this limitation, we introduce SubIMDB, a corpus of everyday language spoken text we created which contains over 225 million words. The corpus was extracted from 38,102 subtitles of family, comedy and children movies and series, and is the first sizeable structured corpus of subtitles made available. Our experiments show that word frequency norms extracted from this corpus are more effective than those from well-known norms such as Kucera-Francis, HAL and SUBTLEX$_{us}$ in predicting various psycholinguistic properties of words, such as lexical decision times, familiarity, age of acquisition and simplicity. We also provide evidence that contradict the long-standing assumption that the ideal size for a corpus can be determined solely based on how well its word frequencies correlate with lexical decision times.

1 Introduction

Large corpora of text are certainly one of the most fundamental resources in the field of Computational Linguistics. In Psycholinguistics, it has been long established that word frequencies from corpora play a very important role in cognitive processes. Brysbaert and New (2009) points out that frequently occurring words are often much more easily perceived, recalled and associated than rare words (Balota and Chumbley, 1984; Rayner and Duffy, 1986). In Text Simplification, researchers have found a strong relationship between frequencies and word simplicity (Devlin and Tait, 1998).

An inherent limitation of work based on word frequency analysis is that the type of resource used as a corpus is often built for a specific communication purpose, such as news (Burgess and Livesay, 1998). This is however not representative of everyday language usage, particularly from a psycholinguistic perspective. The other extreme of the spectrum features resources compiled from user-generated content, such as micro-blogs. However, these resources often suffer from grammar errors and misspellings, excessive use of acronyms and shortenings, partly due to the constrains of the publication means (e.g. limited number of characters) (Pak and Paroubek, 2010).

This is particularly concerning given that previous research has shown that the source from which a corpus was extracted is one of its most important defining traits. For example, the experiments of Brysbaert and New (2009) and Shardlow (2013) reveal that frequencies from spoken text have a much stronger correlation with psycholinguistic word properties than those from other sources. Their findings greatly highlight the potential of spoken language text, but there are very few examples of resources of this kind available for English. SUBTLEX$_{us}$ is a notable exception: it contains texts extracted from 8,388 subtitles of American movies, and is freely available for download. The OpenSubtitles2016 corpus (Lison and Tiedemann, 2016) is another example, featuring sentences extracted from numerous subtitle files aligned at sentence level across 60 languages.

Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pages 1669–1679, Osaka, Japan, December 11-17 2016.

However, since the subtitles in these corpora are not restricted with respect to genre or domain, their proficiency in capturing everyday language can also be limited. Movies and series span from lighthearted productions for toddlers to historic dramas targeting older audiences, with very distinct vocabulary used. In this paper, we explore the use of everyday language corpora in psycholinguistic applications. In an effort to address the lack of reliable everyday language corpora for English, we create SubIMDB, the first structured corpus of subtitles in the literature. SubIMDB is composed of subtitles of movies and series written for the "average audience", and can be downloaded in useful formats. In the sections that follow, we describe the resources and procedures used to build SubIMDB, and evaluate its performance in various tasks.

2 Building SubIMDB

Our goal in creating SubIMDB was to compile and provide freely a large, structured corpus of everyday language. As a data type, we chose subtitles of movies and series, since they are available for dozens of languages. Another advantage of using subtitles as opposed to, for example, chat logs or podcast transcripts, is that movies and series are subject to production standards, and hence the subtitles created for them tend to be composed of linguistically correct constructs.

2.1 Acquiring Subtitles

To create a reliable corpus of subtitles one must take into account that movies and series can be of many different genres, and may target very distinct audiences. The compilation of $\text{SUBTLEX}_{\text{us}}$ involved the download of 8,388 subtitles of U.S films and series released between 1900-2007, with no restriction with respect to genre. We took a different approach when creating SubIMDB. We use OpenSubtitles[1] as a data source. One can download subtitles from their API by providing with a production's unique IMDb[2] identifier.

As the first step in creating SubIMDB, we queried the IMDb platform searching for identifiers of six types of content: family movies, family series, comedy movies, comedy series, movies for children and series for children. We chose these genres because productions of this kind tend to target viewers of either young or all ages, and hence tend to use accessible language. Our hypothesis is that word usage statistics from this type of content correlate better with psycholinguistic properties of words, such as lexical decision times and age of acquisition.

To obtain the identifiers, we used the IMDb engine[3] to search for and parse all pages under the family and comedy feature film pages, as well as the ones under the family and comedy series categories. Since IMDb does not contain a category specific for children movies and series, we resorted to 15 movies and series lists created by IMDb users to obtain them. In total, we obtained the IMDb identifiers of 9,709 family movies, 8,008 family series, 66,411 comedy movies, 24,776 comedy series, 745 children movies and 124 children series.

We then queried the online OpenSubtitles API for each of these 109,773 IMDb identifiers. Surprisingly, we were only able to find subtitles for 12,618 movies and series. On the other hand, since series are comprised of various episodes, we downloaded subtitles for each episode of every season available in OpenSubtitles. A total of 38,102 subtitles were collected in this way.

2.2 Processing Subtitles

In order to make their content more easily accessible, we first tokenized all lines in the subtitles and removed any HTML tags. A filtering algorithm was then applied to discard subtitle lines which:

1. **Refer to metadata or timing indicators**: These lines do not contain meaningful information.

2. **Have more than 80 characters**: In most cases, lines with close to or more than 80 characters are composed of sequences of random spurious characters.

[1] http://www.opensubtitles.org
[2] http://www.imdb.com
[3] http://www.imdb.com/search

Size	HF	LF	All
10M	$\mathbf{-0.393}$	-0.391	-0.576
20M	-0.393	-0.433	-0.601
30M	-0.392	-0.454	-0.613
40M	-0.390	-0.471	-0.624
50M	-0.391	-0.465	-0.620

Size	HF	LF	All
60M	-0.390	-0.468	-0.622
70M	-0.390	-0.469	-0.622
80M	-0.391	-0.470	-0.623
90M	-0.392	-0.470	-0.623
100M	-0.392	$\mathbf{-0.471}$	$\mathbf{-0.624}$

Table 1: Pearson correlation of decision times and high and low frequency words per corpus size.

3. **Have at least one word with more than 15 characters**: Lines with unusually long words tend to be incorrectly formatted sentences.

4. **Contain advertisement**: These lines refer to credits attributed to the creators of the subtitles in question. Some examples of expressions targeted are "synched by" and "opensubtitles.org".

The resulting corpus contains 225,847,810 words in 38,643,849 lines, which is 4.5 times bigger than $\text{SUBTLEX}_{\text{us}}$.

2.3 Reliability Assessment

One of the most popular strategies for frequency norm quality assessment is to evaluate how well they predict lexical decision times. A very popular task in the field of Psycholinguistics, *lexical decision*, also known as *lexical reaction time*, refers to the process of deciding whether or not a given sequence of characters is a real word of the language in question (Balota et al., 2007). Previous work has measured the time taken by subjects to make such a decision for certain words, then used correlation metrics to assess how well their frequencies can predict them (Balota et al., 2004; Van Heuven et al., 2014; Vega et al., 2011; Brysbaert and New, 2009). In this section, we evaluate the reliability of SubIMDB by replicating some of the lexical decision experiments of Brysbaert and New (2009) and Burgess and Livesay (1998).

Brysbaert and New (2009) reveal that the size of a spoken text corpus plays a role in its utility. In general, but not always, larger corpora tend to capture psycholinguistic properties of words more effectively, given that they tend to feature a broader vocabulary and a wider array of distinct contexts from which to extract word usage statistics. But going beyond the "the bigger, the better" assumption, Burgess and Livesay (1998) propose that the ideal corpus size depends on the frequency of the words which one aims to predict the lexical decision times for.

To replicate their experiments, we first sample SubIMDB in portions containing 10 to 100 million words from sentences selected at random. As our test set, we use the MRC psycholinguistic Database (Coltheart, 1981), which provides lexical decision times for 40,468 words. Like in (Brysbaert and New, 2009), we consider only the subset of 38,130 lowercase words in order to avoid most abbreviations and proper nouns.

We split these 38,130 words in two sets: high and low frequency words. A word is considered high frequency (HF) if it is among the 1% most frequently occurring words in SubIMDB, otherwise, it is considered low frequency (LF). This methodology resembles that of Burgess and Livesay (1998). The Pearson correlation between word frequencies and lexical decision times for each corpus size are presented in Table 1.

The scores support the hypothesis in (Burgess and Livesay, 1998): the Pearson correlation for high frequency words peaks at 10 million words, while the correlation for low frequency words continuously grows from 10 to 100 million words. The increase in corpus size reflects positively on the overall performance of SubIMDB for all words, contradicting the results obtained by Brysbaert and New (2009), which suggest that a corpus does not need to have more than 16 million words in order to be cost effective.

As discussed in (Hauk and Pulvermüller, 2004), word length can also influence lexical decision times. Intuitively, one would expect to take longer to read a ten character word than a three character word, for example. Inspecting the word frequencies from SubIMDB, we found that larger words tend to benefit from larger corpora. Table 2 shows the Pearson correlation scores with lexical decision times obtained by SubIMDB samples in different sizes with respect to word length in characters.

	2	3	4	5	6	7	8	9
Count:	38M	85M	77M	18M	11M	8M	5M	3M
10M	**−0.736**	**−0.591**	−0.606	−0.576	−0.552	−0.529	−0.498	−0.455
20M	−0.728	−0.580	−0.608	−0.584	−0.564	−0.545	−0.522	−0.482
30M	−0.727	−0.584	−0.612	**−0.588**	−0.571	−0.556	−0.531	−0.498
40M	−0.716	−0.586	**−0.617**	−0.570	−0.559	−0.546	−0.532	−0.506
50M	−0.723	−0.583	−0.615	−0.583	−0.571	−0.556	−0.535	−0.505
60M	−0.721	−0.581	−0.615	−0.582	**−0.572**	**−0.557**	−0.536	−0.506
70M	−0.712	−0.579	−0.616	−0.581	−0.570	−0.554	**−0.537**	−0.506
80M	−0.713	−0.579	−0.617	−0.581	−0.569	−0.554	−0.535	**−0.508**
90M	−0.714	−0.581	−0.617	−0.579	−0.568	−0.552	−0.536	−0.508
100M	−0.714	−0.581	−0.617	−0.578	−0.568	−0.553	−0.537	−0.508

Table 2: Pearson correlation of decision times and word size per corpus size. Columns represent word length, rows represent corpus size, and cells depict Pearson correlation scores.

Table 2 shows that the scores for long words tend to require larger corpora. This could be explained by the hypothesis of Burgess and Livesay (1998), since the words' length and frequency in SubIMDB are inversely proportional. As illustrated in the second row of the table, shorter words occur much more frequently than longer words in SubIMDB.

Our findings also agree with the ones of Brysbaert and New (2009), who observed that, contrary to norms obtained from news articles and web content, spoken language text norms are better at predicting lexical decision times for shorter words. Notice that, while the correlation for shorter words tend to peak around -0.6, the correlation for longer words peaks around -0.5. This also applies to words with lengths beyond 9 characters: at around 15 characters, correlation values peak around -0.3.

Table 3 shows Pearson correlation scores for the HAL (Burgess and Livesay, 1998) and SubIMDB corpora with respect to word length. Unlike SubIMDB, the HAL corpus is composed of news articles. Much like what is observed in (Brysbaert and New, 2009), while the SubIMDB norm considerably outperforms HAL for words with 2-4 characters, the HAL corpus gives more reliable norms for words with 5+ characters. The reason behind SubIMDB's disadvantage with longer words is explained by the fact that words with 2-4 characters compose 80% of SubIMDB's content. Although this observation may seem puzzling at first, it can be easily explainable. Take, for an example, the sentences "what have you done?" and "what do you mean?". Both these sentences are composed entirely of words between two and four characters, and occur very frequently in SubIMDB. Other notable examples are "come on", "I got to go now" and "have a good one". This difference between HAL and SubIMDB suggests that combining frequency norms from different sources could be a good way of creating even more reliable norms.

	2	**3**	**4**	**5**	**6**	**7**	**8**	**9**
HAL	−0.660	−0.543	−0.598	**−0.604**	**−0.605**	**−0.584**	**−0.574**	**−0.544**
SubIMDB	**−0.716**	**−0.586**	**−0.616**	−0.570	−0.559	−0.546	−0.532	−0.506
	••	•••	•••	•••	•••	•••	•••	•••

Table 3: Correlation comparison between frequencies and decision times on a word size basis. The last line indicates a statistically significant difference with SubIMDB given $p < 0.1$ (•), $p < 0.01$ (••) or $p < 0.001$ (• • •) (F-test).

In sections to come, we compare the performance of SubIMDB and numerous other corpora in various psycholinguistic tasks.

3 Predicting Lexical Decision Times

In this experiment we assess how well frequencies from different sets of SubIMDB subtitles fair against other well-known corpora in how they correlate with lexical decision times. For this experiment, we extracted word frequencies from various SubIMDB subcorpora, as shown in Table 4.

All SubIMDB (SubIMDB)	All Comedy content (SubCOM)	Comedy movies (SubCOM-M)
All movies (SubMOV)	All children content (SubCHI)	Comedy series (SubCOM-S)
All series (SubSER)	Family movies (SubFAM-M)	Children movies (SubCHI-M)
All Family content (SubFAM)	Family series (SubFAM-S)	Children series (SubCHI-S)

Table 4: Subcorpora from SubIMDB used to predict lexical decision times

We compare ours to six frequency norms:

- **KF**: Oldest and most widely used frequency norm, calculated over the Brown corpus (Rudell, 1993; Francis and Kucera, 1979).
- **HAL**: *Hyperspace Analogue to Language* word frequency norm, calculated over the HAL corpus, which contains over 131 million words from Usenet newsgroups (Burgess and Livesay, 1998).
- **Wiki**: Word frequencies from Wikipedia, with 97 million words (Kauchak, 2013).
- **SimpleWiki**: Word frequencies from Simple Wikipedia, with 9 million words (Kauchak, 2013).
- **SUBTLEX**: Word frequencies from $\text{SUBTLEX}_{\text{us}}$, with 51 million words (Brysbaert and New, 2009).
- **Open2016**: Word frequencies from OpenSubtitles2016, with 2 billion words (Lison and Tiedemann, 2016).

We regularise all norms using Equation 1, in which f is the frequency norm value of a word w. This transformation has shown to best represent the relationship between word frequencies and lexical decision times (Balota et al., 2004).

$$norm(f(w)) = \log_{10}(f(w) + 1) \tag{1}$$

We use the same lexical decision dataset from our previous experiments as our test set. The results in Table 5 reveal that, while SubIMDB in its entirety yields the highest Spearman (ρ) correlation scores, the SubMOV corpus, which contains only subtitles of movies, yields the highest Pearson (r) correlation. F-tests show a statistically significant difference between frequencies from SubIMDB and all other corpora.

Norm	Size	ρ	r	F-test
KF	1M	-0.517	-0.486	•••
HAL	131M	-0.641	-0.616	•••
Wiki	97M	-0.531	-0.506	•••
SimpleWiki	9M	-0.560	-0.530	•••
SUBTLEX	62M	-0.653	-0.619	•••
Open2016	2B	-0.657	-0.602	•••
SubIMDB	225M	$\mathbf{-0.659}$	-0.624	-
SubMOV	125M	-0.657	$\mathbf{-0.626}$	•••
SubSER	100M	-0.652	-0.620	•••

Norm	Size	ρ	r	F-test
SubFAM	34M	-0.649	-0.614	•••
SubCOM	199M	-0.657	-0.624	•••
SubCHI	17M	-0.634	-0.592	•••
SubFAM-M	17M	-0.640	-0.596	•••
SubFAM-S	17M	-0.632	-0.590	•••
SubCOM-M	107M	-0.655	-0.623	•••
SubCOM-S	91M	-0.651	-0.618	•••
SubCHI-M	8M	-0.625	-0.572	•••
SubCHI-S	8M	-0.606	-0.556	•••

Table 5: Lexical decision prediction correlation scores. The last column indicates a statistically significant difference with SubIMDB given $p<0.1$ (•), $p<0.01$ (••) or $p<0.001$ (•••) (F-test).

Unlike what was reported in (Brysbaert and New, 2009), the HAL norm achieved lower correlation scores than the SUBTLEX norm, despite the fact that the HAL corpus is twice as large as $\text{SUBTLEX}_{\text{us}}$. This contrast highlights the potential of spoken language corpora in lexical decision prediction.

Our results also indicate a poor performance for the Kucera-Francis coefficient. Despite its use in numerous previous contributions (Burgess and Livesay, 1998; Zevin and Seidenberg, 2002; Brysbaert and New, 2009), more modern resources proved more effective. We believe this is caused by the fact that these coefficients are calculated from a corpus that is very small when compared to the other resources presented in this paper.

4 Predicting Psycholinguistic Properties

In addition to lexical decision times, other psycholinguistic properties of words have been studied in terms of their correlation with frequency norms (Paetzold and Specia, 2016a). In this experiment, we evaluate how well the norms described in Section 3 correlate with four psycholinguistic properties extracted from the MRC psycholinguistic Database:

- **Familiarity**: Available for 9,392 words – frequency with which a word is seen, heard or used daily.
- **Age of Acquisition**: Available for 3,503 words – age at which a word is learned.
- **Concreteness**: Available for 8,228 words – how "palpable" the object the word refers to is.
- **Imagery**: Available for 9,240 words – intensity with which a word arouses images.

The results in Table 6 reveal that SubFAM-M (family movies) performs better than all other norms in predicting age of acquisition and concreteness, although it is 117 times smaller than OpenSubtitles2016 (Open2016). F-tests reveal a statistically significant difference between SubIMDB and all other corpora.

		Age of Acquisition		Familiarity		Concreteness		Imagery	
	Size	r	F-test	r	F-test	r	F-test	r	F-test
KF	1M	-0.447	•••	0.669	•••	$\mathbf{-0.180}$	•••	-0.045	•••
HAL	131M	-0.511	•••	0.732	•••	-0.064	•••	0.086	•••
Wiki	97M	-0.412	•••	0.676	•••	-0.043	•••	0.084	•••
SimpleWiki	9M	-0.486	•••	0.667	•••	0.011	•••	0.129	•••
SUBTLEX	62M	-0.676	•••	0.774	•••	0.017	•••	0.190	•••
Open2016	2B	-0.666	•••	**0.799**	•••	-0.003	•••	0.185	•••
SubIMDB	225M	-0.698	-	0.781	-	0.037	-	0.213	-
SubMOV	125M	-0.705	•••	0.777	•••	0.031	•••	0.212	•••
SubSER	100M	-0.687	•••	0.777	•••	0.038	•••	0.207	•••
SubFAM	34M	-0.723	•••	0.758	•••	0.038	•••	0.217	•••
SubCOM	199M	-0.696	••	0.781	•••	0.037	•••	0.211	•••
SubCHI	17M	-0.709	•••	0.735	•••	0.028	•••	0.201	•••
SubFAM-M	17M	$\mathbf{-0.746}$	•••	0.742	•••	0.043	•••	**0.220**	•••
SubFAM-S	17M	-0.685	•••	0.743	•••	0.007	•••	0.178	•••
SubCOM-M	107M	-0.698	•••	0.777	•••	0.027	•••	0.207	•••
SubCOM-S	91M	-0.690	•••	0.777	•••	0.042	•••	0.209	•••
SubCHI-M	8M	-0.728	•••	0.723	•••	0.026	•••	0.191	•••
SubCHI-S	8M	-0.670	•••	0.704	•••	-0.006	•••	0.158	•••

Table 6: Pearson correlation of norms with respect to psycholinguistic properties. Columns following correlation scores indicate a statistically significant difference with SubIMDB given $p<0.1$ (•), $p<0.01$ (••) or $p<0.001$ (•••) (F-test).

Perhaps most surprising is the performance of the SubIMDB subset of children movies (SubCHI-M) in predicting age of acquisition. Despite its small size, its performance is still much superior than almost

all corpora, including OpenSubtitles2016, which is over 250 times larger. Comparing word frequencies from SubCHI-M with the ones in OpenSubtitles2016, we found interesting differences. Table 7 shows the most over and underrepresented words in SubCHI-M based on percentages of variance with respect to OpenSubtitles2016.

It can be noticed that while overrepresented words ("turtles", "hedgehog", etc.) are mostly innocent in nature, underrepresented words describe mostly sexual and/or thought-provoking concepts ("vagina", "abortion", etc.). These differences reveal that, although subtitle corpora may share traits in general, the domain from which the subtitles are extracted plays an important role. This highlights the often disregarded advantages of a structured, raw text subtitle corpora like the one we collected here. By making subtitles available in their raw form along with metadata about their source of origin, future research can explore different ways of building the ideal corpus for a given task, e.g. by employing clever subtitle selection and filtering techniques.

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Over	hoagy	flintstone	turtles	potter	fantasia	hedgehog	hiccup	dialogue
Under	vagina	abortion	cartel	intercourse	rapist	overdose	porn	pimp

Table 7: Representation contrast between the SubCHI-M and OpenSubtitles2016 corpora.

Inspecting our data, we also found further evidence that, unlike what was found by Brysbaert and New (2009), it is unfeasible to predict the ideal size of a corpus by simply looking at frequency correlation with lexical decision times. Table 8 illustrates Pearson correlation scores of different SubIMDB sample sizes for all aforementioned psycholinguistic properties. The correlation scores all behave differently: while familiarity benefits from larger corpora, the remaining properties do not.

	Age of Acquisition		Familiarity		Concreteness		Imagery	
	ρ	r	ρ	r	ρ	r	ρ	r
10M	-0.686	-0.703	0.770	0.724	0.067	0.018	0.225	0.186
20M	$\mathbf{-0.691}$	$\mathbf{-0.711}$	0.782	0.745	$\mathbf{0.072}$	$\mathbf{0.032}$	$\mathbf{0.234}$	0.207
30M	-0.688	-0.710	0.796	0.761	0.070	0.031	0.233	0.211
40M	-0.677	-0.698	0.804	0.768	0.069	0.030	0.227	$\mathbf{0.213}$
50M	-0.683	-0.706	0.805	0.769	0.066	0.030	0.229	0.211
60M	-0.680	-0.703	0.808	0.772	0.065	0.030	0.229	0.211
70M	-0.679	-0.701	0.809	0.772	0.063	0.028	0.228	0.209
80M	-0.678	-0.701	0.811	0.774	0.063	0.028	0.227	0.209
90M	-0.677	-0.700	0.811	0.774	0.063	0.029	0.227	0.210
100M	-0.676	-0.700	$\mathbf{0.811}$	$\mathbf{0.775}$	0.063	0.029	0.227	0.210

Table 8: Pearson correlation per corpus size for different psycholinguistic properties.

5 Predicting Simplicity

Everyday language corpora can also be useful in predicting word simplicity. In this experiment, we evaluate how well SubIMDB fairs against other corpora when employed as a solution to Lexical Simplification.

As our test set, we use the one from the English Lexical Simplification task of SemEval 2012, which contains 1,710 instances composed of a sentence, a target word, and candidate substitutions ranked by simplicity. This dataset has been widely used and hence allows the comparison of SubIMDB against state-of-the-art solutions for the task. For evaluation, we use Spearman (r) and Pearson (ρ) correlation, as well as the TRank metric proposed by Specia et al. (2012), which measures the rate with which a candidate substitution with the highest gold rank i.e. the simplest, was ranked first by the system.

We compare the performance of all frequency norms described in Section 3 to Google 1T, a corpus composed of over 1 trillion words (Evert, 2010), and the winner system in the SemEval 2012 task, which

Norm	r	ρ	TRank	F-test
KF	0.619	0.626	0.589	•••
HAL	0.630	0.633	0.598	•••
Wiki	0.575	0.583	0.516	•••
SimpleWiki	0.626	0.632	0.570	•••
SUBTLEX	0.649	0.649	0.619	•••
Open2016	0.650	0.647	0.619	•••
SubIMDB	0.654	0.652	0.622	-
SubMOV	**0.660**	**0.658**	**0.623**	•••
SubSER	0.648	0.647	0.619	•••
SubFAM	0.649	0.650	0.615	•••

Norm	r	ρ	TRank	F-test
SubCOM	0.655	0.653	**0.623**	•
SubCHI	0.643	0.645	0.611	•••
SubFAM-M	0.653	0.653	0.618	•••
SubFAM-S	0.647	0.650	0.620	•••
SubCOM-M	**0.660**	**0.658**	**0.623**	•••
SubCOM-S	0.647	0.648	0.618	•••
SubCHI-M	0.650	0.654	0.600	•••
SubCHI-S	0.640	0.644	0.608	•••
Google 1T	N/A	N/A	0.585	-
Best SemEval	N/A	N/A	0.602	-

Table 9: Correlation and TRank scores for frequency norms with respect to simplicity. The fifth column indicates a statistically significant difference with SubIMDB given $p<0.1$ (•), $p<0.01$ (••) or $p<0.001$ (•••) (F-test).

employs a Support Vector Machine ranker that uses a wide array of features (Jauhar and Specia, 2012). The results in Table 9[4] reveal that SubIMDB outperforms all baselines, including Google 1T and the former state-of-the-art for the task in TRank. Nonetheless, some SubIMDB subcorpora are even more effective than using our corpus in its entirety, despite being much smaller.

Work in Text Simplification has, however, explored more than single-word frequency norms, considering for example raw n-gram frequencies and language model probabilities (Horn et al., 2014; Baeza-Yates et al., 2015; Paetzold and Specia, 2016b). Table 10 shows TRank scores obtained on the SemEval 2012 task when using 3-gram and 5-gram raw frequencies and language model probabilities extracted from various corpora. The 3-grams and 5-grams consist in a candidate substitution surrounded by one and two tokens, respectively. For probabilities, we trained 5-gram language models using SRILM (Stolcke, 2002). For the Kucera-Francis (KF) norms we use the Brown corpus (Francis and Kucera, 1979). The HAL corpus is not available for download and hence it could not be tested here.

Table 10 shows that single word frequencies are more effective than both 3-grams or 5-grams in the SemEval 2012 task. We believe that the reason for this lies in the fact that almost all candidate substitutions in each instance of the dataset perfectly fit the context in which the target word was found, both with respect to grammaticality and meaning preservation. This setup disregards the need to account for context, which hence makes the use of n-grams less crucial. Since the representative sparsity of a corpus inherently grows as sequences of words become longer, n-grams with $n \geq 1$ are consequently much less reliable than single-word frequencies for this task in particular. This hypothesis is also supported by the fact that 3-gram frequencies achieved considerably higher scores than 5-gram frequencies.

Nonetheless, there is a clear advantage to using language model probabilities as opposed to raw frequencies for larger n-grams, since language models employ sophisticated smoothing techniques to reduce issues due to sparsity. These findings highlight again how important it is for corpora to be released in raw format to make it possible to train language models.

6 Conclusions

In this paper we presented a study on the application of everyday language corpora in the prediction of psycholinguistic properties of words. For our experiments, we created SubIMDB: a large structured corpus of subtitles of movies and series for the average audience. It contains 38,102 subtitles, each individually annotated with metadata about the movie or series for which they were created. Altogether, our corpus has 225,847,810 words in 38,643,849 lines, which is 4.5 times larger than the widely used $\text{SUBTLEX}_{\text{us}}$ corpus (Brysbaert and New, 2009).

We found that word frequencies from SubIMDB capture lexical decision times more effectively than various other frequency norms. Additionally, we found that using only certain types of subtitles can yield

[4]Specia et al. (2012) only provides results for TRank.

		Frequency				Probability			
		3-grams		5-grams		3-grams		5-grams	
Norm	Size	TRank	F-Test	TRank	F-Test	TRank	F-Test	TRank	F-Test
KF	1M	0.234	●●●	0.234	●●●	0.234	●●●	0.234	●●●
Wiki	97M	0.388	○	0.257	○	0.528	●●●	0.520	●●●
SimpleWiki	9M	0.354	●●●	0.247	●●●	0.557	●●●	0.560	●●●
SUBTLEX	62M	0.402	●●●	0.261	●	**0.588**	○	0.586	○
Open2016	2B	**0.461**	●●●	0.234	●●●	0.564	○	0.550	○
SubIMDB	225M	0.425	-	**0.264**	-	0.582	-	0.564	-
SubMOV	125M	0.401	●●	0.262	○	0.582	○	0.580	○
SubSER	100M	0.399	●●●	0.254	●	0.575	●●●	0.567	●●●
SubFAM	34M	0.379	●●●	0.251	●●	0.577	○	0.569	○
SubCOM	199M	0.416	○	0.261	○	0.577	●●●	0.566	●●●
SubCHI	17M	0.354	●●●	0.246	●●●	0.572	○	0.572	○
SubFAM-M	17M	0.357	●●●	0.248	●●●	0.589	○	**0.587**	○
SubFAM-S	17M	0.364	●●●	0.246	●●●	0.574	○	0.574	○
SubCOM-M	107M	0.398	●●●	0.259	●	0.582	●●●	0.572	●●●
SubCOM-S	91M	0.396	●●●	0.253	●	0.570	●●●	0.564	●●●
SubCHI-M	8M	0.329	●●●	0.242	●●●	0.572	○	0.569	●
SubCHI-S	8M	0.334	●●●	0.243	●●●	0.569	○	0.569	○

Table 10: TRank scores for n-grams. Columns following TRank scores indicate a statistically significant difference with SubIMDB given $p < 0.1$ (●), $p < 0.01$ (●●) or $p < 0.001$ (●●●) (F-test).

noticeable increase in performance. The same was observed for the prediction of other psycholinguistic properties, such as age of acquisition.

Our experiments provided evidence to support (Burgess and Livesay, 1998)'s hypothesis, which states that the ideal size of a corpus depends on the overall frequency of the words which one aims to predict lexical decision times for. Nonetheless, our results also reveal that, unlike what is claimed by Brysbaert and New (2009), one should not attempt to quantify the ideal corpus size based solely on correlation scores with lexical decision times.

Finally, we found that in English Lexical Simplification both word frequencies and language model probabilities from SubIMDB outperform the ones extracted from all other corpora available, as well as the state-of-the-art method for the task. Through these findings, we hope to encourage other researchers to collect and release corpora in more flexible, useful forms rather than simply providing with pre-computed single-word frequency counts.

In future work, we aim to add other types of subtitles to SubIMDB and to study smarter subtitle selection and filtering strategies. We also intend to study the use of other types of spoken text corpora, such as tweets and conversations from Facebook (Herdağdelen and Marelli, 2016), in improving the performance of Natural Language Processing tasks. We released the SubIMDB corpus in both raw form, containing subtitles individually annotated with metadata, and in compiled form. Both versions are freely available for download at `http://ghpaetzold.github.io/subimdb`.

Acknowledgements

This work has been partially supported by the European Commission project SIMPATICO (H2020-EURO-6-2015, grant number 692819).

References

Ricardo Baeza-Yates, Luz Rello, and Julia Dembowski. 2015. CASSA: A context-aware synonym simplification algorithm. In *Proceedings of the 2015 NAACL*, pages 1380–1385.

David A Balota and James I Chumbley. 1984. Are lexical decisions a good measure of lexical access? the role of word frequency in the neglected decision stage. *Journal of Experimental Psychology: Human perception and performance*, 10:340.

David A Balota, Michael J Cortese, Susan D Sergent-Marshall, Daniel H Spieler, and MelvinJ Yap. 2004. Visual word recognition of single-syllable words. *Journal of Experimental Psychology: General*, 133(2):283.

David A Balota, Melvin J Yap, Keith A Hutchison, Michael J Cortese, Brett Kessler, Bjorn Loftis, James H Neely, Douglas L Nelson, Greg B Simpson, and Rebecca Treiman. 2007. The english lexicon project. *Behavior research methods*, 39:445–459.

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*, 41:977–90.

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*, 30:272–277.

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

Siobhan Devlin and John Tait. 1998. The use of a psycholinguistic database in the simplification of text for aphasic readers. *Linguistic Databases*, pages 161–173.

Stefan Evert. 2010. Google web 1t 5-grams made easy (but not for the computer). In *Proceedings of the 2010 NAACL*, pages 32–40.

W Nelson Francis and Henry Kucera. 1979. Brown corpus manual. *Brown University*.

Olaf Hauk and F Pulvermüller. 2004. Effects of word length and frequency on the human event-related potential. *Clinical Neurophysiology*, 115:1090–1103.

Amaç Herdağdelen and Marco Marelli. 2016. Social media and language processing: How facebook and twitter provide the best frequency estimates for studying word recognition. *Cognitive Science*.

Colby Horn, Cathryn Manduca, and David Kauchak. 2014. Learning a Lexical Simplifier Using Wikipedia. In *Proceedings of the 52nd ACL*, pages 458–463.

S. Jauhar and L. Specia. 2012. Uow-shef: Simplex–lexical simplicity ranking based on contextual and psycholinguistic features. In *Proceedings of the 1st SemEval*, pages 477–481.

David Kauchak. 2013. Improving text simplification language modeling using unsimplified text data. In *Proceedings of the 51st ACL*, pages 1537–1546.

Pierre Lison and Jrg Tiedemann. 2016. Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles. In *Proceedings of the 10th LREC*.

Gustavo Henrique Paetzold and Lucia Specia. 2016a. Inferring psycholinguistic properties of words. In *Proceedings of the 2016 NAACL*, pages 435–440.

Gustavo Henrique Paetzold and Lucia Specia. 2016b. Unsupervised lexical simplification for non-native speakers. In *Proceedings of The 30th AAAI*.

Alexander Pak and Patrick Paroubek. 2010. Twitter as a corpus for sentiment analysis and opinion mining. In *Proceedings of 2010 LREC*, volume 10, pages 1320–1326.

Keith Rayner and Susan A Duffy. 1986. Lexical complexity and fixation times in reading: Effects of word frequency, verb complexity, and lexical ambiguity. *Memory & Cognition*, 14(3):191–201.

Allan P. Rudell. 1993. Frequency of word usage and perceived word difficulty: Ratings of kuera and francis words. *Behavior Research Methods*, pages 455–463.

Matthew Shardlow. 2013. A comparison of techniques to automatically identify complex words. In *Proceedings of the 51st ACL Student Research Workshop*, pages 103–109.

Lucia Specia, Sujay Kumar Jauhar, and Rada Mihalcea. 2012. Semeval-2012 task 1: English lexical simplification. In *Proceedings of the 1st SemEval*, pages 347–355.

Andreas Stolcke. 2002. SRILM-an extensible language modeling toolkit. In *Proceedings of the 2002 ICSLP*, pages 257–286.

Walter JB 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*, 67:1176–1190.

Fernando Cuetos Vega, María González Nosti, Analía Barbón Gutiérrez, and Marc Brysbaert. 2011. Subtlex-esp: Spanish word frequencies based on film subtitles. *Psicológica: Revista de metodología y psicología experimental*, 32:133–143.

Jason D Zevin and Mark S Seidenberg. 2002. Age of acquisition effects in word reading and other tasks. *Journal of Memory and language*, 47:1–29.