Mapping of American English vocabulary by grade levels

Michael Flor, Steven Holtzman, Paul Deane and Isaac Bejar
Educational Testing Service, Princeton, NJ, USA

GR4L2 workshop, December 7, 2021, UCLouvain
Mapping of American English vocabulary by grade levels

• What we developed
• Why it was developed
• How it was done
• Notes
• CEFR connection
What
What

We developed a mapping of American English vocabulary by U.S. school grade levels (for native English speakers)

Comprehensive coverage: 126,260 words (forms)

Range:
GL from 0 to 16 (from kindergarten till end of undergraduate studies)
Why
Why

Why express vocabulary difficulty in terms of grade levels? There already are several types of scales and approaches:

• **Word frequency** – can be established with a variety of corpora.

• Binning vocabulary into **frequency bins** of 1000 words (Waring & Nation, 1997), often also with morphological word-family relatives.

• **Age of Acquisition** (AoA) which is expressed in years of age, (Kuperman et al, 2012).

• **Word familiarity**, measured via crowdsourcing and expressed as proportion of people who know the word (Brysbaert et al, 2019).

• A variety of Computational (NLP) measures of word complexity (shared tasks: Shardlow et al., 2021, Yimam et al., 2018; etc.).
Why
Expressing vocabulary difficulty in terms of grade levels is very useful and convenient for educators!

For example, teachers and test-developers often need to check that their texts adhere to certain grade-level expectations, and decide whether or not a word is appropriate for a given grade level.

Consider statements like

“word frequency of 6 per million”, or 0.000006, or log=-5.2,
“is known by 70% of the general population”,
“is learned on average at age 8”, “is more difficult than 34% of other words”
such statements have only limited utility in many educational situations.

These are not the scales that educators are used to or are convenient for them to operate with.
Why

A grade level scale for vocabulary is what educators often need.

We are not the first in this enterprise.

A book called "EDL Core Vocabularies" has seen multiple editions from 1949 to 1989, is widely used by teachers (and at ETS). But it is old, and limited (10K words). There are other books, but often also limited in scope.

Living Word Vocabulary is an excellent resource (Dale & O’Rourke, 1981).

We felt there is a need for an up-to-date wide-coverage resource that maps words to grade levels.
How
How

What does it mean to map words to grade levels?

There are different approaches:

• Some researchers ask which words are known in different grade levels (known to a certain degree, and on average in population). LWV belongs to that camp.

• Other researchers focus on which words are expected to be known or expected to be learned in different grade levels (on average). Our study belongs in this group.
How

Our study involved two major stages:

• Empirical collection – to establish ‘ground truth’
• Statistically-based expansion (model and predict)
How

Data Collection

• We collected grade-level vocabulary lists from 70 sources, such as teachers and school websites, state education departments in USA, and published data by researchers and organizations working on vocabulary in USA.

• Overall, we collected data for >80K tokens (strings), each word with an expected grade level designator (a number, call it XGL). This reduces to 21K different word forms (types).

• Many words appear in multiple lists, with different XGL, and we generally averaged XGL values (thus we got fractional numbers, e.g. 3.56).

• From this effort we got the initial list of words mapped to grade levels, (with values in range 0-16).
How

- **A validation** study: how our grade levels compare to some other scales.

- Compared to data from Brysbaert & Biemiller (2017); 2 variables: AoAR (AoA rated) are the AoA norms from Kuperman et al. 2012; AoAT (AoA testing) is based on data from LWV.

- Original LWV has only even grades 4-6-8-10-12 (13&16), and B&B added GL2. LWV works with word senses (44K), and was reduced to 30K wordforms.

- B&B 2017 report Pearson correlation of AoAR to AoAT $r=0.757$ (n=18K)

- Our collected VXGL list has an overlap of about 15K words with those lists.

  Pearson correlations:

<table>
<thead>
<tr>
<th></th>
<th>AoAT</th>
<th>VXGL</th>
</tr>
</thead>
<tbody>
<tr>
<td>AoAR</td>
<td>0.8112</td>
<td>0.8107</td>
</tr>
<tr>
<td>AoAT</td>
<td></td>
<td>0.7792</td>
</tr>
</tbody>
</table>

- Well, it looks that our scores are comparable to the others, at least in the relative ordering of words.
How Prediction

• We have 21K words with VXGL scores, and want to expand to 126K.
• To provide coverage to many more words we use the usual statistical magic: build a prediction model with the existing data, and if it works well enough, use it to predict data for unknown cases.
• Initial prediction model uses 15K words.
• Our logic: utilize many variables in this prediction exercise, but we also want to utilize AoAR and AoAT, as strong predictors
• But for thousands of other words we won't have AoAT or AoAR scores. So we will see how we are doing with and without AoA.
How

Variables used in the prediction model. Some of them are 'usual suspects'.

- Square root of word length
- Syllable count per word (algorithmic)
- LSdiff = (num. letters – num. syllables)
- TASA corpus: Log-transformed freq. value (SFI), IDF
- Wikipedia corpus: SFI, lemmatized SFI, IDF
- Gigaword corpus: SFI
- Google Books Ngrams data (>400 billion words) (using unigrams only): IDF, SFI, lemmatized SFI
- Number of senses in WordNet
- Number of inflected word-family members
- Number of lemmas in derivational family
- What major POS the word takes (noun/verb/adjective/adverb) (binary variables)
- Associtative Estimate of GL
- AoAT and AoAR
How

Prediction Results

• We used GBM regression. N=15K words. Training: 90%, Testing: 10%.

<table>
<thead>
<tr>
<th></th>
<th>Pearson Correlation to GL</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>With AoAR and AoAT</td>
<td>0.790</td>
<td>1.412</td>
</tr>
<tr>
<td>Without</td>
<td>0.601</td>
<td>1.948</td>
</tr>
</tbody>
</table>

RMSE for ranges of true VXGL scores

<table>
<thead>
<tr>
<th>GL ≤ 3.5</th>
<th>3.5 – 6.5</th>
<th>6.5 – 9.5</th>
<th>&gt; 9.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>With AoAR and AoAT</td>
<td>0.996</td>
<td>1.340</td>
<td>1.266</td>
</tr>
<tr>
<td>Without</td>
<td>1.727</td>
<td>1.817</td>
<td>1.398</td>
</tr>
</tbody>
</table>
How

Associative Estimate of Grade Level

A version of the Distributional Hypothesis:

A word's GL can be known by the GLs of words with which it keeps company.

• ‘the company’: take a fixed set of common words, for which VXGL are known.
• To predict VXGL for a new word, check it's co-occurrence (in large corpora) with the reference ‘company’, and use the average (mean) value as predictor.
• We used: Longman Dictionary of Contemporary English (LDOCE) defining vocabulary list (2000 words) and the Oxford 3000 "most important words". The unified list has 3259 unique word forms for which we have VXGL values.
• For association measure we used mutual conditional probability, of the form:

  \[ MCP = \frac{p(A,B)}{P(A) \times P(B)} \]

• The data comes from a word-to-word co-occurrence database trained on Wikipedia (co-occurrence within paragraphs – Flor & Beigman Klebanov, 2014)
How

**Associative Estimate of Grade Level**

- And thus for a new word we take average of 3259 MCP values
- For the Unified List (the reference set itself), we check self-consistency: predict VXGL by using AEGL of all other words on this list (leave-one-out). Pearson correlation of predicted value with VXGL is **0.738**.

- AEGL is a decent VXGL predictor: For 15K words, Pearson correlation of predicted value with VXGL is **0.392**.

- And it is not correlated with frequency: Correlation with Google-Books Lemmatized SFI is **0.037**.
Notes

Limitations

• We should improve ‘gold’ VXGL values by collecting more data.
• We should improve the prediction model to further reduce RMSE.
• Our data is per word-forms, not senses ...
• We did not include multi-word expressions yet.

But even the current version is useful and is already utilized at ETS
Example use of VXGL data:

• Coloring words in text by grade level (rounded VXGL values)

Crop Rotation

Conventional farmers can grow the same crop year after year on the same piece of land. If they lose specific nutrients from the soil, they can add chemical fertilizers. But organic farmers don’t use chemical fertilizers. In addition to relying on natural fertilizers, farmers rely on a technique called crop rotation to keep soil healthy. According to Susan Windmere, an expert on crop rotation techniques, Crop rotation means changing the kind of crop that is planted in a plot of land after one or two harvests. The idea behind crop rotation is to make sure that one kind of crop doesn’t use up all of the nutrients in the soil by growing in the same place for many years in a row.

The diagram to the right shows a simple example of how an organic farmer might rotate crops to keep the soil healthy. In Year 1 in Plot 1, the farmer plants corn. Corn uses up a lot of nitrogen to grow. In Year 1 in Plot 2, the farmer plants peas, which release nitrogen into the soil. The next year, the farmer switches, or rotates, the crops that are planted in each plot. The farmer plants the corn in Plot 2, where last year’s peas left plenty of nitrogen in the soil. The farmer plants peas in Plot 1 so that nitrogen will be released into the soil. In Year 3, corn can go back in Plot 1. Peas go back in Plot 2.

Legend:
Colors of vocabulary grade levels: 1 2 3 4 5 6 7 8 9
CEFR connection
How does VXGL data relate to CEFR?

- We used the English Profile CEFR vocabulary (American English version)
- We have 6313 single words in common with VXGL
- For each word, we used the lowest available CEFR level.

- Pearson correlation of VXGL to CEFR is 0.63
- Wikipedia IDF values for same set correlate to CEFR at 0.46.
- Adding VXGL and Wikipedia IDF values into a linear regression raises the correlation with CEFR to 0.67 (adjusted R-square is 0.45).
Thank you

contact: mflor@ets.org
References


English Vocabulary Profile Online. [https://www.englishprofile.org/wordlists/evp](https://www.englishprofile.org/wordlists/evp)


