Accurate Measurement of Lexical Sophistication with Reference to ESL Learner Data

Ben Naismith
University of Pittsburgh
Department of Linguistics
4200 Fifth Ave, Pittsburgh, PA 15260
1-412-624-5900
bnaismith@pitt.edu

Na-Rae Han
University of Pittsburgh
Department of Linguistics
4200 Fifth Ave, Pittsburgh, PA 15260
1-412-624-5900
naraehan@pitt.edu

Alan Juffs
University of Pittsburgh
Department of Linguistics
4200 Fifth Ave, Pittsburgh, PA 15260
1-412-624-5900
juffs@pitt.edu

Brianna Hill
University of Pittsburgh
School of Computing and Information
4200 Fifth Ave, Pittsburgh, PA 15260
1-412-624-5900
blh82@pitt.edu

Daniel Zheng
University of Pittsburgh
Department of
Electrical and Computer Engineering
4200 Fifth Ave, Pittsburgh, PA 15260
1-412-624-5900
daniel.zheng@pitt.edu

ABSTRACT

One commonly used measure of lexical sophistication is the Advanced Guiraud (AG; [9]), whose formula requires frequency band counts (e.g., COCA; [13]). However, the accuracy of this measure is affected by the particular 2000-word frequency list selected as the basis for its calculations [27]. For example, possible issues arise when frequency lists that are based solely on native speaker corpora are used as a target for second language (L2) learners (e.g., [8]) because the exposure frequencies for L2 learners may vary from that of native speakers. Such L2 variation from comparable native speakers may be due to first language (L1) culture, home country teaching materials, or the text types which L2 learners commonly encounter. This paper addresses the aforementioned problem through an English as a Second Language (ESL) frequency list validation. Our validation is established on two sources: (1) the New General Service List (NGSL; [4]) which is based on the Cambridge English Corpus (CEC) and (2) written data from the 4.2 million-word Pitt English Language Institute Corpus (PELIC). Using open-source data science tools and natural language processing technologies, the paper demonstrates that more distinct measurable lexical sophistication differences across levels are discernible when learner-oriented frequency lists (as compared to general corpora frequency lists) are used as part of a lexical measure such as AG. The results from this research will be useful in teaching contexts where lexical proficiency is measured or assessed, and for materials and test developers who rely on such lists as being representative of known vocabulary at different levels of proficiency. This research applies data-driven exploration of learner corpora to vocabulary acquisition and pedagogy, thus closing a loop between educational data mining and classroom applications.

Keywords

Advanced Guiraud, corpus linguistics, English as a Second Language, ESL, learner corpora, lexical sophistication, vocabulary lists

1. INTRODUCTION

An enduring concern of researchers in second language (L2) vocabulary development is the basic set of words learners should know; moreover, having acquired this vocabulary, what kinds of intervention are best for promoting acquisition of the additional words that learners need in order to function professionally and academically [8, 23]? Thus, establishing the correct set of basic words that learners already know is important to be able to measure subsequent development in productive vocabulary knowledge. In order to accurately track the acquisition of new vocabulary over time, researchers have focused on quantitative measures that can be used to examine different aspects of the 'lexical richness' of learner output, including *lexical diversity*, which uses text internal measures such as VocD (D) and MTLD (e.g., [17, 21]); *lexical sophistication*, which makes reference to frequencies in corpora with measures like the Advanced Guiraud (AG) (e.g., [10, 28]); and *lexical depth*, which measures knowledge of usage (e.g., [6, 11]). In this paper, we focus on lexical sophistication because (1) the calculation of AG depends on the establishment of the correct set of high-frequency words that the learners may (already) know; (2) the frequency bands of 3000-9000 words are lexical items that researchers advocate should be the focus of instruction [25]; and (3) teacher perceptions of lexical proficiency have been shown to correlate strongly with lexical sophistication [10].

2. LITERATURE REVIEW

Vocabulary knowledge in a second language is a vital component in the development of L2 proficiency [23]. As a result, accurate measurement of vocabulary is important for all language learning stakeholders including learners, teachers, material developers, developers of standardized tests, and educational institutions. One common context of English as a Second Language (ESL) learning, and that of this study, is in tertiary education intensive English programs (IEPs). Most students entering IEPs already know some English, typically placing at the low-intermediate level and above. As a corollary, learners are expected to already know high-frequency English vocabulary such as the first 2000 words of the New General Service List (NGSL; [4]).

The stakes are high in that most students have a short time to prepare for academic work, and as such, the targeting of instruction to students' needs is important. Yet, this task is difficult for teachers because the first languages (L1s) of the students vary, and students may in fact not know all of the basic words assumed by frequency lists of basic vocabulary. Such lack of certainty makes measuring vocabulary development beyond the basic list challenging because at the higher levels learners may not be given credit for acquiring high-frequency words they are assumed to know, but in fact do not control in their productive lexicon. In contrast, low-frequency words that they already know, based on their own cultural or educational background, may wrongly be treated as newly acquired. This issue reflects a general concern that materials written for learners may not consider broader linguistic needs of the students [18] and that frequencies from large corpus analyses may not always reflect linguistic challenges (e.g., [16]).

The literature on vocabulary development has shown that Advanced Guiraud (AG) can be an effective method of measuring of lexical sophistication [12, 19], but may not always reflect development [11]. In essence, AG is a form of Type/Token ratio (TTR) [28] with two key differences. First, it takes as the denominator the square root of the total tokens, a measure designed to neutralize TTR's sensitivity to text length. Second, types that are very frequent, for example the 2000 most frequent words on the NGSL, are removed from the total types [28, 12]. As a result, AG incorporates frequency information, while other measures do not.

In [12], Daller and Xue compared two groups of Chinese-speaking learners, one in China and the other in the UK. They found that Guiraud (all types/√tokens) and AG were both effective at distinguishing the China group from the UK group, whose mean (stdev) AG scores were 0.72 (.2) and 0.94 (.29) respectively. However, when Daller et al. [11] investigated the longitudinal development of 42 Arabic-speaking ESL learners, the values of AG were low and increased minimally, ranging from an average of about 0.20 to 0.25 [11]. In neither study was the composition of the AG list of 2000 basic types specified, referred to only as 'the 2000 frequency band.' Considering, as [16] says, that the needs of the users should be accounted for when replicating a word list, knowing such information would be of great use to researchers seeking to evaluate and replicate previous results.

Supporting Daller and Xue's findings, Juffs [19] analyzed a subset of the Pitt English Language Institute Corpus (PELIC) data. He found that AG (using the 2000 frequency bands of the BNC-COCA at http://lextutor.ca as a lexical sophistication metric) was a better measure than D (a lexical diversity metric) in distinguishing progress in lexical development of Arabic, Chinese, and Korean learners who studied throughout the upper-intermediate (level 4) and advanced (level 5) levels in the Pitt IEP. Juffs found that the level 4 learners' AG scores ranged from 1.32 to 1.53 on average, whereas the level 5 learners' scores ranged from 1.90 to 2.12. However, Juffs' study, while suggestive, only included 254,055 tokens and did not fully utilize PELIC's written sub-corpus which actually consists of more than 4.2 million tokens when all L1s are included.

The studies reviewed here demonstrate large variability in terms of how frequency data are measured and collected. Not only are the 2000-word lists for AG inconsistent or unknown across studies, but so too is the definition of the 'types' which form the basis of many lexical measures. Although a full discussion of this area is beyond the scope of this paper (see, e.g., [22]), it directly impacts all measures using frequency lists. On one end of the spectrum, measurements such as TTR count types mechanically without grouping different forms in anyway, so that 'dog' and 'dogs' would be counted as two distinct types. In this approach, the value lies in the ease with which data can be analyzed automatically with no need for human judgements. However, should a learner who produces 'mango' and 'mangos' be said to have the same lexical range as someone who produces 'mango' and 'pomegranate', or can we assume that the latter student will also know the plural forms?

At the other extreme, many researchers (e.g., [1]) advocate for *word families* to be the base counting unit, i.e., a word plus its derivational and inflectional forms. For example, 'happy', 'happiness', 'unhappy', and comparative 'happier', would be one unit. While this solves the previous issue, it means that a learner would not be given credit for knowing words related by derivation to a common word, with, for example, 'actresses', 'actionable', and 'inaction' all belonging to the word family 'act' (http://lextutor.ca).

A third 'middle ground' approach advocated by Schmitt [24] uses lemmas as a measurement unit. A lemma typically refers to a word plus its inflected forms only; lemma information has accompanied various resources, including the Brown Corpus and the New-GSL (not to be confused with the NGSL) [3]. Thus, 'act', 'acted', and 'acting' would be one unit, but 'act' and 'actionable' separate units. In sum, when creating a word list there are numerous decisions to make regarding not only the relative value of word frequency, range, and dispersion, but even the unit of counting must be considered and justified [16].

Given the challenges in data collection and analysis, the lack of consensus as to best practice is unsurprising. Comparisons across studies are further complicated by small sample sizes, limited L1 backgrounds, and different learning contexts, all of which threaten the external validity and thus the generalizability of the results. The reported scores in this literature do, however, give this study a range of reasonable AG scores that one might expect.

In contrast, PELIC is a multi-million-word learner corpus representing learners from different L1 backgrounds who have studied together in the same location, using similar materials, and in the same educational context. Exploiting this unique dataset, we seek to address the following research questions:

(1) How can data mining tools be applied to a learner corpus to produce effective vocabulary lists?

(2) Do the different types that are removed for the purposes of the AG have an effect on the measurement of lexical sophistication across levels (and by proxy lexical development)?

(3) Which 2000-lemma vocabulary list reveals level differences in lexical sophistication most clearly?

3. METHOD

3.1 Selection of frequency lists for AG

The first list that was selected for AG was the NGSL. This list, released in 2013, is an updated version of the General Service List from 1953 [31]. Unlike many publicly-available word lists, the NGSL is specifically designed with second language learners in mind, and therefore, relevant to Pitt IEP students. To achieve validity, the NGSL is based on a subset of the large Cambridge English Corpus (CEC) which contains two billion words; the subset selected consists of 272 million words, representative of a number of sub-corpora, most notably 38 million words from the Cambridge learner corpus. As a result of this careful corpus composition, the overall coverage of the NGSL exceeds 90% of the CEC texts. The NGSL was also selected due to its public availability in useful Excel file format and clear division of the lemmas into their headwords and inflected forms. In total, for the AG calculations, we used the 2000 highest-frequency lemmas (in keeping with the standard AG formula), as well as an additional 52 basic lemmas from the NGSL supplementary list such as the months of the year and numbers up to one hundred. In the upcoming version 2.0 of the NGSL, these supplementary items will be included in the overall frequency list [5].

The second list was derived from data from PELIC. This corpus contains both written and spoken data that were collected via a web interface and initially stored in a MySQL database. Students may have contributed data from one to three terms, with an average of two terms. For our dataset, we used only the written data from writing classes at the most common levels, levels 3 (intermediate), 4 (upper-intermediate), and 5 (advanced). The written data are 4.2 million tokens from several L1 backgrounds, but primarily Arabic, Chinese, Korean, Spanish, and Japanese learners. The written data were extracted from the MySQL database and analyzed in Python.

To create a high-frequency list from PELIC, which we call the Pitt Service List Level 3 (PSL-3), we used the same 52 supplementary items from the NGSL (for consistency) and added the next most frequent 2000 words in the learners' output at the intermediate level (level 3). When comparing the two lists, the analysis revealed that in terms of identical lemmas, only 1317 of the PSL-3 are found in the NGSL top 2000, with an additional 178 of the PSL-3 in the NGSL top 3000. Words in the PSL-3 that were not in the NGSL top 2000 fell into three broad categories: (i) cultural: e.g., 'camel', 'pyramid', 'spicy', 'tofu', and 'kimchi'; (ii) names: e.g., 'Japan', 'Colombia', 'Pittsburgh'; and (iii) student life: e.g., 'campus', 'admission', 'visa', and 'homework'.

3.2 ETS Comparison-Validation

For comparative purposes, we ran the same AG calculations on a different, but comparable learner corpus: the ETS Corpus of Non-Native Written English (ETS; [2]). This corpus consists of 12,100 English essays written by TOEFL test-takers in 2006-2007. These test-takers have 11 different L1s (many the same as in PELIC), and the texts are divided equally amongst them (1100 per L1). ETS split test takers into proficiency rankings of 'low', 'medium', or 'high'. As such, overall differences in AG lexical sophistication could be measured across proficiency bands.

ETS and PELIC share some similarities since both are learner corpora, contain a variety of L1s, and divide into three proficiency levels. However, they differ in that ETS data were collected under test conditions, whereas PELIC data were collected from day-to-day assignments. Nevertheless, we would expect any patterns found in lexical sophistication in one to be mirrored in the other if the underlying learner-corpus-based frequency lists are generalizable beyond our local context. That is to say, the PELIC-based and NGSL-based AG should equally indicate differences in lexical sophistication on both, despite PELIC and ETS not sharing any of the same learners, tasks, or specific writing prompts.

3.3 PELIC data processing

To preprocess the PELIC data samples for AG analysis, various Python libraries such as pandas, spaCy, and NLTK were used. We filtered out all texts with less than 70 words, following [12], who had a minimum of 66-word texts in their corpus. This process reduced the number of texts from 48,384 to 16,227, but only reduced the token count by 13% from 4,232,746 to 3,736,556. Further filtering of the data was then required as learners in the Pitt IEP revised and re-submitted assignments, often resulting in multiple versions of the same text; the dataset was therefore screened to include only the first version each essay. In addition, within each level and L1 group, there is variance in terms of proficiency and the number of texts and tokens produced. To account for this variation, we calculated average AG scores for individuals to prevent any skewing of data by prolific writers.

Manipulation of the texts was kept to a minimum, and we made a conscious decision to not correct some spelling errors. For example, if a student meant to write 'pot' or 'raw' but due to potential phonological influence on spelling wrote 'port' or 'row', these contextual spelling errors were neither screened nor corrected. However, misspelled tokens were excluded from analysis if they resulted in a non-word (as determined by NLTK's WordNet Synsets as a spellchecker). Such a step was necessary in order to avoid having misspelled basic words like 'thier' register as an advanced type, thereby inflating the AG score. To illustrate the significant effect that misspellings which create non-words can have on lexical sophistication measures, in the ETS data, Arabic low-proficiency texts had an average AG of 1.3 when misspellings were included, whereas this figure dropped to 0.37 when non-word misspellings were excluded from calculation.

Another consideration was advanced-level lexical items found in the writing prompts, which are frequently repeated in student responses. After considering removal of such lexical items from calculations, we ultimately decided to leave them in because the fact that the student 'took up' and used the words in their writing suggests that some learning may have occurred.

Each text was then tokenized using regular expressions. Finally, these tokens were lemmatized, taking the third approach described in section 2. Having completed the above data cleaning process, the resulting data for analysis was comprised of the numbers of texts in Table 1 and individual students in Table 2.

Table 1. Number of texts > 70 words by L1 and level

Level	Arab	Chin	Japan	Korea	Span
3 (Intermediate)	844	307	89	408	116
4 (Upper-Int.)	1659	1001	400	1191	234
5 (Advanced)	1229	851	271	797	184

Table 2. Numbers of students by L1 and level

Level	Arab	Chin	Japan	Korea	Span
3 (Intermediate)	131	48	14	63	13
4 (Upper-Int.)	210	101	39	120	29
5 (Advanced)	141	71	27	86	20

4. RESULTS

4.1 AG measurements of PELIC data

To reiterate, AG is defined as:

$$AG = \frac{Advanced\ Types}{\sqrt{Tokens}}$$

Section 4 describes the results of computing AG using two different high-frequency lists: NGSL and PSL-3. Tables 3 and 4 report the results in that order and the corresponding figures display the mean AG data with standard error bars indicating variability.

Table 3. AG with NGSL on PELIC mean (stdev)

Level	Arab	Chin	Japan	Korea	Span
3	0.63 (0.23)	0.64 (0.23)	0.66 (0.17)	0.77 (0.22)	0.67 (0.15)
4	0.75 (0.25)	0.80 (0.28)	0.83 (0.26)	0.78 (0.21)	0.88 (0.31)
5	0.85 (0.33)	1.06 (0.32)	1.05 (0.29)	0.94 (0.31)	1.03 (0.23)

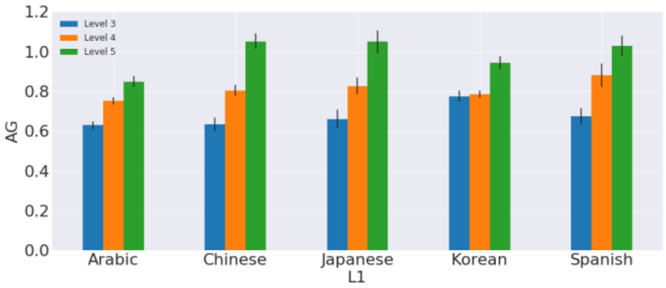

Figure 1. Average AG (using NGSL) on PELIC

Table 4. AG with PSL-3 on PELIC mean (stdev)

Level	Arab	Chin	Japan	Korea	Span
3	0.33 (0.19)	0.31 (0.16)	0.32 (0.19)	0.47 (0.22)	0.33 (0.10)
4	0.57 (0.28)	0.64 (0.31)	0.63 (0.30)	0.59 (0.23)	0.72 (0.39)
5	0.74 (0.37)	0.99 (0.40)	0.94 (0.40)	0.88 (0.37)	0.97 (0.34)

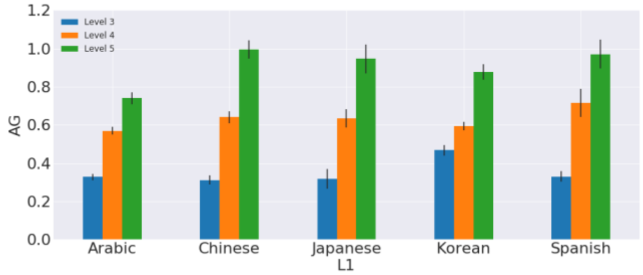

Figure 2. Average AG (using PSL-3) on PELIC

The results in Tables 3 and 4 show that for all L1s, some reliable and consistent group increases are evident in AG as proficiency level increases, regardless of whether NGSL or PSL-3 are used in the AG calculations. Thus, the NGSL means and PSL-3 means distinguish AG among levels. Although standard deviations are high, hand-calculated Confidence Intervals (CI) at the 95% critical value (1.96) show mostly non-overlapping means. This is true for all L1 groups with the exception that the Spanish speakers show an overlap of upper and lower CI for levels 4 and 5 with NGSL. Also noticeable is the difference between levels 3 and 4 for Koreans when using NGSL, as the increase in AG is not significant unlike for the other L1s. However, when PSL-3 is used, this lack of increase is corrected, showing greater increase as would be expected.

However, NGSL and PSL-3 differ in the AG scores that they produce. PSL-3 returns lower AG scores overall, but shows greater range, e.g., approximately 0.31 (Chinese level 3) to 0.99 (Chinese level 5) (a range of 0.67), compared to 0.64 (Chinese level 3) to 1.06 Chinese level 5 (a range of 0.42) for NGSL. The AG scores being lower overall for PSL-3 confirms that PSL-3 includes more words that the learners already know. However, by level 5, AG scores are comparable regardless of the high-frequency list used, indicating that they receive credit for high-frequency words which they later learn. Additionally, with PSL-3, level scores across all L1s appear more distinctly and uniformly segregated: all Level 5 scores regardless of L1 are higher than Level 4 scores. This was not the case with NGSL: the Arabic Level 5 score, for instance, is seen on par with Level 4 scores of other L1s, suggesting (incorrectly) that Arabic Level 5 students are at a similar level of lexical sophistication to, say, Spanish Level 4 students.

In terms of specific L1 differences, there are clear effects for Arabic and Spanish speakers. Overall, Arabic speakers have a lower range and Spanish speakers have a higher range. This lower range in the Arabic speakers' data is manifested across both AG measures, but the upper bound CI for level 5 with PSL-3 was lower than the lower bound CI at level 5 when using NGSL. This result again suggests that PSL-3 is appropriately discounting low-frequency, culture-specific words which learners already know that would otherwise inflate their AG score.

4.2 AG measurements of ETS data

For comparative purposes, we then measured AG in the same way using NGSL and PSL-3, but this time on the ETS corpus. Tables 5 and 6 report the results in that order and the corresponding figures present the mean AG data with standard error bars.

Table 5. AG with NGSL on ETS mean (stdev)

Level	Arab	Chin	Japan	Korea	Span
low	0.34 (0.22)	0.37 (0.26)	0.34 (0.20)	0.38 (0.21)	0.43 (0.23)
medium	0.48 (0.26)	0.58 (0.31)	0.51 (0.20)	0.60 (0.21)	0.55 (0.23)
high	0.82 (0.44)	0.91 (0.42)	0.68 (0.30)	0.83 (0.41)	0.79 (0.38)

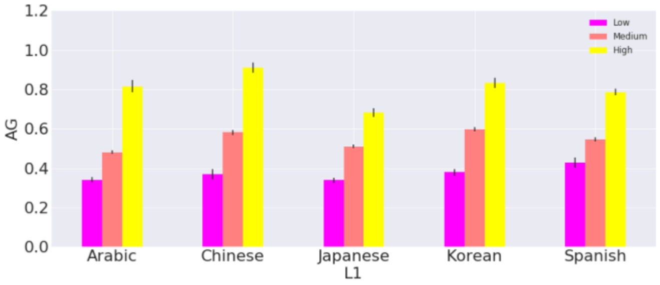

Figure 3. Average AG (using NGSL) on the ETS Corpus

Table 6. AG with PSL-3 on ETS mean (stdev)

Level	Arab	Chin	Japan	Korea	Span
low	0.31 (0.24)	0.35 (0.27)	0.30 (0.22)	0.33 (0.21)	0.44 (0.25)
medium	0.49 (0.32)	0.60 (0.37)	0.548 (0.28)	0.57 (0.31)	0.57 (0.33)
high	0.95 (0.51)	1.02 (0.53)	0.74 (0.38)	0.87 (0.46)	0.93 (0.48)

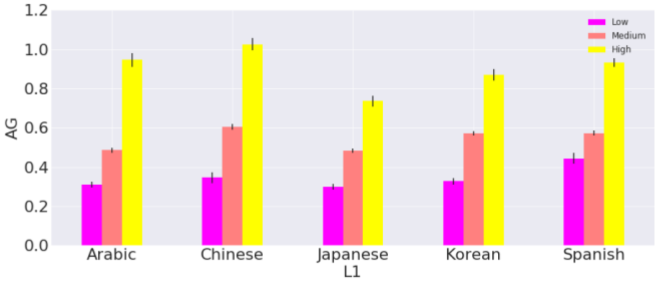

Figure 4. Average AG (using PSL-3) on ETS

These results from the comparison ETS corpus reveal a great deal of consistency in terms of the trends described in 4.1. We acknowledge that the essays in ETS are labelled ‘low’, ‘medium’, and ‘high’, and as such are not strictly comparable to the level system in PELIC. Nevertheless, the AG which was based on PSL-3 appears more effective at showing differences in lexical sophistication than NGSL, as would be expected for learners of different proficiency levels completing an international proficiency exam like TOEFL. This pattern suggests that the findings in 4.1 are not purely specific to the Pitt IEP context, but importantly can be generalized to other learner datasets (though not as effectively as compared to the local context).

5. DISCUSSION

5.1 Differences in frequency lists

To return to our research questions, for question 1, we have demonstrated how data science methods, and specifically natural language processing (NLP) suites such as spaCy and NLTK in Python, can be successfully used to automatically produce vocabulary lists through lemmatization, removal of non-word spelling errors, and token frequency counts.

Regarding research question 2, we showed in answer to question 1 that different frequency lists could be created and deployed and that the choice of corpus affects which high-frequency words are included. In our analysis of our two high-frequency word lists for calculating AG, we found that both NGSL and PSL-3 can show reliable increases as proficiency level increased. These increases in lexical sophistication were detected in both the local learner corpus, PELIC, and the international learner corpus, ETS, validating PSL-3. In addition, the analysis shows that for each L1, AG increases significantly from level to level. (The exception was Spanish-speaking learners from level 4 to 5; this result may be due to low-frequency words being based on Greek and Latin roots which the Spanish speakers control more easily.)

In answer to question 3, we found that the results from the two frequency lists differ in terms of the degree to which AG levels increased with proficiency levels. Overall, the learner-corpus based frequency list yielded more distinct AG differences from level to level, indicative of how we would expect AG to increase with a learner’s overall lexical development over time in an instructed context. Here we acknowledge that the level-by-level data described is cross-sectional, but it can serve as a proxy for longitudinal growth; in future work, hierarchical linear modeling (HLM) will be used to statistically confirm this claim. (HLM is appropriate as not all learners provide a data point at each level, but this statistical approach allows one to compensate for this issue, e.g., [29]) Instead, at present we are restricting the analysis to the calculation of mean scores with confidence intervals, thereby allowing us to provide descriptive evidence of differences in AG when different lists are used.

Our explanation for this finding is that learners may already know and control some less frequent NGSL words at a low-intermediate stage due to cultural background but may not know some words that occur in the 2000 most frequent words in a native speaker corpus. This knowledge inflates AG at lower proficiency levels. In other words, when measuring lexical development against a native-speaker corpus, learners incorrectly get credit for less frequent words that they already know (items not in the frequency list from their culture or educational context), but do not get credit for words that they learn when these more frequent items become known to them. Thus, native speaker-based frequency measures may present a less nuanced picture of the L2 productive lexicon. The learner-corpus frequency list provides more differentiated AG scores, resulting in a more clearly stratified picture of learner knowledge across levels, and by extension, predicted longitudinal growth.

5.2 Importance of data science tools

These observations were made possible by data analysis of very large numbers of texts and tokens. To our knowledge, data mining analysis of a corpus of learner data of this nature, with a variety of L1s and a similarity of educational experience in an IEP, has not been reported before in the literature. Although a subset of the

PELIC spoken data was hand-coded and made public (see, e.g., http://alpha.talkbank.org/data-cmdi/talkbank-data/SLABank/English/Vercellotti/) and several articles published since [20, 29, 30], the potential for far greater insights into development in an IEP are possible from analysis of the whole dataset. Therefore, the ability to analyze a learner corpus of this size is an important step forward in more precise characterization of 'academic readiness', which is an issue in IEP programs that prepare international students for academic programs [15].

5.3 Limitations and L1 effects

We acknowledge that there are limitations at this early stage of exploration. For example, we have yet to determine the exact effect of task prompts or the most reliable manner of lemmatizing our own high-frequency lists with open-source tools. Another area for investigation is the degree to which specific L1 characteristics affect their AG measurements. For example, it has been documented in PELIC that Arabic learners tend to misspell more than other L1s [14]. By excluding all non-word misspellings, Arabic learners may not receive credit for words they may know in all senses except for the spelling. This finding is important as Arabic speakers' knowledge of the L2 may be underestimated and thus put them at a disadvantage in standardized proficiency tests, which are the gateway to quality higher education programs.

6. CONCLUSION

This paper used data mining techniques to provide evidence that AG measures of lexical sophistication will provide more accurate descriptive data if they are based on learner corpora (e.g., PSL-3) rather than frequency lists based on native speaker corpora (e.g., NGSL). The work presented here shows that mining a large dataset that has been collected from an L2 population can provide more fine-grained insight into level differences, and by implication development, than data that are less closely associated with the learners. This research is also a good example of how applied linguists and data scientists can collaborate to provide results from very large datasets, combining linguistic theory with data analysis.

As a next step, we plan to conduct further analysis and comparisons using other corpora and word lists as the basis for calculations. The Cambridge English: Preliminary and Preliminary for Schools Vocabulary List (PET; [7]) which is based on the Cambridge Learner Corpus, a subset of the CEC, is an obvious choice. As this list is intended to indicate words that a learner at CEFR level B1 should possess, it would seem a well-suited comparison to PSL-3. It may be that an ideal frequency list would consist of a combination of a local (like PSL-3) and a global (like PET) list in order estimate learner knowledge and their lexical needs.

We will also explore additional quantitative validation metrics, such as comparing AG scores with various frequency lists to general proficiency measures. We would also like to know whether culture-specific words such as 'camel', 'pyramid', 'tofu' and 'spicy' should be counted for all L1s. It is natural that Arab-speaking learners already know 'camel', but perhaps not Japanese learners, who are more likely to be familiar with 'tofu'. Would L1 specific versions of PSL-3 change the outcomes for each L1 and would materials writers for each L1 context find such L1-specific lists useful?

Overall, this research has the potential to inform numerous areas of language teaching. For materials writers, curriculum planners, and teachers, there is great value in having easy access to a valid list of level- and context-appropriate vocabulary on which to base classroom lessons. For testing services such as ETS or other institutions interested in automated assessment of proficiency levels, such lists can improve the reliability and validity of measurements related to lexical sophistication, and by extension, overall lexical development. Finally, in terms of research in this field, transparent and theoretically-motivated list selections allow for improved comparisons and reproducibility across studies. We therefore see this paper as a step in closing the gap between educational data mining research, classroom instruction, and assessment in the ESL industry.

7. ACKNOWLEDGMENTS

We would like to thank the teachers and students of the English Language Institute at the University of Pittsburgh and grants from the National Science Foundation via the Pittsburgh Science of Learning Center (http://learnlab.org), funded award number SBE-0836012. (Previously NSF award number SBE-0354420.)

8. REFERENCES

[1] Bauer, L. and Nation, P. 1993. Word families. *International Journal of Lexicography*, 6, 4, 253-79.

[2] Blanchard, D., et al. 2014. ETS Corpus of Non-Native Written English LDC2014T06. Philadelphia: Linguistic Data Consortium, 2014.

[3] Brezina, V. and Gablasova, D. 2015. Is There a Core General Vocabulary? Introducing the *New General Service List*. *Applied Linguistics*, 36, 1, 1, 1–22. DOI: https://doi.org/10.1093/applin/amt018

[4] Browne, C., Culligan, B. and Phillips, J. 2013. The New General Service List. Retrieved from http://www.newgeneralservicelist.org.

[5] Browne, C. 2014. A New General Service List: The Better Mousetrap We've Been Looking For? *Vocabulary Learning and Instruction*, 3, 2, 1-10.

[6] Bulté, B. and Housen, A. 2014. Conceptualizing and measuring short-term changes in L2 writing complexity. *Journal of Second Language Writing* 26, 42-65. DOI: http://dx.doi.org/10.1016/j.jslw.2014.09.005

[7] Cambridge English: Preliminary and Preliminary for Schools Vocabulary List. 2012. Retrieved from http://www.cambridgeenglish.org/images/84669-pet-vocabulary-list.pdf

[8] Cobb, T. 2016. Numbers or numerology? A response to Nation (2014) and McQuillan (2016). *Reading in a Foreign Language* 28, 2, 299-304.

[9] Daller, H., van Hout, R. and Treffers-Daller, J. 2003. Lexical Richness in the Spontaneous Speech of Bilinguals. *Applied Linguistics* 24, 2 (Jun. 2003), 197-222. DOI: https://doi.org/10.1093/applin/24.2.197

[10] Daller, H. and Phelan, D. 2007. What is in a teachers' mind? In Daller, Milton, Treffers-Daller (Eds.) *Modelling and Assessing Vocabulary Knowledge*, (234-244). Cambridge University Press, Cambridge.

[11] Daller, M., Turlik, J., and Weir, I. 2013. Vocabulary acquisition and the learning curve. In S. Jarvis & M. Daller (Eds.), *Vocabulary Knowledge: Human ratings and*

automated measures, 185-218). John Benjamins, Amsterdam.

[12] Daller, H. and Xue, H. 2007. Lexical richness and the oral proficiency of Chinese EFL students. In H. Daller, J. Milton, and J. Treffers-Daller (Eds.), *Modelling and assessing vocabulary knowledge* (150-164). Cambridge University Press, New York.

[13] Davies, M. 2008-. The Corpus of Contemporary American English (COCA): 560 million words, 1990-present. Available online at https://corpus.byu.edu/coca/.

[14] Dunlap, S. 2012. *Orthographic quality in English as a second language.* (PhD), University of Pittsburgh, Pittsburgh, PA.

[15] Hoekje, B.J., & Stevens, S.G. 2017. Creating a culturally inclusive campus: A guide to supporting international students. Routledge, New York.

[16] Gibson, E. and Schütze, C.T. 1999. Disambiguation preferences in noun phrase conjunction do not mirror corpus frequency. *Journal of Memory and Language,* 40, 263-279.

[17] Jarvis, S. 2013. Defining and measuring lexical diversity. In S. Jarvis & M. Daller (Eds.), *Vocabulary knowledge: human ratings and automated measures,* 13-44. John Benjamins, Amsterdam.

[18] Juffs, A. 1998. The acquisition of semantics-syntax correspondences and verb frequencies in ESL materials. *Language Teaching Research,* 2, 93-123.

[19] Juffs, A. (in press). Lexical development in the writing of English Language Program Students. In R. M. DeKeyser and G.P. Botana (Eds.), *Reconciling pedagogical demands with pedagogical applicability*. John Benjamins, Amsterdam.

[20] Li, N., and Juffs, A. 2015. The influence of moraic structure on English L2 syllable final consonants. P. *2014 Annual Meeting on Phonology*. DOI: http://dx.doi.org/10.3765/amp.v2i0.3767

[21] Malvern, D., Richards, B.J., Chipere, N. and Durán, P. 2004. *Lexical diversity and language development.* Palgrave, Basingstoke.

[22] Nation, P. and Waring, R. 1997. Vocabulary Learning Strategies. In N. Schmitt and M. McCarthy (Eds.) *Vocabulary: Description, Acquisition and Pedagogy,* 6-19. Cambridge University Press, Cambridge.

[23] Nation, I.S.P. 2001. *Learning vocabulary in another language*. Cambridge University Press, Cambridge.

[24] Schmitt, N. 2010. Researching Vocabulary. A Vocabulary Research Manual. Palgrave Macmillan, Basingstoke.

[25] Schmitt, N. and Schmitt, D. 2014. A reassessment of frequency and vocabulary size in L2 vocabulary teaching. *Language Teaching,* 47, 4, 484–503. DOI:10.1017/S0261444812000018

[26] Schmitt, D. 2016. Beyond Caveat Emptor: Applying Validity Criteria to Word Lists. In Proceedings of Vocab@TOKYO: Current Trends in Vocabulary Studies. September 12-14, 2016, 17.

[27] Tidball, F. And Treffers-Daller, J. 2008. Analysing lexical richness in French learner language: what frequency lists and teacher judgements can tell us about basic and advanced words. Journal of French Language Studies 18, 3, 299-313. DOI: https://doi.org/10.1017/S0959269508003463.

[28] van Hout, R. and Vermeer, A. 2007. Comparing measures of lexical richness. In H. Daller, J. Milton, and J. Treffers-Daller (Eds.), *Modelling and Assessing Vocabulary Knowledge*, 93-115. Cambridge University Press, Cambridge.

[29] Vercellotti, M.L. 2017. The development of complexity, accuracy and fluency in second language performance. *Applied Linguistics*, 38, 90-111. *Doi.org/10.1093/applin/amv002*

[30] Vercellotti, M.L., & Packer, J. 2016. Shifting structural complexity: The production of clause types in speeches given by English for academic purposes students. *Journal of English for Academic Purposes,* 22, 179-190. DOI: dx.doi.org/10.1016/j.jeap.2016.04.004

[31] West, M. 1953. A General Service List of English Words. London: Longman, Green and Co.