IRI Joint Research Series IRI-2020-03

An Exploratory Study in the Visualization of Corpus Data for the Design of an English Placement Test

Daniel Parsons
International University of Japan

Russell Mayne
International University of Japan

Michael Mondejar
International University of Japan

Richard Smith
International University of Japan

March 2020

IUJ Research Institute
International University of Japan

An Exploratory Study in the Visualization of Corpus Data for the Design of an English Placement Test

Daniel Parsons, Russell Mayne, Michael Mondejar, Richard Smith

Abstract

This report outlines an exploratory method for designing a grammar test for the purpose of placing university graduate students into an English support program. The method involved three interrelated steps. First, instructor insights and the current program syllabus were utilized in creating constructs. Second, learner corpus data were used to create visual supports through residual plots, mosaic plots, concordance lines and word frequency charts. These can facilitate test writers in making visual judgements about what items to include in a test and the types of multiple-choice distractors it might be appropriate to use. Finally, sample TOEFL tests were analyzed to provide a point of reference for our own test design. The main contribution of this research is to propose a procedure for test design that is based on evidence from learner corpus data while also accounting for local institutional needs.

Introduction

The visualization of large amounts of data to assist decision making has been growing in many fields in recent times. The field of corpus linguistics, which handles millions of words to enhance our understanding of language structure and use, has, since the 1960s, made use of concordance lines to visualize language structure.[1] Visualizing repeating language patterns with concordance lines has proved fruitful in shedding new light on the relationship between grammar and meaning.[2] While many of these new insights are now verified statistically, visualization is often the first step in the analysis, allowing the language researcher to establish hypotheses for further investigation.[3]

[1] John Sinclair, *Corpus, Concordance, Collocation* (Oxford: Oxford University Press, 1991), 32-4.

[2] Sinclair, 36.

[3] Vaclav Brezina, *Statistics in Corpus Linguistics: A Practical Guide* (Cambridge: Cambridge University Press, 2018).

The research reported here outlines how visualizations drawn from corpus data can be applied to the design of a placement test, focusing specifically on the design of the grammar component of the placement test. There are no concrete guidelines in the language testing literature that can help teachers to design grammar placement tests. More often than not, test writers must rely on their intuitions and experience of learners' language knowledge and language errors in order to write test items. However, a great deal of learner language is available through learner corpora, which can be useful for enhancing the test writer's intuitions of learners' language knowledge. By visualizing learner language with the corpus data, it is possible to make decisions about test items and multiple-choice distractors which are based on evidence about the kind of language learners are expected to know.

After outlining the need for this project, this paper will provide a brief review of the literature on testing and corpus linguistics. This is followed by an overview of the methodology used in the research. The main part of this paper provides an example of how a test questions can be constructed through reference to visual aids drawn from learner corpus data.

Impetus for the project

Our graduate-level institution, the International University of Japan (IUJ), currently utilizes the TOEFL as a major contributor in determining whether or not incoming students require extra English language support during the first year of their studies. However, given that proficiency tests like the TOEFL are designed to assess a wide band of abilities, they are not effective in determining whether students entering a particular institution have the appropriate academic background and skills to thrive in that particular institution.[4] In fact, Fulcher stresses

[4] James Dean Brown, *Testing in Language Programs: A Comprehensive Guide to English Language Assessment* (New York, NY: McGraw-Hill, 2005).

the dangers of using general "off-the-peg" tests, noting that these may not be able to provide information which is specific to a specific context.[5] In IUJ's case, the use of TOEFL has sometimes led to situations where students who may not need support are required to join our specific support courses, while other weaker students have missed out on much needed support. The project reported here attempts to address this weakness, focusing on the design of a grammar component of a placement test.

Placement tests

Over 20 years ago Fulcher noted that in contrast to their widespread usage the literature on placement testing was scant.[6] In 2020 the situation is not much improved and EAP professionals looking for practical advice as to how to write placement tests may be disappointed. For instance, Fulcher's[7] *Practical Language Testing* does not even discuss placement tests. Testing books that do mention them often do not offer any practical guidance as to how to actually write them[8] as is the case with Davidson and Lynch's[9] *Testcraft* which only mentions the term to define it. Bachman and Palmer, however, do provide some useful insights into writing and designing placement tests for a university program[10]. However, while there was useful information on writing appropriate specifications, the book mainly describes the creation of a writing test rather than a grammar test and was thus of limited utility.

[5] Glenn Fulcher, *Practical Language Testing* (London: Hodder education, 2010), 10.

[6] Glenn Fulcher, "An English Language Placement Test: Issues in Reliability and Validity." *Language Testing* 14, no. 2 (1997), 113-39.

[7] Glenn Fulcher, *Practical Language Testing* (London: Hodder Education, 2010).

[8] See for example Anthony Green, *Exploring Language Assessment and Testing* (New York: Routledge, 2014).

[9] Fred Davidson and Brian Lynch, *Testcraft* (New Haven: Yale University Press, 2002), 131.

[10] Lyle Bachman and Adrian Palmer *Language Testing in Practice* (Oxford: Oxford University Press, 1996), 253.

Corpus linguistics in language assessment

The application of corpus linguistics methods to language assessment only began in the 1990s, but since then large testing companies have used a diverse range of corpus resources and analytical approaches.[11] One resource available for language assessors is native speaker corpora. These corpora, such as the British National Corpus, consist of a large number of texts which represent speakers of English as a first language and which cover both spoken and written texts from a variety of genres, including newspapers, TV interviews, and academic journals. These kinds of corpora have been used to create word lists and develop rating scales that are useful in test design.[12] They have also been used to check whether a test item's construct properly represents correct usage by native speakers.[13] A further resource available for language assessors is learner corpora. These corpora, such as the Open Cambridge Learner Corpus[14], allow language assessors to observe both the language competence of learners and the errors that they make at different proficiency levels, which in turn allow assessors to validate their intuitions about the proficiency of learners.[15] They are also particularly versatile for creating tests of enabling skills such as grammar and vocabulary.[16]

The main purpose of this research was to use evidence drawn from a learner corpus to aid the design of multiple-choice grammar items and distractors. The approaches to drawing on learner corpora can be classified into corpus-informed, corpus-based, and corpus-driven

[11] Fiona Barker. "How Can Corpora Be Used in Language Testing?", in *The Routledge Handbook of Corpus Linguistics,* ed. Ann O'Keeffe et al. (Abingdon: Routledge, 2010), 637.

[12] Fiona Barker, Angeliki Salamoura and Nick Saville, "Learner Corpora and Language Testing," in *The Cambridge Handbook of Learner Corpus Research*, ed. Sylviane Granger et al. (Cambridge: Cambridge University Press, 2015), 512.

[13] Barker, "How Can Corpora Be Used in Language Testing?" 637.

[14] OpenCLC (v1), 2017.

[15] Barker, Salamoura & Saville, "Learner Corpora and Language Testing," 512.

[16] Barker, "How Can Corpora Be Used in Language Testing?" 634.

approaches.[17] In a corpus-informed approach, the corpus is used in the process of validating the assessor's claims about a test taker's proficiency, and this can be applied at any stage of the test cycle from planning to rating the test.[18] A corpus-based approach compares learner and native-speaker corpora to examine similarities and differences in usage, again facilitating the classification of proficiency levels.[19] A corpus-driven approach uses statistical methods on learner and native corpora to address researchers' hypotheses.[20]

Test construction with corpus linguistics

The important issues in testing are that a test measures what it is purported to measure (validity) and that the result did not come about by chance (reliability). A third important consideration for test writers is how much can be achieved with the available resources (practicality).[21] A corpus-informed approach can provide the evidence necessary to strengthen validity and reliability early in the test design process. This potentially mitigates the need to re-write substantial parts of a test.

After establishing the test purpose and criterion, in this case to allocate university resources accurately to the students with the greatest language needs, the next stage in creating a test is to define the constructs that the test will be designed to measure. Constructs are a way to measure a student's ability or skill in a chosen domain of knowledge.[22] Fulcher notes that construct definition must be undertaken with care.[23] If we wish to measure something, we must

[17] Marcus Callies, "Using Learner Corpora in Language Testing and Assessment: Current Practice and Future Challenges," in *Studies in Learner Corpus Linguistics: Research and Applications for Foreign Language Teaching and Assessment*, ed. Erik Castello et al. (Bern: Peter Lang, 2015), 23.

[18] Callies, "Using Learner Corpora in Language Testing and Assessment," 23.

[19] Callies, 23-24.

[20] Callies, 24.

[21] Bachman and Palmer "Language Testing in Practice," 35.

[22] Fulcher, "Practical Language Testing," 96.

[23] Fulcher, "Practical Language Testing," 97.

clearly delineate the boundaries of the qualities we seek to measure. Evidence from a learner corpus can facilitate the necessary care in construct design and can delineate the test takers' proficiencies.

Method

The process of selecting, collating, and analyzing data was broken down into three main steps. The first step involved faculty consultations in order to determine the main areas of grammar to focus on; the second step was a process of collecting and collating data from a learner corpus; the third step involved analyzing sample TOEFL tests to help mitigate limitations associated with the first and second steps.

Step 1: Faculty Consultations and Workshops

Most useful for this project was a paper describing the steps taken by a university academic program to create a new in-house placement test.[24] At one point Fulcher notes that the test items were chosen from the in-sessional grammar revision course by their difficulty.[25] This provided us with a reasonable model to emulate. Largely, due to the paucity of specific practical literature on this subject, the process was conducted somewhat "blind". This limitation is addressed later by analyzing TOEFL sample tests.

The main construct that we wanted to measure was "knowledge of language". In order to operationalize this construct, it was necessary to define what particular aspects of this we wanted to focus on. As the courses that students will go on to are designed to remedy the gap between

[24] Fulcher, "An English Language Placement Test: Issues in Reliability and Validity," 113-39.
[25] Fulcher, "An English Language Placement Test: Issues in Reliability and Validity," 115.

where their English is and where it needs to be (among other things), it seemed reasonable to define the construct in light of the syllabus areas of the courses that students would have to take were they to not be exempted. Basing constructs on syllabus features is considered acceptable practice.[26,27] Following on from this, tutors were asked to categorize a list of possible areas (grammar and vocabulary) as in Table 1.

We would expect all students to be able to do this reasonably well.	*Students we want to see are likely to struggle with this, but better students would likely not.*	*Only very good or native students would be able to do this well.*

Table 1: The grid used by instructors to categorize the importance of different grammar areas.

Tutors could also add or change categories. At the end, the results of this exercise were then compiled. If all tutors believed that the feature was in the second category, then it was seen as useful for the definition of the construct. For example, *articles* were not considered to be a useful indicator of whether students needed support whereas all tutors agreed that *word form* may be a good indicator. The final list of areas is given in Table 2.

[26] Fulcher, "An English Language Placement Test: Issues in Reliability and Validity," 113-39.

[27] Albert Weideman, "Validation and Validity Beyond Messick," Per Linguam: Tydskrif vir Taalaanleer 28, no. 2 (2012), 1-14.

	Grammar area considered to be important to learners at IUJ
1	Passive structures
2	Word order
3	Word form
4	Pronouns
5	Conditionals - specifically those using past tense
6	Question forms
7	Conjunctions (linkers, linking words etc) (vocab cohesive devices)
8	Determiners
9	Count/non-count nouns
10	Modality - *should* in particular (cautious language)
11	Trend language
12	Cause and effect language
13	Set phrases (e.g. *not only...but also*)
14	Language for comparing and contrasting
15	Sentence ordering (within a paragraph)

Table 2: Main areas of grammar considered to be important to learners at IUJ. The ranking of these is arbitrary.

Step 2: Corpus search and retrieval

Corpus description

The Open Cambridge Learner Corpus (OCLC) was selected as the reference corpus to search for specific language items.[28] The corpus was accessed through Sketch Engine, for which an individual user subscription was purchased. Additionally, in order to gain access to the OCLC

[28] OpenCLC (v1), 2017.

through Sketch Engine, a policy that Cambridge University Press will monitor all use of the corpus as the copyright holder of the corpus was agreed to.[29]

This corpus consists of 2.9 million words of learner examination texts and represents learners from a wide range of countries in Europe, some Asian countries and various countries in Africa and South America. The corpus contains texts for exams that target proficiency levels B2, C1 and C2 on the Common European Framework of Reference (CEFR) scale; in other words, only the intermediate to advanced level tests are represented. However, in terms of learner proficiency, learners who are assessed to be at B1 level (low intermediate) are also represented in the corpus, albeit only comprising a small proportion of the corpus. These are students who took the B2 level exam and did not do well enough to be awarded B2, and thus remained at B1 level. While this is suitable enough for the purposes of exploring the language use of intermediate level learners, the difference in size of the sub-corpus for each proficiency level should be noted.

The proportion of documents represented at each proficiency level is given in Table 3. It is clear from the table that B1 (non-pass B2 exam students) contains the smallest proportion of documents at 12.3% of the total number of documents. The total number of sentences for each proficiency level is not provided in the metadata for the corpus. However, it can be calculated via a corpus query language (CQL) search for the sentence tag, which appears once per tagged sentence. The total number of sentences can then be calculated through Sketch Engine's concordance frequency calculation tool.

[29] Lexical Computing CZ, *Sketch Engine*, https://www.sketchengine.eu

Learner Proficiency Levels	Average word count per document (standard deviation)	Total number of sentences (percent of total)	Total number of documents (percent of total)
B1	179 (39)	15,260 (9.0%)	1,416 (12.3%)
B2	209 (61)	44,191 (27.5%)	3,711 (32.2%)
C1	277 (84)	60,150 (37.4%)	4,096 (35.5%)
C2	350 (72)	41,196 (25.6%)	2,316 (20.0%)
TOTAL		160,797	11,539

Table 3: Summary data for word count, sentence-counts and document-counts in the Open Cambridge Learner Corpus. Data generated from OpenCLC (v1), 2017, Distributed by Lexical Computing Limited on behalf of Cambridge University Press and Cambridge English Language Assessment.

Corpus search

The process of searching the corpus involved three sub-steps: grammar enrichment, corpus query language construction, and data handling.

Grammar enrichment. Grammar enrichment was a necessary step in operationalizing the list of grammar constructs into groups of search terms written in corpus query language. This enrichment process involved narrowing down the constructs to more precise instances of the grammar. Because the learner corpus was not tagged for errors, some constructs (word order, word form and sentence ordering) had to be excluded from this step as these were not easily operationalized into corpus query language. However, these constructs could be investigated in the context of other constructs at a later stage in the analysis. Additionally, some items on the list were consolidated into one set. For example, the constructs conjunctions, cause-effect language and compare-contrast language were consolidated into a set which we labelled stance and linking adverbials. As part of this enrichment process, learner corpus literature and grammar dictionaries were consulted.[30,31,32,33]

Corpus Query Searches. Through grammar enrichment, the original list of constructs was transformed into ten broad sets (see Appendix 1), each with a varying number of searchable patterns (modal + adverb / adverb + modal / modal + not + adverb). In total, 175 instances of the ten grammar constructs were identified and translated into corpus query language. An example of the corpus query language search for modal-adverb clusters is:

[tag="MD"][word="n't|not"]{0,1}[tag="RB.*"]within<s/>.

This search term tells Sketch Engine to search for any modal verb, followed by an optional negative in the form of n’t or not, and then followed by an adverb. The expression within<s/> was applied to remove any instances in which the pattern crossed sentence boundaries. These search terms were entered into Sketch Engine’s concordance interface and all the sentences that contained the 175 instances were downloaded along with metadata about proficiency level and document numbers.

Corpus Data Handling. After retrieving the full data set from Sketch Engine, totaling approximately 139,000 sentences, three computer programs were written in R to carry out the following tasks: 1) pre-process the meta-data; 2) extract exemplar sentences at each proficiency level; 3) collate the data into types of grammar and proficiencies of test-takers.

[30] John A. Hawkins and Paula Buttery, “Criterial Features in Learner Corpora: Theory and Illustrations,” *English Profile Journal* 1(1) (2010).

[31] Anne O’Keeffe and Geraldine Mark, “The English Grammar Profile of Learner Competence,” *International Journal of Corpus Linguistics* 22(4) (2017).

[32] Biber et al., *Longman Grammar of Spoken and Written English.* (Essex: Pearson Education Limited, 1999).

[33] English Grammar Profile. http://www.englishprofile.org/english-grammar-profile (accessed September 1, 2019)

Step 3: TOEFL sample test analysis

The first step of faculty consultation was essentially a blind process due to a lack of literature and other guidance in this area. This implies that the list created might not be exhaustive. Additionally, the second step of corpus data collation, while involving enrichment of grammar, cannot itself provide guidance on sentence types and stem lengths. Therefore, the overall purpose of this third step was to provide a confirmatory dimension to the test design and provide a template for how to write suitable multiple-choice grammar questions. This process of reverse engineering a test allows test writers to enhance the validity, particularly the face validity, of the new test.[34]

Due to our lack of access to authentic TOEFL ITP tests, we utilized six sample tests from TOEFL preparatory textbooks for our test analysis. Three sample tests were selected from the *Longman Introductory Course for the TOEFL Test: The Paper Test.*[35] Note that the TOEFL paper-based test (PBT), for which this text was written, is essentially identical to the TOEFL ITP. Two of these tests were introductory level, while the third was "TOEFL-level", or of the same difficulty as an actual TOEFL test in terms of assessed breadth of English grammar knowledge. The introductory level tests were included in the analysis in order to examine the differences in grammatical range and stem length between the introductory level and TOEFL level of tests. In addition, one "TOEFL-level" test was selected from the text *ETS TOEFL ITP*

[34] Glenn Fulcher and Fred Davidson, *Language Testing and Assessment: An Advanced Resource Book*, (Abingdon, Oxford: Routledge, 2007), 56-7.

[35] Deborah Phillips, *Longman Introductory Course for the TOEFL Test: The Paper Test*, (New York: Pearson Education, 2004).

Official Guide, while two were selected from *Kanzen Kouryaku! TOEFL ITP Test Moshi 4-Kaibun* for test analysis.[36,37]

From each sample test, we only analyzed Section 2, which consisted of 15 multiple-choice structure items and 25 multiple-choice written expression (error identification) questions. For each item, the grammatical focus of the question and stem length was determined. The constructs were then compiled into a database for descriptive statistical analysis. The final data set can be found in Appendix 2.

Designing the test

Here we report the design of questions for the first set of grammar constructs: modal-adverb clusters. Modal-adverb clusters consisted of three types or patterns:

S1_T1: modal + adverb

S1_T2: modal + not + adverb

S1_T3: adverb + modal

The normalized (per 10,000 sentences) frequencies of each type of modal-adverb cluster at each proficiency level is provided in Table 3. A Pearson's chi-square test of independence was carried out to examine the relation between the use of the three modal-adverb cluster types and the proficiency levels. The relation between the variables was significant with χ^2 (6, N = 1584.5) = 20.8, p = .002. Cramer's V = 0.08 (a small effect size).

[36] Atsushi Tajino and Toshiyuki Kanemaru, *ETS TOEFL ITP Official Guide,* (Educational Testing Service, 2012).

[37] Paul Walden, Robert Hilke and Tetsuro Fujii, *Kanzen Kouryaku! TOEFL ITP Test Moshi 4-Kaibun*, (Tokyo: Alc Press, 2015).

Proficiency Level	S1_T1: modal + adverb	S1_T2: modal + not + adverb	S1_T3: adverb + modal
B1	171.04	7.21	47.84
B2	252.09	12.22	59.74
C1	374.07	16.96	59.02
C2	495.92	26.70	61.66

Table 4: Observed normalized frequencies (per 10,000 sentences) of modal-adverb clusters at four proficiency levels in OCLC.

Data visualization

After initial examination of the data, a procedure was designed to help facilitate the creation of placement test questions. The procedure is outlined in Table 5 and exemplified in detail below for the case of modal-adverb clusters.

Step	Purpose	Contribution to test-writing
Visualize residuals	To gauge the proficiency levels at which a particular grammar pattern contributes significantly.	Helps to determine the type of test suitable for the grammar pattern, e.g. act as a distractor in a multiple-choice question, be a gap fill question, etc.
Visualize mosaic charts	To gauge how the proportions of each grammar type varies within each proficiency level and how those proportions change with increasing proficiency	Guides us in judging the difficulty level of a particular grammar pattern. Guides us in determining whether or not to test the pattern.
Visualize concordance lines	To observe repeating patterns of usage.	Highlights errors in usage such as word order and sentence structure; shows use of fixed phrases; shows semantic preferences and variation
Visualize word frequencies with charts and word clouds	To observe changes in lexico-grammar with proficiency	Guides us in grading the vocabulary used in individual questions.

Table 5: An outline of the procedure for visualizing data in order to write placement test questions.

The first step is to visualize the residuals. Figure 1 shows a plot of the standardized residuals, which explains how each element contributes to the significance of the result of the chi-square test.

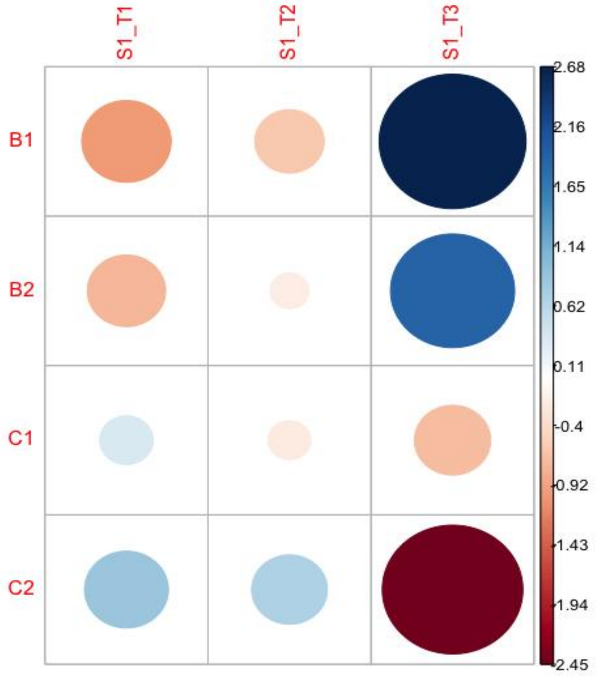

Figure 1: Standardized residuals for modal-adverb clusters at each proficiency level.

The colour blue shows a positive contribution to the significant result, and the colour red shows a negative contribution. The darker the colour, the greater the contribution. This implies that the S1_T3 pattern (adverb + modal), such as in the phrase *"I really would like some help"*, is used more frequently than expected by the least proficient learners and less frequently than expected by the highest proficient learners. This observation suggests that this pattern could be used as a distractor in a multiple-choice question that less proficient learners may be tempted to select and higher level learners would reject. The chart also shows that the pattern S1_T1 (modal + adverb), such as in the phrase *"I would really like some help"*, becomes more familiar to learners as their proficiency grows. This observation suggests that a gap fill question to elicit usage could also help to distinguish between lower and higher proficient learners. This inference can also be made about the pattern S1_T2 (modal + not + adverb), such as in the phrase *"We can't simply accept this idea"*.

The second step is to visualize the mosaic chart. Figure 2 presents a mosaic chart for the modal-adverb clusters. The chart shows the relative proportions of occurrence of each grammar type across proficiencies. The width of each bar indicates the relative proportion of all modal-adverb clusters at each proficiency. The heights of the bars within one proficiency level show the proportion of each type used by test takers at that proficiency level.

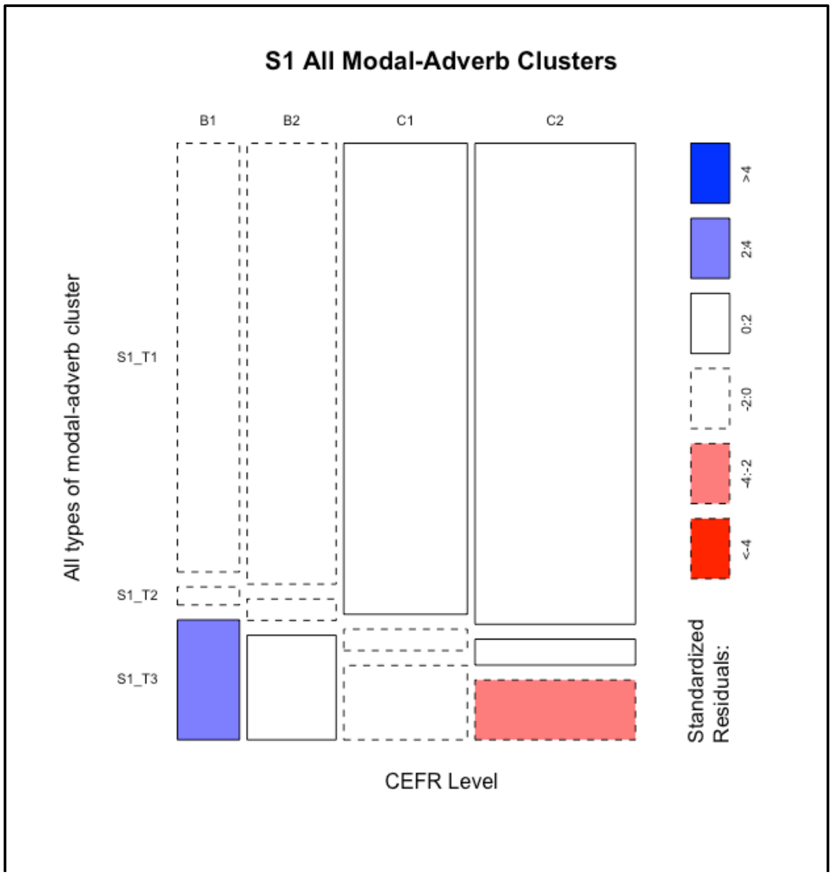

Figure 2: Mosaic chart showing the relative frequencies of three types of modal-adverb cluster

We can see immediately that the pattern S1_T2 is used much less frequently than other modal-adverb clusters at all levels. This suggests that expressions such as *"We can't simply accept this idea"* are rarely used among learners, and that testing this particular instance would likely not help us to easily discriminate low from high proficiency among learners.

The pattern S1_T1 is the most familiar pattern of the three main adverb-modal cluster patterns investigated at all proficiency levels. Looking vertically, the relative frequency of S1_T1 at each proficiency level grows with increasing proficiency (B1 = 75.7%, B2 = 77.8%, C1 = 83.1%, C2 = 84.9%). Looking at the areas of the rectangles, we can judge that C1 level learners use this pattern more than twice as frequently as B1 level learners. Since this pattern clearly grows with proficiency, it can be argued that it represents language development in learners, and thus, testing this item can indicate where learners are in their development. A correct answer in a well-written multiple-choice question should indicate familiarity, and a correct answer in a gap-fill question should indicate knowledge of usage.

The pattern S1_T3 shows a decrease in the proportion of usage as proficiency increases, but almost no change in relative frequency across proficiencies. This suggests that B1 learners may be making errors in usage of this pattern, whereas C2 learners may be using this pattern more judiciously. Therefore, as indicated earlier, this could be a good distractor in a multiple-choice question to discriminate those at lower levels.

Traditional visualization of the language data through concordance lines can also be carried out. Some representative examples of language use with modal-adverb clusters are given in Table 6.

As can be seen in Table 6, there is variation in acceptable usage. For example, the pattern S1_T1 (Modal + Adverb) captures “will always” and the pattern S1_T3 (Adverb + Modal) captures “always will”, both used by B1 level learners. The latter, while not completely unacceptable, is probably not acceptable in the context provided and could negatively affect the writer’s score on a writing test. Similarly, for the pattern S1_T2 (Adverb + not + Modal), the position of “not” in “won’t maybe” by a B2 level learner might be considered unusual. These types of instances, therefore, can act as good distractors in multiple choice questions.

Set_Type	**Pattern**	**Example**
S1_T1	Modal + Adverb	I will always prefer travelling by bicycle (B1) I would also like to thank you. (B2)
S1_T2	Modal + not + Adverb	they won’t maybe believe me (B2) You cannot always have a perfect meal. (C1)
S1_T3	Adverb + Modal	I always will prefer books (B1) What else can I say? (C2) People nowadays can appreciate… (C1)

Table 6: Examples extracted from concordance line data of modal-adverb clusters. Examples taken from OpenCLC (v1), 2017.

The second thing to notice is that the sentence types in which these patterns occur can vary. For example, pattern S1_T3 (Adverb + Modal) can occur in question forms. In this case, gap-fill questions that include question forms could help to eliminate higher proficient learners from the requirement for support.

The final visualization is that of vocabulary frequency. Bar charts and word clouds are helpful here. For example, as the bar charts in Figures 3 and 4 show, the word “can” is more familiar to lower level learners than the word “would”, but “would” grows in usage at higher proficiencies to overtake the use of “can”. Given that “would” has a wider semantic range than “can”, the use of “would” could be useful as a distractor, since less proficient learners may likely only know a restricted range of uses for “would”.

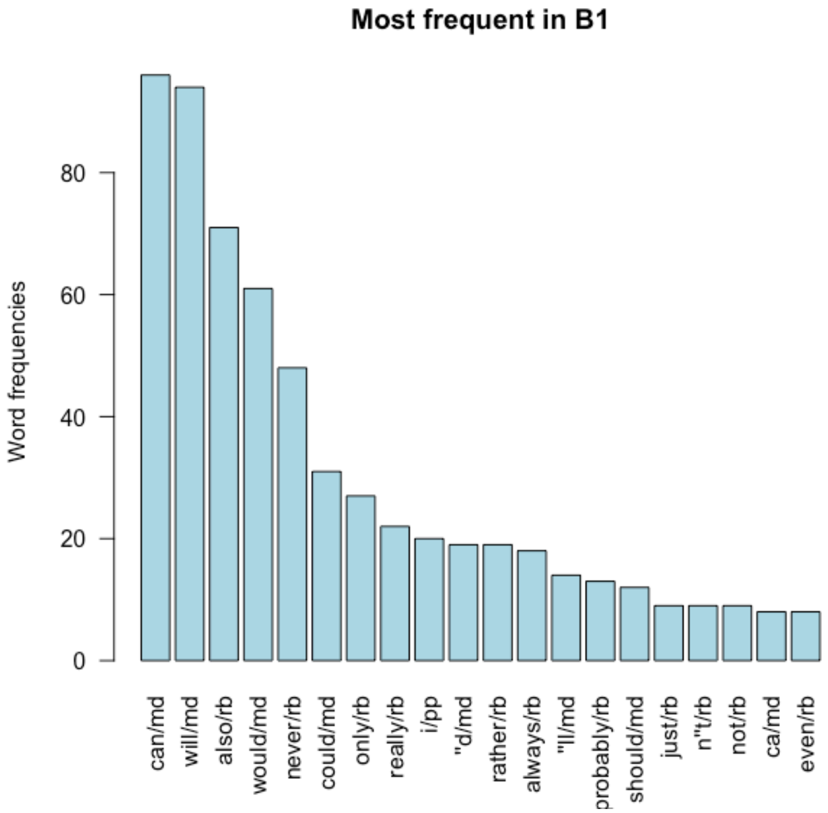

Figure 3: Bar chart showing the top twenty words, based on raw frequencies, used in the modal-adverb clusters at level B1.

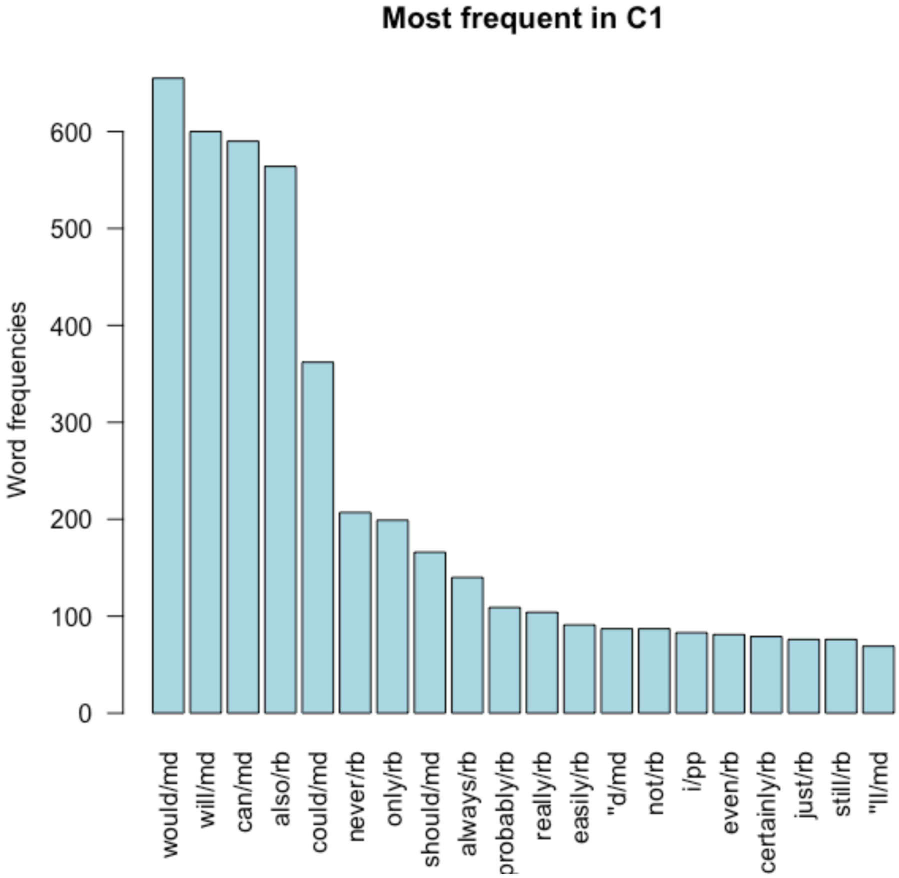

Figure 4: Bar charts showing the top twenty words, based on raw frequencies, used in the modal-adverb clusters at level C1.

Word clouds showing the prominence of words at each proficiency level can also be used to make decisions about items. One observation that can be made from these word clouds is that there is a growth in the range of adverbs as proficiency grows. Knowing the specific words that appear at each proficiency level can help the test writer to select vocabulary in order to adjust the difficulty of a question. For example, from Figures 5 and 6 we can see that the use of the word "not" appears to grow in prominence by the time learners reach C1 level. This implies that a test of the pattern with this word could help discriminate between intermediate and more advanced learners.

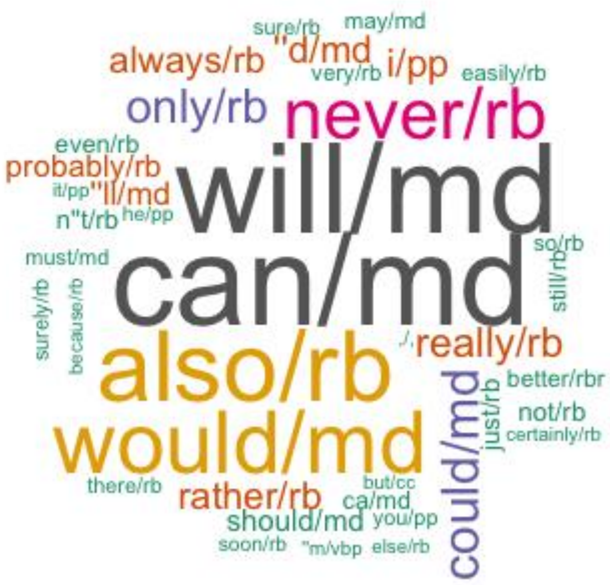

Figure 5: Word cloud showing the prominence of modal verbs and adverbs in the modal-adverb cluster at level B1.

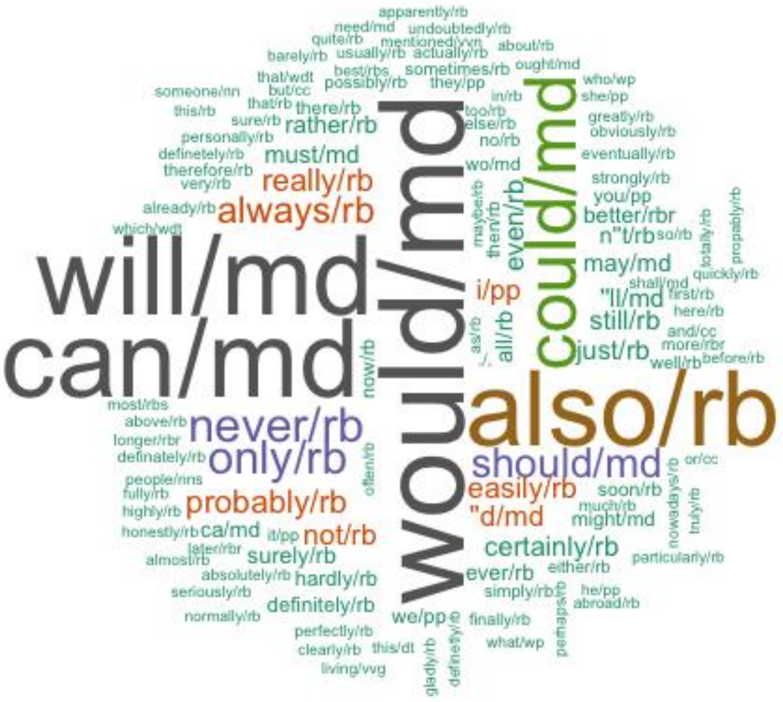

Figure 6: Word cloud showing the prominence of modal verbs and adverbs in the modal-adverb cluster at level C1.

Question writing

Based on the descriptions and explanations provided about the data so far, it is now possible to write some questions. Four example multiple-choice questions are described here.

1. I practiced hard last night for the test, so I ___________ do well tomorrow.

a. *might only*

b. *will really*

c. *would like*

d. *should never*

This question uses only B1 level vocabulary. The word order is modal + adverb with no errors in word order. The choice of “would like” is based on the hypothesis that learners might only know a restricted semantic usage of “would” at lower levels, and so they might be distracted by this option. We would expect this to be a simple question. Learners making mistakes with this type of question would very likely need support.

2. It was raining too much so I ___________ see where I was driving.

a. *hardly would*

b. *could hardly*

c. *must hardly*

d. *hardly must*

This question mixes B1 and B2 level vocabulary, particularly the use of hardly at B2. Word order is varied to distract lower level learners who may be tempted to mix word order in an unnatural way, as seen in the concordance lines for B1 level learners. We would expect this question to be slightly more difficult than the first question and incorrect answers from students would indicate a need for support.

3. The investors had a lot of money, but they_____________ make a one percent profit.

a. could not even

b. would like

c. always can

d. also never might

Question 3 again focuses on B2 level vocabulary, but this time offers different length options and possibly attracts more correct answers from C1 level learners because of the use of “not”. The use of “would like” is intended as a distractor that might attract lower proficiency learners, suggesting the need for support.

4. Because of the falling population and shortage of labor in the workforce, the Japanese government ___________ allow more foreigners into the country.

a. not only can

b. could well

c. either can

d. will not ever

The final question involves longer noun phrases as part of the question, and uses B1, B2 and C1 level vocabulary in the choices. Choices a and c are both found among learner writing at B2 level and above as can be seen by examining the concordance lines, but they are much less frequent and, in this context, they are syntactically incorrect. Similarly, the use of “could well” is also very low frequency and has a specific nuance that only higher-level learners would be expected to know. Therefore, this question is expected to be difficult for test takers, with correct answers suggesting that the learners certainly do not need remedial support.

Discussion and Conclusion

The procedure described above used visualizations to help design questions in a language test. The visualizations provided evidence of both correct and incorrect language use by learners, as well as guidance on what types of language learners might or might not be familiar with. Sinclair pointed out that the processing of large amounts of data in this way offers a new approach that can both counter and support our intuitions.[38] By accounting for the local institutional needs based on instructor experience and intuitions and by examining learner corpus data, the decisions about question types and multiple choice distractors can be made with much more confidence. In much the same way that visualization in corpus linguistics can help researchers make hypotheses about language usage, the design of test questions are essentially hypotheses based on a combination of the evidence from learner corpus data and the traditional intuitions of instructors and test-writers.

However, the question to ask ourselves at this stage is does our test have concurrent validity alongside other tests? Table 7 shows the frequency of English grammar constructs in the entire sample of two introductory-level and four TOEFL-level tests. Several constructs, namely coordination, agreement, reference, participles, determiners, SVO structure, and noun clauses, were assessed at a high frequency in the sample tests; these structures tended to be assessed in more than two items per test. Word forms, passives, collocations, comparatives, relative clauses, subordination, verb tenses, prepositional phrases, adjectives, adverbials, and clefts were assessed at a moderate frequency, appearing 1-2 times per test. Finally, modals, copulas, indefinite subjects, infinitives, inversions, object complements, superlatives, and discourse markers, were assessed at a low frequency on the tests. These low-frequency constructs tended to appear only in

[38] Sinclair, *Corpus, Concordance, Collocation*, 36.

the more difficult TOEFL-level sample tests (see Appendix 2) and were each utilized at most once a test. Commonalities exist between the grammar constructs generated through faculty consultations in the first part of this project and those derived from the TOEFL-test analysis. However, inclusion of high- or moderate-frequency constructs from the TOEFL analysis (but absent from the consultations) in the test generated by this project may improve our test's validity.

Construct	**#**	**%**	**Construct**	**#**	**%**
Coordination	30	12.5	Verb tense	8	3.3
Agreement	29	12.1	Prep phrase	6	2.5
Reference	22	9.2	Adjective	5	2.1
Participle	16	6.7	Adverbial	5	2.1
Determiner	14	5.8	Cleft	5	2.1
SVO structure	14	5.8	Modal	4	1.7
Noun clause	13	5.4	Copula	3	1.3
Word form	11	4.6	Indefinite subject	3	1.3
Passive	9	3.8	Infinitive	3	1.3
Collocation	8	3.3	Inversion	3	1.3
Comparative	8	3.3	Object complement	2	0.8
Relative clause	8	3.3	Superlative	2	0.8
Subordination	8	3.3	Discourse marker	1	0.4

Table 7: Summary statistics of all sample TOEFL ITP tests

Stem length was also examined in the test samples; the average length of a stem in all of the samples was 17.4 words, with the stems in the TOEFL-level tests being slightly longer than those in the introductory ones (18.2 words and 15.7 words, respectively). Based on this data, the optimal average length of a stem may be 16 to 18 words. However, there was some variability in the length of stems in the sample; stem lengths ranged from 8 to 33 words in the TOEFL-level tests, and from 8 to 29 words in the introductory tests.

This research had a number of limitations. The first is that there is no available literature on how to construct a test by using evidence visually from corpus data. Therefore, there was no "road map" to follow and we were unable to compare our approach with others. The success of this approach will be hard to judge without first piloting the test on students. This will provide us with initial insights into the performance of the test and help to identify any major issues with items. We should be able to see, to some extent, how well our hypotheses about the difficulty of certain items are borne out. However, testing is not an exact science and the particular construction of items may lead to results that were not expected. For instance, an item which is considered fairly "easy" may be written in a confusing way. The opposite situation may also arise. This all means that the post-test item analysis has to be carried out with care.

A second issue is that the ability to test our hypothesis regarding student outcomes is confounded by a number of factors. Students may not be successful on courses for a variety of reasons that are not related to their raw English ability. In some cases, we have heard from students who have performed poorly on the placement test on purpose in order to be able to take the extra English courses. In testing terms, this would appear to be a false positive. In contrast, a student may be exempted from the English course but find university study overwhelming. This student may not be able to produce good academic work for reasons other than her English ability. This may appear to be a false negative to the university stakeholders. It is not clear at this point how or if issues like these can ever be resolved satisfactorily.

Bibliography

Alexopoulou, Theodora, Helen Yannakoudakis, and Angeliki Salamoura. "Classifying Intermediate Learner English: A Data-Driven Approach to Learner Corpora." In *Twenty Years of Learner Corpus Research: Looking Back, Moving Ahead. Corpora and Language in Use - Proceedings 1,* edited by Sylviane Granger, Gaetanelle Gilquin, and Fanny Meunier, 11-23. Louvain-la-Neuve: Presses Universitaires de Louvain, 2013.

Bachman, Lyle, and Adrian Palmer. *Language Testing in Practice*. Oxford: Oxford University press, 1996.

Barker, Fiona. "How Can Corpora Be Used in Language Testing?" In *The Routledge Handbook of Corpus Linguistics*, edited by Anne O'Keeffe and Michael McCarthy, 633-646. Abingdon: Routledge, 2010.

Barker, Fiona, Angelika Salamoura, and Nick Saville. "Learner Corpora and Language Testing." In *The Cambridge Handbook of Learner Corpus Research,* edited by Sylviane Granger, Gaetanelle Gilquin, and Fanny Meunier, 511-533. Cambridge: Cambridge University Press, 2015.

Biber, Douglas, Susan Conrad, and Randi Reppen. *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge: Cambridge University Press, 1998.

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad, and Edward Finegan. *Longman Grammar of Spoken and Written English*. Essex: Pearson Education Limited, 1999.

Brezina, Vaclav. *Statistics in Corpus Linguistics: A Practical Guide.* Cambridge: Cambridge University Press, 2018.

Brown, James Dean. *Testing in Language Programs: A Comprehensive Guide to English Language Assessment*. New York: McGraw-Hill, 2005.

Callies, Marcus. "Using Learner Corpora in Language Testing and Assessment: Current Practice and Future Challenges." In *Studies in Learner Corpus Linguistics: Research and Applications for Foreign Language Teaching and Assessment,* edited by Erik Castello, Katherine Ackerly, and Francesca Coccetta, 21-36. Bern: Peter Lang, 2015.

Davidson, Fred, and Brian Lynch. *Testcraft.* New Haven: Yale University Press, 2002.

English Grammar Profile. http://englishprofile.org/english-grammar-profile (accessed September 1, 2019).

Fulcher, Glenn. "An English Language Placement Test: Issues in Reliability and Validity." *Language Testing* 14, no. 2 (1997): 113-39.

Fulcher, Glenn. *Practical Language Testing.* London: Hodder education, 2010.

Green, Anthony. *Exploring Language Assessment and Testing.* New York: Routledge, 2014.

Hawkins, John, A., and Paula Buttery. "Criterial Features in Learner Corpora: Theory and Illustrations." *English Profile Journal* 1(1) (2010): 1-23.

Lexical Computing CZ. *Sketch Engine*, https://www.sketchengine.eu

O'Donnell, Mick. "Using Learner Corpora to Order Linguistic Structures in Terms of Apparent Difficulty." In *Studies in Learner Corpus Linguistics: Research and Applications for Foreign Language Teaching and Assessment*, edited by Erik Castello, Katherine Ackerly, and Francesca Coccetta, 71-86. Bern: Peter Lang, 2015.

OpenCLC (v1). Distributed by Lexical Computing Limited on Behalf of Cambridge University Press and Cambridge English Language Assessment, 2017.

O'Keeffe, Anne, and Geraldine Mark. "The English Grammar Profile of Learner Competence." *International Journal of Corpus Linguistics* 22(4) (2017), 457-89.

Phillips, Deborah. *Longman Introductory Course for the TOEFL Test: The Paper Test.* New York: Pearson Education, 2004.

Sinclair, John. *Corpus, Concordance, Collocation.* Oxford: Oxford University Press, 1991.

Tajino, Atsushi and Toshiyuki Kanemaru. *ETS TOEFL ITP Official Guide*. Educational Testing Service, 2012.

Wadden, Paul, Robert Hilke, and Tetsuro Fujii. *Kanzen Kouryaku! TOEFL ITP Test Moshi 4-Kaibun.* Tokyo: Alc Press, 2015.

Weideman Albert. “Validation and validity beyond Messick” *Per Linguam: Tydskrif vir Taalaanleer* 28, no. 2 (2012): 1-14.

Acknowledgements

This research report has made use of the English Grammar Profile. This resource is based on extensive research using the Cambridge Learner Corpus and is part of the English Profile programme, which aims to provide evidence about language use that helps to produce better language teaching materials. See http://www.englishprofile.org for more information.

Appendix 1: Sets (S) of grammar constructs and the types (T) chosen to represent the set.

Sets		**Types**	
S1	Modal-Adverb Clusters	T1	Modal + Adverb
		T2	Modal + not + Adverb
		T3	Adverb + Modal
S2	Complex tenses	T1	Present perfect simple
		T2	Present perfect continuous
		T3	Present perfect passive
		T4	Past progressive
		T5	Past perfect simple
		T6	Past perfect continuous
S3	Complex nominal reference	T1	Quantifiers: Negative member or amount
		T2	Quantifiers: Large quantity
		T3	Quantifiers: Small quantity
		T4	Quantifiers: Inclusive quantifiers
		T5	Quantifiers: Phrasal quantifiers
		T6	Quantifiers: Comparative
		T7	Pronouns: one/ones
		T8	Nouns with gerunds
S4	Coordination	T1	Correlative conjunctions: both … and
		T2	Correlative conjunctions: either … or
		T3	Correlative conjunctions: neither … nor
		T4	Correlative conjunctions: not only … but also...
S5	Subordination	T1	Time subordinators
		T2	Concession subordinators
		T3	Contingency subordinators
		T4	Result/Cause subordinators
		T5	Contrast subordinators
		T6	"As"
		T7	Complex subordinators
S6	Linking and Stance Adverbials	T1	Sequencing
		T2	Adding emphasis
		T3	Contrasting
		T4	Adding
		T5	Comparing
		T6	Summarizing
		T7	Apposition
		T8	Result/Inference
		T9	Stance

Sets		Types	
S7	Determiners	T1	Possessives (its, their, of, ‘s)
		T2	Articles (another, the other)
		T3	Demonstrative pronouns (one of these)
S8	Conditionals	T1	If + present tense + will
		T2	If + present tense + modals
		T3	If + present continuous
		T4	If + past simple + would
		T5	If + past simple + could
		T6	If + past perfect + would have + -ed
		T7	Unless + present simple
S9	Passive Structures	T1	Past simple
		T2	Past continuous
		T3	Past perfect simple
		T4	Present continuous
		T5	Infinitive after verbs, adjectives and nouns
		T6	Modal perfect
S10	Relative Clauses	T1	Past perfect continuous + past perfect passive
		T2	Past perfect simple
		T3	Past progressive
		T4	Present perfect simple
		T5	Present perfect continuous
		T6	Present perfect passive

Appendix 2: Summary data for introductory and TOEFL-level tests.

Summary statistics of introductory-level tests

Construct	#	%
Agreement	14	17.5
Coordination	11	13.8
Reference	10	12.5
SVO structure	7	8.8
Noun clause	5	6.3
Participle	5	6.3
Word form	5	6.3
Determiner	4	5.0
Passive	4	5.0
Subordination	4	5.0
Verb tense	4	5.0
Modal	2	2.5
Relative clause	2	2.5
Cleft	1	1.3
Collocation	1	1.3
Copula	1	1.3

Summary statistics of TOEFL-level tests

Construct	#	%
Coordination	19	11.9
Agreement	15	9.4
Reference	12	7.5
Participle	11	6.9
Determiner	10	6.3
Comparative	8	5.0
Noun clause	8	5.0
Collocation	7	4.4
SVO structure	7	4.4
Prep phrase	6	3.8
Relative clause	6	3.8
Word form	6	3.8
Adjective	5	3.1
Adverbial	5	3.1
Passive	5	3.1
Cleft	4	2.5
Subordination	4	2.5
Verb tense	4	2.5
Indefinite subject	3	1.9
Infinitive	3	1.9
Inversion	3	1.9
Copula	2	1.3
Modal	2	1.3

Construct	#	%

Construct	#	%
Object complement	2	1.3
Superlative	2	1.3
Discourse marker	1	0.6