Computerized Grading of Simple Sentence Forming Skill

Towards the Development of a Method for Rapid Assessment and Diagnosis

コンピューターでの文を作るスキル評価

—アセスメントと診断の早い方法の開発へ—

BRINKMAN Michael

ブリンクマン　マイケル

Computerized Grading of Simple Sentence Forming Skill

Towards the Development of a Method for Rapid Assessment and Diagnosis

コンピューターでの文を作るスキル評価

―アセスメントと診断の早い方法の開発へ―

BRINKMAN Michael

ブリンクマン　マイケル

Abstract：A method for the computerized analysis and diagnosis of basic sentence grammar is outlined.

Keywords：computerized grading, grammar diagnosis, grammar error analysis

1．Introduction

Assessment of student English skill is a very complex area of study, with different methods and parameters, depending on the learner's level, age, and the purpose of the testing. Tests may be broadly divided into two categories: (1) Macroscopic tests which seek to assess learner's overall language/communication ability (skill), or (2) smaller tests designed to test more specific areas of understanding (knowledge). Tests may also be characterized by amount of subjectivity. If humans are involved in the grading process, there will be an element of subjectivity. A rubric and/or careful specifications will reduce this somewhat, but there is

a limit to how many specifications humans can remember as they listen to or read student work, and each criterion will be somewhat subjectively applied. Of course, macroscopic tests are very useful, as evidenced by the import companies place on candidates' test scores, and how much students are willing to pay. Research also shows that at least the TOEIC Speaking and Writing tests are reasonably stable (Qu et al). Certainly, these tests are valuable for assessing language skills as they would be used at work or school. At the other end of the spectrum, multiple-choice or fill-in-the-blank tests have clear right and wrong answers. They are best suited for assessing student understanding of specific language details (knowledge). These tests are necessary for assessing students' grasp of the basic building blocks of English, and are simple to mark and can often be easily marked by machine, but of course real-world communication is not like this at all. Various writing tasks or translating into English are more real-world, but marking them can become subjective again, since there is quite likely more than one way to correctly form an answer. This is further compounded by the ability of real-life conversation partners to understand what the speaker means in spite of mistakes, and some speakers are able to creatively overcome their language deficiencies to successfully communicate anyway (especially in a face-to-face situation) by somehow providing enough information to communicate the information, or at least enough for the communication partner to hazard a guess, which can then be further clarified by the original speaker. This sort of creativity-enhanced communication technique is often used by travelers to a country with a different language who attempt to shop or ask for directions. Adding gestures and drawings can further improve communications speed and accuracy. Of course, the advantage of improving grammar and increasing vocabulary is that communications become smoother and more reliable the more grammar and vocabulary one can use.

One more way to categorize tests and testing methods is by how much diagnostic information they provide, and about what. The choices looked at so far are either appropriate for high level assessment and diagnosis, where basic sentence-forming ability is assumed, (and there is little or no opportunity for diagnosis of problems at low level), or require selection of a correct answer from a limited set of options, which checks understanding of a specific language detail without requiring the student to be able to integrate it with other details to make an actual useful sentence. The limited options of the test question control the

variables so that it is easy to come up with a number score, which is useful for assessing student progress, but it is somewhat artificial, as the question itself provides a lot of information not provided in real life conversations.

Among the testing options we have not discussed yet are having students generate full sentences in response to some stimuli, and marking them by hand. To simplify marking, one could give full points for a perfect answer, and zero for any deviation; but as mentioned above, there are many kinds of mistakes which do not impair communication, or which are easily recoverable. It would be perfectly reasonable to award part marks in those cases, but if one has many tests to mark, maintaining consistency in the face of such a wide variety of partially correct answers is a big challenge, even if the same person marks all the tests. Tabulating data about the sorts of mistakes individual students make on each question would be more time consuming yet, and having multiple marking assistants at multiple locations would be even more challenging.

One example of a method to partially address this concern is to give a Japanese sentence as a stimulus, and provide the words of an English translation in a mixed up order. The task is to arrange the words in the correct order, and then (for example) the student would write down the second and fifth words on the test paper. Under this schema, the student could get zero points, one point (in two different ways), or two points. There is some tolerance for mistakes and some credit can be given for partially correct answers, but most of the information about the rest of the words and their relations is discarded, along with most of the diagnostic potential.

Testing using a communications task with a partner (successful completion of the task indicates that any errors made by either partner were corrected sufficiently to enable completion of the task, and therefore indicate communicative competence on that point) is another useful testing technique, in addition to being good practice, but the limitation is that there is no diagnostic information about either low level grammar mechanics or about who made the mistakes, and who discovered them and how they corrected for them. There are a lot of uncontrolled variables, which do not interfere with macroscopic assessment, but do make enough 'noise' to easily mask small effects at a low level if used before and after a unit for assessment.

Using some sort of artificial intelligence system might seem appropriate, and indeed there are systems used commercially in which computers grade essays for the GRE or other high stakes tests, but there are definite problems so far, as studied by MIT's Les Perelman and commented on by Orion Taraban (executive director of a tutoring company in San Francisco) as discussed in an NPR article (Smith). Using AI would involve spending a lot of time training the system, and even after that, it is often unclear as to what exactly the AI program is responding to. And in any case, a one-number score does not provide any diagnostic information.

2. The Problem I Propose to Address

In this paper, I propose a way to develop a computer program to test students' basic sentence forming abilities in a communicative way. By "communicative" I mean two things: (1) It will require students to generate sentences from scratch, which is what people must do in normal communications, and (2) the testing method will be able to give partial marks for answers that are not fully correct, in a consistent way. Communications involve mistakes, but listeners are able to tolerate a variety of mistakes; a communicative testing method should be able to distinguish an answer with 'recoverable' mistakes versus a completely unintelligible one. The idea of a 'recoverable' error in real life is dependent on many factors. In this paper, it will be taken to mean an error in word order or auxiliary verb usage or other ideas developed below. To limit the length of the discussion, I will use the simple present tense as our domain/area of assessment. It is hoped that other tenses and other complex and interrelated grammar points could be tested in a similar way.

This method is being developed to test the efficacy of a language practice drill technique that I have developed (which will be the subject of a later paper). The technique is designed to help learners (re) learn the basics of English sentence formation, and become fluent/automatic at that basic level. After working through the technique, student skill improvement was noticed informally; the time taken to perform the drill became shorter as a function of repetition, and it was noticed that students in one class became able to correct their

own mistakes in writing that they had done prior to doing the drill practice. What I wanted was a more detailed method to see exactly what students could do or not do with regards to sentence forming skill; hard data from both before and after the drill. To that end, what I propose in this paper is a method somewhat similar to the re-arrange-the-words method discussed above, but having students enter all the words, selecting from a large pool of words and inflections, with many distractors, in order to make a sentence.

The program would, in its limited domain, check for student understanding of basic mechanics of basic sentence formation, looking at such things as selection and placement of auxiliary verbs, proper placement of 'not' in negative formation, usage and placement of third person 's', and other grammar points, as outlined in section 5. It would also collect and organize data for each student about what they do or do not understand/use correctly. To further aid the teacher in spotting common problems, answers should be able to be divided into groups of similar mistakes. Teachers will be able to make tests for the specific points they want to check for, and control the length of the test, to fit their needs. For grading purposes, a single-number score could also be generated.

It is hoped that this will be a useful test for rapid formative or summative assessment of basic student English competence. It will be machine-gradable like multiple-choice or matching questions, with the same advantages of quick turn-around time and ease for teachers. It will also have the advantage of being able to test how well students understand the whole picture of sentence formation, in a way that, until now, has required much more time to grade. Test problem design will also be simplified. It is not necessary to think up credible but wrong options for each multiple-choice question, but merely prompts and/or sentences to translate. In this way, it will be easy for teachers to tailor tests to their curriculum.

3. Test Question Design and Format

The basic format of test questions is uniform throughout the test, as influenced by the input method discussed in section 4:

_____ (subject)_____ _____ _____ _____ (complement/object (if any)).

To avoid giving hints, blank length is uniform, as is the number of blanks. Not all blanks need to be filled. Changing the input method will allow an arbitrary number of blanks/words to be used (as discussed in section 6). The instructions to students can be of two kinds; In type (1) the subject and object are given along with instructions to form an affirmative sentence, a negative sentence, or a yes/no question. In this type of test problem, there is nothing to translate. If expanded to multiple tenses, hints as to tense such as 'always', 'yesterday', or 'right now' can be given to check understanding of which tense is best.

Sample question (1)

31-35　(wearの正し形をつかって) 肯定文 (affirmative sentence)
_____ Nancy always _____ _____ _____ _____ nice clothes.
(Note that 'always' is a hint to use the the present simple tense).

Instruction type (2) provides a Japanese sentence to translate into English.

Sample question (2)

46-50 Bennaはコーヒーを好きですか？
_____ Benna _____ _____ _____ _____ coffee?

Only 2 problems are provided as an example here, but the scope envisioned for this method is present tense regular affirmative sentences, negative sentences, and yes/no questions. Additionally, each of those three cases can have one of at least two different types of subjects: third person singular (he/she/it) and everything else (I/you/we/they). That is six cases for regular verbs, and the "you/we/they" case needs to be separately checked if you are using the 'be' verb, for another 9 cases, for a total of 15 questions to test every case in the

present tense. If included at a later stage, the past tense would require 3+6 questions to completely check and the present continuous would require another 9 questions to fully check.

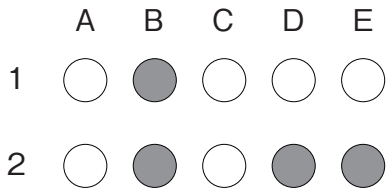
4. Input Method

The input method initially used was based on students' coding of their input choices on machine readable multiple-choice test papers. This method was chosen because students do not need a computer to write the test, and the questions can be of a wide variety. As the computer program is not complete, I will refrain from rigorously evaluating test validity here, but informally, the input method works, and when test questions as described in this paper were included as part of a final exam even the simple analysis method (in section 5 below) generally gave good results for students who scored well on the other part of the final exam, and low scores for students who did poorly. However, the input method is complex, and requires a lot of effort from students over and above using the English, and it is prone to serious errors. Therefore, a different input method will be used as discussed in section 6. The original method is included for completeness immediately below in this section, but may be skipped. The analysis methods and goals, in section 5 below, will be valid regardless of input method.

The idea for the method grew out of the capabilities of ZipGrade, an inexpensive "Scantron"-like smartphone application (ZipGrade Top Page). It uses the phone's camera to scan a multiple-choice test paper, and compares it to a pre-entered answer key to give an instant score. There are many such applications, but this one is useful for two specific reasons (among others);

(1.) It can read up to three circles per five circle set (one question), as in Figure 1.

Figure 1: Examples of acceptable answers using ZipGrade.



This is very useful because instead of students having to choose from 5 options, now there can be up to 21 choices, (a b c d e, ab ac ad ae bc bd be cd ce de, abc abd abe acd ace ade) plus the option of leaving it blank. This reduces the chances of randomly selecting the correct answer, and gives more choices of words to use, thus increasing authenticity.

(2.) It will export the contents of the student answers to a CSV file. There are a few formats available, including one in which the exact answers the students entered are all available if desired. This is useful because although the software is useful enough for everyday tests, the data can also be more deeply analyzed with a spreadsheet or other software.

Students complete the test by entering the codes for the words they wish to use selected from the ' Word Bank' (Table 1 below), and entering the codes onto the multiple-choice sheet. The format of each question is the same: a blank in front of the subject (which is given), and 4 blanks between the subject and object, if any (as mentioned in section 3). There can be more blanks per question, but to avoid confusion, it is useful to go up by fives. (Even so, it is easy to get confused). That way, test problems will follow the pattern of starting at position 1, 6, 11, 16, 21 and so on, with one blank corresponding to one answer position on the mark sheet, as

Table 1. ' Word Bank' (words and inflections students can use to make sentences, with their codes)

左詰めに書いて下さい。(enter blanks from the left)
使わない欄は　なにも書かない。 (leave unused blanks blank)

not	is	am	are	has	have	do	does	~s/es (does 以外)	did	~ed	~ing
A	B	C	D	E	AB	AC	AD	AE	BC	BD	BE

you	to	how	who	wear	meet	come	watch	want	it
CD	CE	DE	ABC	ABD	ABE	BCD	BCE	BDE	CDE

added to sample question (1) below Table 1. In this manner, a 100-question form could test 20 test problems/sentences.

Sample question (1) with numbers added to specify test sheet positions

31-35 (wearの正し形をつかって) 肯定文 (affirmative sentence)

_____ Nancy always _____ _____ _____ _____ nice clothes.

Blank： 31 32 33 34 35

5. Analysis Methods and Goals

After students complete the test, the answers will be analyzed by the computer program.

Of course, the easiest thing to analyze is the completely correct answer, which checks that each of the blanks is correctly filled and gives a point only if they are all correct. Next, partially correct answers could be checked using a simple algorithm that gives points for each blank that has the correct word in it. However, that gives less diagnostic information, and a less full picture of student understanding.

A more complex algorithm is to look at how students form sentences, and extract pinpoint information as to what they do and don't understand. Specifically, what I want to check when testing student understanding of the present simple tense for example, would be

(1.) Use of correct verb. Usually this will be simple, but teachers can write problems to test commonly confused verbs, for example 'have' vs 'be verb' e.g. "She is long hair." or "She has cute." or "She has a teacher" when talking about their occupation.

(2.) Presence/absence of third person 's' (depending on subject and tense).

(3.) Proper placement of third person 's' i.e. on the main verb or auxiliary verb (do) if there is one.

(4.) Proper use of auxiliary verb 'do'.

(5.) Proper selection of auxiliary verb. Present tense needs 'do', but students often use a form of 'be'. Other tenses may require other verbs.

(6.) Proper placement of auxiliary verb.

(7.) Correct placement and usage of 'not'.

By breaking down student answers and checking many different points for each answer, and then tabulating the results, it will be possible to understand exactly what students do and don't understand, in which types of sentences. This can lead to more efficient instruction. For example, a student who often uses the wrong auxiliary verb in the correct position and case (i.e. negative or question) at least partially understands usage of auxiliary verbs. The next step would be to help students understand which auxiliary verb to use.

It will also be possible for teachers to give positive feedback to students; it is encouraging for students to realize that they understand at least part of what they need to know. By looking at results for the class, and having the software group common mistakes, the teacher can easily decide which points are worth working on as a class, and which are better dealt with on a small group or individual basis.

Here I will present an outline for a method and computer program that will be able to analyze sentences in the simple present tense, in questions, negatives, and affirmative sentences, in first, second, or third person, using the 'be' verb or selected regular verbs (with dictionary form not ending in 's' or 'd'). When generating a test question, the teacher will enter the information as per Table 2, and the program will look at a student's answers and provide a detailed analysis, using various 'flags' that indicate correct or incorrect usage of various details of sentence construction and what they did correctly and incorrectly on each question, and use those flags to provide statistics to the teacher about their overall strengths and weaknesses. If used on a group of students, it will provide information for the group organized by test question and overall group strengths and weaknesses.

Table 2. Data to be Entered for Each Test Question

Problem type:		question, negative, affirmative
Tense	:	present, (past)
Subject	:	first (I), (you/we/they), or 3rd person singular (he/she. . .)
Verb	:	be, regular verb (not ending in s or d)

We will use sample 1 from above as an example.

31-35 （wearの正し形をつかって）肯定文 (affirmative sentence)

_____ Nancy always _____ _____ _____ _____ nice clothes.

Blank: 31 32 33 34 35

If the code/mark sheet is used, the correct answer would be

31-35 （wearの正し形をつかって）肯定文 (affirmative sentence)

_____ Nancy always ABD AE _____ _____ nice clothes.

Blank: 31 32 33 34 35

Which, per Table 1 above, corresponds to

31-35 （wearの正し形をつかって）肯定文 (affirmative sentence)

_____ Nancy always wear s _____ _____ nice clothes.

Blank: 31 32 33 34 35

(A handwriting font used to signify that it is a student answer).

From this point we will use the words from the table versus the multiple-choice answers, for ease of visualization. Using Google forms or other software, input would be in the form of words with an arbitrary number of blanks before and after the subject. In this case the 's' in the example above (or any other inflections) would be directly attached to the word before it, which we will assume for the remainder of this outline. Whatever input method is chosen, it might be expedient to limit input to the contents of a box like Table 1, with a variety of

common words, inflections and auxiliary verbs; both the necessary ones and distractors.

Check 1: Perfect Answer

The computer would first of all compare the student answer to the correct answer,

∅, wears

∅, called the 'null' character, is used to signify that the correct answer is a blank. The program will look at two positions, before and after the subject. If they match, the program will set the 'perfect answer' flag and also other relevant flags indicating understanding, and go to the next question.

Check 2: Correct Main Verb Choice

If the answer is not perfect the program will check for the presence of the correct main verb, setting a flag 'chose correct main verb' if it finds a form of it. Usually the choice will be clear, but as mentioned above, some pairs are easy to confuse. Inflections are checked later. It would also check the main verb is entered after the subject, and set a relevant flag as appropriate.

Check 3: Question formation

Next, the computer goes on to check for a ∅ in the first position (before the subject). If the test problem is not a question, as entered into the information box for that problem, that place should be empty. If it is empty, it implies that the student understands that the subject must come first and a flag would be marked as such. If the test problem is a question, the program would check for the presence of a auxiliary verb, mark a flag to indicate if there is, next it would check for the correct auxiliary verb ('be' or 'do'), setting flags as necessary. (third person 's' would be analyzed later.)

Check 4: Negative formation

The program needs to check for proper usage of the negative (i.e. 'not'). It will check the teacher-entered data about the test problem. If it is a negative, it will ensure that 'not' is used once and only once. Flags would be set for 'student did not use 'not'', 'student used 'not' twice

or more', or 'student used 'not' in an affirmative or question' as appropriate. Next it will check placement. If the test question is a 'be' verb, the 'not' needs to come directly after the 'be' verb, or directly after 'do'/ 'does' if it is not a 'be' verb'. Flags would be set for correct placement. Finally, auxiliary verb usage would be tested for. Regular verbs would use 'do' in front of the 'not', (third person 's' is checked for later) and a form of the 'be' verb also would be in front of the 'not'. Flags would be set for correct use, and also in case of incorrect use of 'do' in a 'be' verb sentence. Note that it would simplify program construction if the contraction "n't" were disallowed.

Check 5: Singular Third person 's'

Next, the program could count incidents of third person 's'. The easiest way to do this is to check how many times the final letter in each word is 's', correcting for the presence of 'was'. (Note: It simplifies the program if no words from the selection box end in 's' in the dictionary form.) If the test problem uses the present tense, and a singular third person subject, the correct answer is 1. If it is one, it indicates the student correctly understand that a third person 's' would be necessary if this was a present tense. If the test problem does not have a singular third person subject, there should be no 's'. Two words ending in 's' would indicate a problem in understanding as well. If the test problem does not have a singular third person subject, there should be no 's'.

Next, the program would need to check for placement of the third person 's'. In an affirmative sentence, the 's' belongs on the end of the main verb. In a negative sentence, it belongs on the auxiliary verb, in front of 'not', and in a yes/no question it belongs on the first word in front of the subject. If the relevant condition is met, the appropriate spreadsheet flag 'understands placement of 's' in affirmative/negative/question formation' would be set. Failing that, the program could check the main verb in a question/negative sentence, setting flags as appropriate.

Check 6: Auxiliary Verb (Re)check and Final Check

Next the program should check for the presence/absence of auxiliary verbs, their correct choice, and correct location. In general, if the problem is a question, the auxiliary verb (main

verb if it is the 'be' verb) needs to come before the subject. The flag in this case would be, 'auxiliary verb correctly before subject in question' or ' "be" verb correctly before subject in question' if it is correct, and 'auxiliary verb ["be" verb] not in front of subject in question' if it is not. If it is not a question but the auxiliary verb is in front of the subject, flags would be set for those cases. If it is a negative sentence, the auxiliary verb should be after the subject and before 'not' (flag 'auxiliary verb in correct place in negative sentence'), as analyzed in check 4. If it is a regular affirmative sentence in present simple tense, there should be no auxiliary verb (flag 'understands that an auxiliary verb is not necessary in an affirmative present sentence'). Finally, the program would check for incorrect helper verbs and other words that should not be there, for easy display and later human analysis.

As the program is written and field tested, the outline above will surely have to be modified and more detail added, both to correctly analyze as many cases as possible, and also to provide the maximum useful and easy to understand information to students and teachers who use it.

6. Conclusions and Ideas for the Next Step

The idea of being able to flexibly analyze and gather detailed data on students' mistakes to see where they do and don't understand the interconnected rules of grammar is very attractive. Since a computer per student is not necessary, using the multiple-choice sheets and powerful optical reader capabilities of ZipGrade seemed to be a way to enter data to this end. However, after administering the test, it became clear that although many students could successfully fill in the test, there are many ways for students to make mistakes in entering their answers, even though they are able to answer the question correctly (as evidenced by looking at their question sheet), and in general, it is too much extra work for the students.

A second problem is that of how to use a computer to analyze the data. Initially I used a spreadsheet, but it was very difficult to build a spreadsheet to adequately analyze the students' work, and to help visualize common problems.

Going forward, I propose that using a word processor, or something like Google forms, or even incorporating a program/web page with a 'drag and drop' interface, would be a much more user friendly input method for the students. The use of the word box developed above would still be useful to limit the number of possibilities somewhat, and the question designs can be kept as well. Numbered blanks would no longer be necessary, nor would coding. Students could use their smartphones to do the test, or if possible, a computer with a keyboard would be the easiest. To help the teacher grade the answers and analyze the data, a computer program designed according to the above outline needs to be written.

Thinking even farther, by using the diagnostic output to give feedback to the student and automatically select sentence types that the student doesn't do well with for further review, students will be able to study by themselves in an efficient and effective fashion. This would combine learning with assessment. In this day and age of increased testing for assessment, it is necessary to combine learning with assessment where possible to make efficient use of limited class and homework time. The method of assessment described here will be a useful step towards this purpose.

Works Consulted

Menu page for a free trial English lesson. FREE TRIAL intermediate level lesson 2www.quia.com/pages/schi33/page5

Qu, Yanxuan., et al. "Evaluating the Stability of Test Score Means for the TOEIC® Speaking and Writing Tests." *ETS Research Report* RR-17-50 (2017):1-10. Web. 25 Sep. 2018.

Singh, Upasana & Villiers, Ruth. "Investigating the use of different kinds of multiple choice questions in electronic assessment (e-assessment)." *Progressio*. 34 (2012): 125-143. Web. 20 Feb. 2018.

Smith, Tovia. "More States Opting to 'Robo-Grade' Student Essays By Computer." *NPR*, National Public Radio, Inc., 30 June 2018, www.npr.org/2018/06/30/624373367/more-states-opting-to-robo-grade-student-essays-by-computer.

Van Els, Theo., et al. *Applied Linguistics and the Learning and Teaching of Foreign Languages*. Kent:Arnold, 1984. Print.

ZipGrade Top Page. *ZipGrade* , ZipGrade LLC, Zipgrade.com.