

Automatic scoring of foreign language textual and spoken responses



Automatic scoring of foreign language textual and spoken responses

What are foreign language textual and spoken responses?

“Foreign language textual and spoken responses” refers to the production by students of written sentences in a second language as well as referring to the production of spoken responses to listening exercises conducted in a language other than the test takers’ native language.

This study reports on how two teams have designed and implemented technology-based solutions to provide automatic scoring of students’ spoken and written responses.

Pearson’s Versant™ Tests

Contact details

Jared Bernstein and Alistair Van Moere,
Address: Pearson, 299 S California Avenue, Palo Alto, CA 94306, USA.
Email: avanmoere@ordinate.com jared@ordinate.com

Brief details

Versant™ Tests (published by Pearson) provide an assessment of students’ spoken language proficiency. In this case study we look in depth at one of Versant’s spoken language tests – the Versant Aviation English Test (VAET) which measures the ability of test takers to understand aviation and plain English and to respond intelligently in spoken English at a full-functional pace. Pearson produces a number of other assessments in the Versant brand covering English, Arabic, Dutch and Spanish. These assessments all follow a similar approach, but their uses include: assessment of primary school children’s reading fluency; Dutch

language proficiency tests used by the Dutch government in their citizenship requirements; spoken language tests to measure students’ language gains in Korea and Japan; assessment of aviators’ language skills for professional licensing; and decision-making for admissions to colleges in the USA. One other high-stakes use of the test is work recruitment and placement (eg call centre staff), as implemented by a variety of household names, such as Dell, IBM, Accenture, Network Rail, Malaysian Airlines, and Emirates Airlines. The tests are also used in program evaluation in higher education courses (Blake et al, 2008).

What was the problem?

The Versant assessment system was developed to fill a major gap in the test supply market. Although there are multiple providers of assessments and tests of written language proficiency, these tests are all based on requiring test takers to write their responses. Traditionally there have been very few assessments and tests of students’ spoken language proficiency or of their listening skills. The major reason for this gap was related to cost – in a human-based assessment and testing environment, it is often prohibitively expensive to provide a human marker to listen to a test taker, even if the test takers’ responses are recorded and sent to the human marker. However, with a technology-based solution, the economics of speaking and listening tests become significantly more acceptable.

The Versant assessments originate from the late 1980s. Originally the idea was to develop an assessment system to support teaching and learning – ie a formative assessment tool. However, since those early developments, the emphasis has moved to stand-alone testing.

The solution

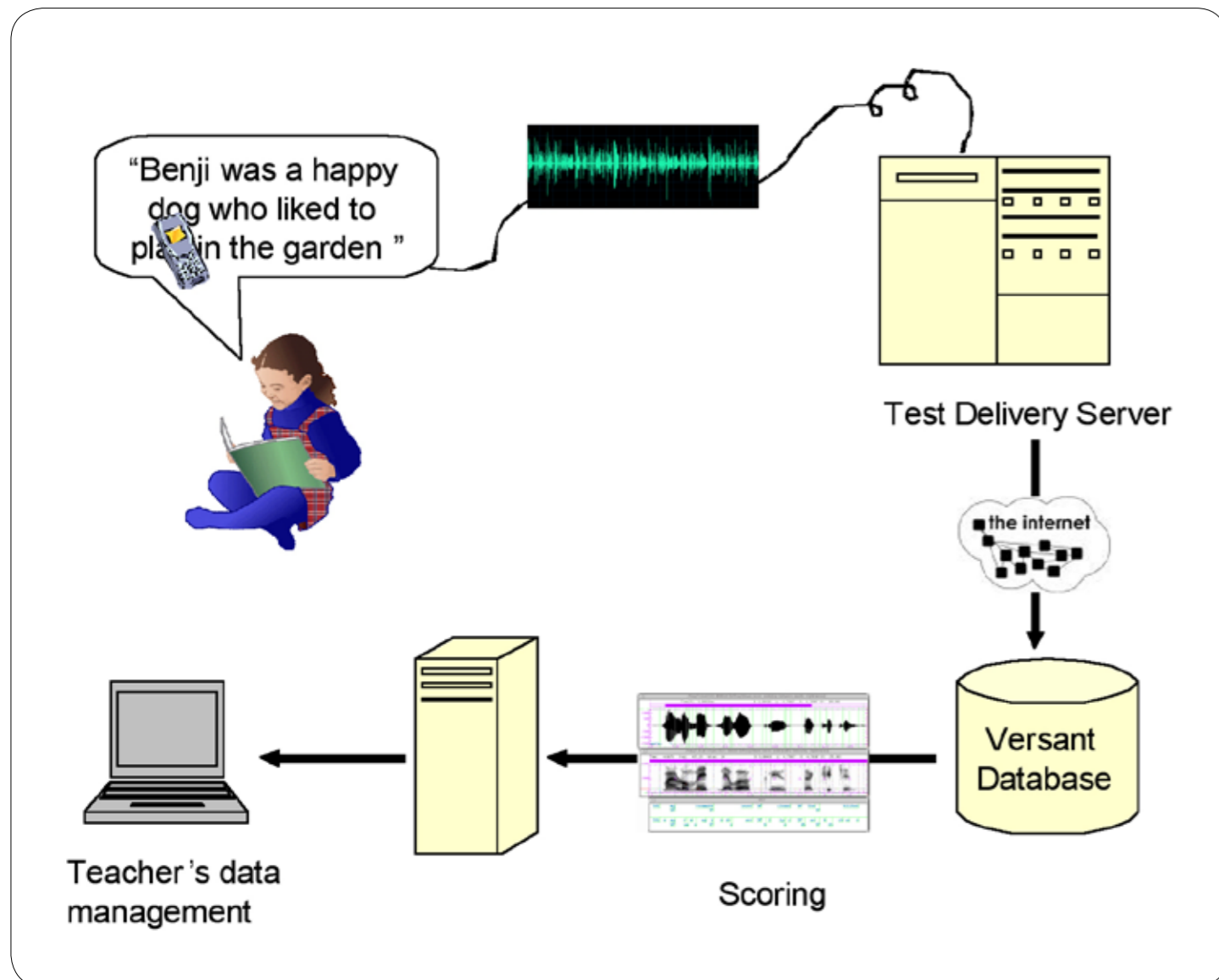
The Versant speaking tests are delivered over the telephone and/or on a computer. During the test, the test taker hears a number of prompts and questions designed to elicit responses in the target language.

The test taker commences a test session by opening a test paper (on screen) and dialling a telephone number given on the paper. The system prompts the test taker to provide an identification number. Once this information is supplied, the test commences. Instructions are spoken by an examiner voice. The test items (or questions) are spoken by numerous different speakers, thereby exposing the test taker to a range of different speech speeds and styles. Test lengths vary: up to 2 minutes for a child's reading

test, up to 15 minutes for a general language proficiency test, and up to 30 minutes for a high-stakes aviator's test.

The Versant system design

The test sessions include a range of tasks, depending on the skills that are to be tested. The Versant Aviation English Test serves as useful example for demonstrating the various kinds of tasks which can be administered. Behind the technology are language assessment specialists who customise the technological capabilities for specific purposes; the aviation test, for example, was designed in collaboration with the Federal Administration Authority (FAA) of the United States to ensure that it assessed the desired abilities in test takers.



For the Versant Aviation English Test there are a total of eight sections, each with their own type of activity. The following describes the sections, with examples of each type of activity.

Section A: Aviation Reading

The test taker reads aloud sentences that are written on the test paper. The sentences use aviation phraseology and vocabulary. These sentences are of varying length and become progressively more demanding. For example, the candidate might see three sentences written on the test paper:

1. World Air 891, request descent.
2. World Air 891, maintain flight level 280 expect descent shortly.
3. Maintaining flight level 280, World Air 891.

and will hear "Please read sentence number three."

Section B: Common English Reading

In contrast to Section A, Section B comprises commonly used English language sentences. For example, the test taker might see three sentences written on the test paper, and hear "Please read sentence number one."

Section C: Repeat

In the Repeat task, the test taker listens to a complete sentence then attempts to repeat it accurately. The sentences are based on commonly-used phrases and are of increasing difficulty. For example, the test taker might hear sentences such as: "Visibility is very poor today", "Do you happen to know what time it is?" or "I try to get there a couple of hours before the scheduled departure." In Part C, the test taker is instructed to repeat the sentences heard.

Section D: Short Answer Questions

In the Short Answer Questions section, the test taker listens to information and then responds to spoken questions which are based on the given information. For example, the test taker might hear "I have a pair of suitcases. How many suitcases do I have?" and would answer "Two", "You have two" or similar answer. As a second example of a short answer question, the test taker might hear "Is land that's almost entirely surrounded by water a peninsula or a pond?" and would answer "Peninsula", "A peninsula" or similar answer.

Section E: Readback

In the Readback section, the test taker is provided with a test paper containing a few relevant phrases. The test taker will then hear a short message and is required to provide an appropriate

response using some of the phrases given on the test paper. For example, the test taker might be provided with the following air traffic control-related phrases:

1. Coastal Air 315
2. World Air 395
3. Coastal Air 405

The test taker would then hear a message such as "Coastal Air 315, maintain flight level 070", and would be expected to respond by saying "Maintain flight level 070, Coastal Air 315", or "Maintaining flight level 070, Coastal Air 315."

Section F: Corrections and Confirmations

In the Corrections and Confirmations section, the test taker listens to a message, which can take either the pilot's perspective or the air traffic controller's perspective. The test taker would be provided with a number of call signs on a test paper:

1. Charlie Romeo 4013
2. Coastal Airline 445
3. World Air 2043

Then the test taker hears a read-back of the message, which will contain the correct information, wrong information, or a request for more information. The test taker must respond appropriately, using accepted air traffic phraseology. For example, the test taker might hear the following two speakers:

(Speaker 1) "Charlie Romeo 4013, continue descent to flight level 110, report passing 150."

(Speaker 2) "Descending to flight level 10 thousand, report passing 15 thousand, Charlie Romeo 4013."

In this example, the test taker would be expected to correct the errors in the second speaker's response, using accepted air traffic controller phraseology, by saying "Charlie Romeo 4013, negative, continue descent to flight level 110, report passing 150", or alternatively by saying "Charlie Romeo 4013, I say again, continue descent to flight level 110, report passing 150."

In this section of the test, some items reflect routine communications/situations, while others cover less routine communications/situations. A small proportion explores unexpected communications/situations as well. The immediacy and appropriateness of the responses as well as the information conveyed in them are important factors in estimating the test taker's ability to manage speaker/listener interactions.

Section G: Story Telling

In this section of the test, the test taker is required to listen to a short passage, then retell the passage in their own words. For example, the test taker might hear the following passage:

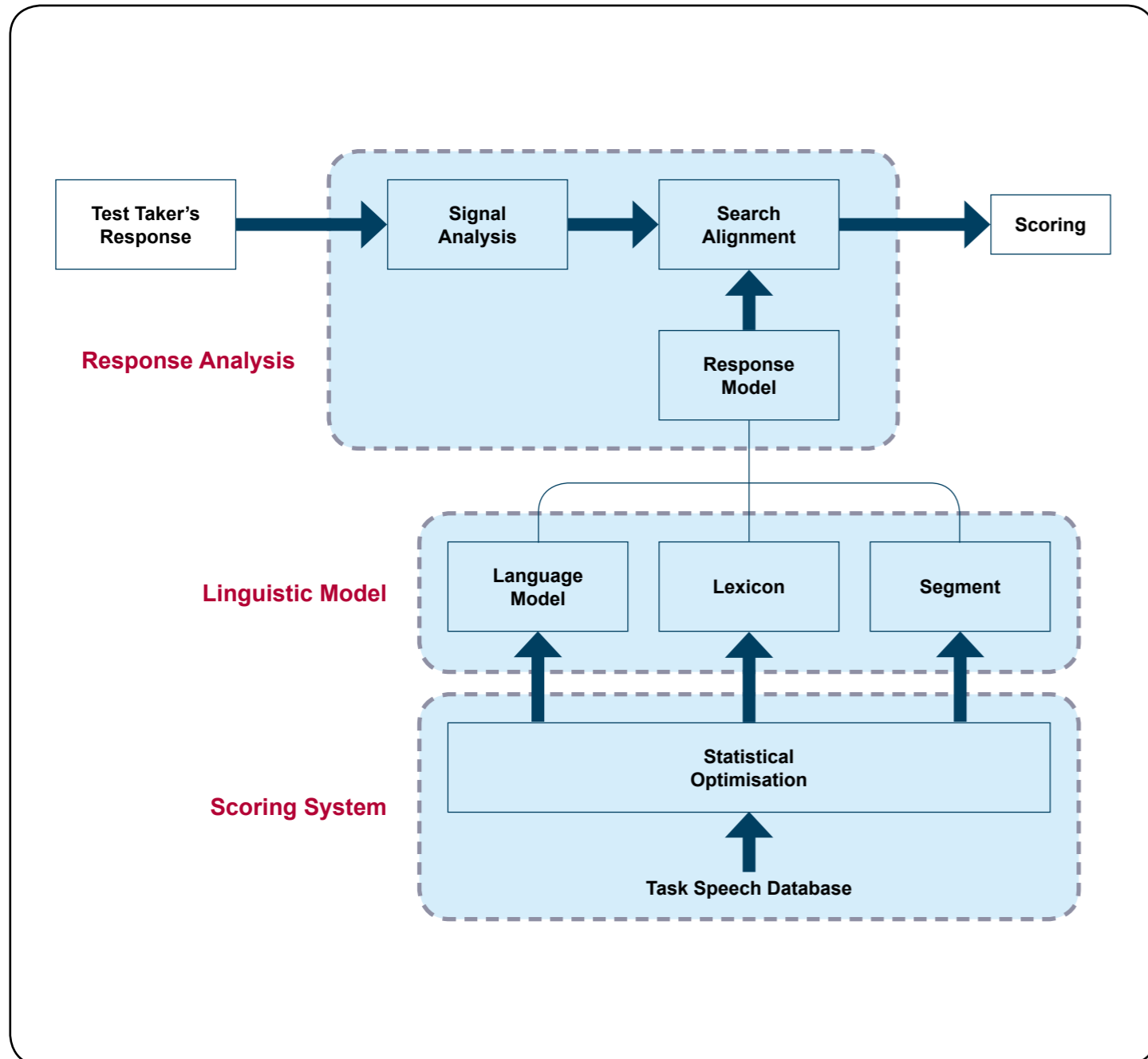
"Most of the flight between Tokyo and Hawaii was calm. An hour into the flight, the pilot encountered a line of storms which she had to fly over. The flight became bumpy and the passengers were nervous but the flight attendants kept everyone in their seats. After a few minutes, the plane cleared the storms and the rest of the flight was smooth. Several hours later, the plane arrived safely in Hawaii."

There is one further section in the Versant Aviation English Test. This requires test takers to respond to an open question, such as "In your experience, what kinds of weather are the most difficult for air traffic management and why? Please explain." Test takers' responses in this section are not currently scored by Pearson, although the responses are made available to auditors.

Scoring the Pearson Versant test spoken tests

Test takers' responses are scored by the Versant system by comparing the test taker's response to a large database of proficient responses. This enables the system to respond "intelligently" to the dialect of the test taker, without requiring test takers to use any form of standard pronunciation. Currently, the database contains over 50 million responses, produced by a mix of proficient native speakers and also learner non-native speakers, which have been scored by the automated scoring algorithms.

The system includes a number of components, represented in the diagram below:



The system has three major components:

- The **Response System**, which captures and analyses the test taker's responses, and returns a score;
- The **Linguistic Model**, which hosts the linguistic speech models;
- The **Scoring System**, which holds statistical scoring models that were "trained" with a large sample of both learners' and proficient speakers' responses.

The Response System

The purpose of the Versant Response System is to provide a reliable measure of the test takers' spoken language proficiency. Once a test taker has provided a response, two linguistic measures are applied to the test takers' response:

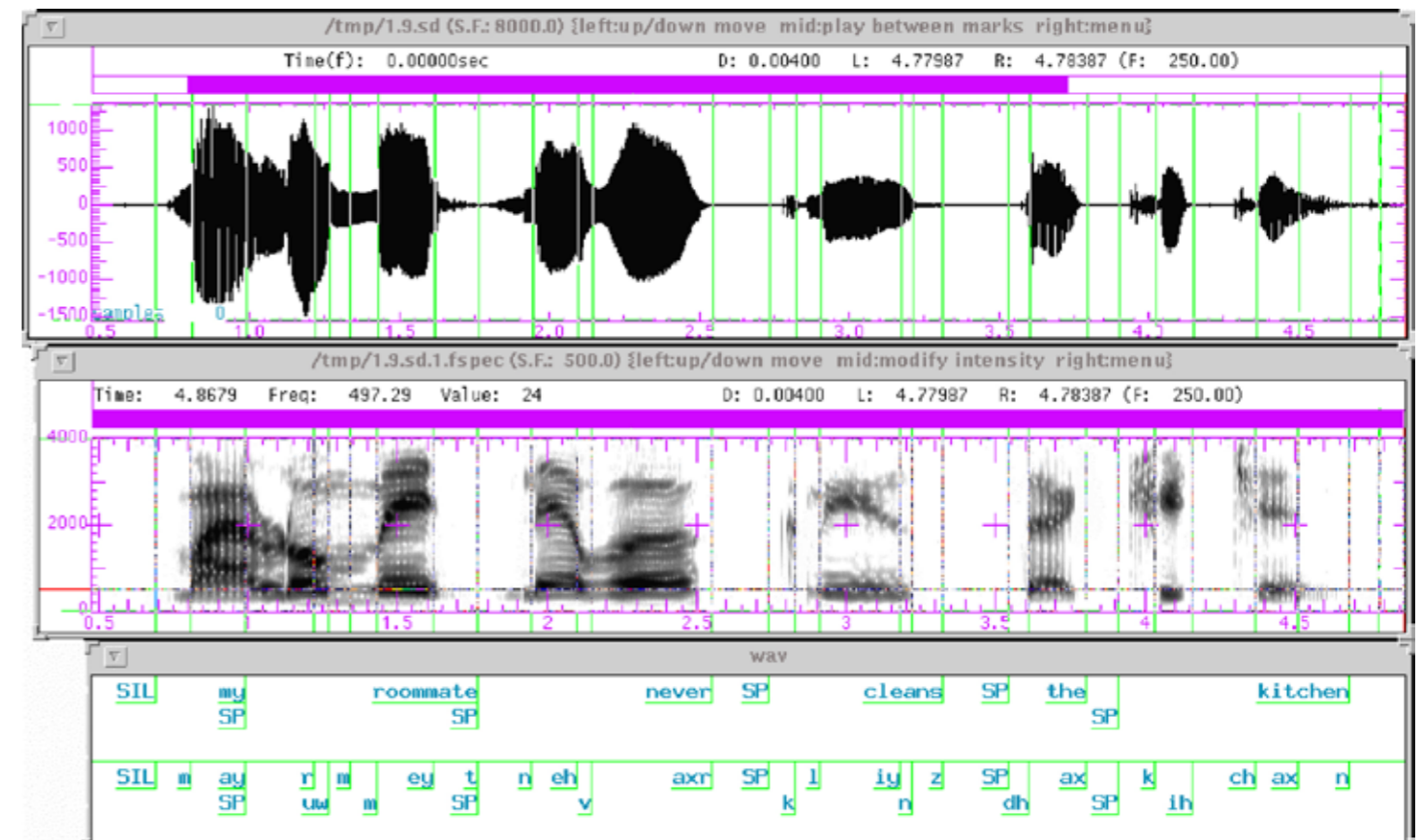
1. The content of the response. This measures aspects of the responses such as word order.
2. The manner of the response. This covers the fluency and pronunciation quality of the response.

The ways in which these measures are applied can be shown through the following representations of a test taker's response. In this example, the system is analysing the sentence "My roommate never cleans the kitchen." The first mapping consists of a waveform representation of the spoken sentence; this consists of a time-series plot of the energy of sounds produced by the test taker. The second mapping is a spectrogram that represents the pitch pattern in the response, where darker

shading represents more stress placed on the phoneme. The third plot represents the phone sounds which were actually "understood" by the speech recogniser ("My roommate never cleans the kitchen"). From this, the Versant speech processors can disambiguate a number of features of the spoken responses, such as duration, hesitations, syllables, clarity of phone production and some subphonemic sounds are recorded and measured.

Pearson uses a patented system to capture and evaluate the test taker's response. The speech processing capability is based on a Hidden Markov Model (HMM)-based speech recognizer developed from the HMM Tool Kit (Young, Kershaw, Odell, Ollason, Waltchew, and Woodland, 2000).

The acoustic models for the speech recogniser (models of each sound in the language) have been trained on data from a diverse sample of non-native speakers of English. In this way, the speech recogniser is optimized for various types of non-native speakers' accented speech patterns and the machine generally recognises response words as well as or better than a naive listener, but does not generally do as well as a trained listener who knows the language content. The speech recogniser also uses language models that represent not only the correct answers, but also the errors and disfluencies that are common for non-native English speakers for each item.



HTK – Hidden Markov Model Toolkit

HTK (Hidden Markov Model Toolkit) is a software toolkit for handling Hidden Markov Models (HMMs). It is mainly intended for speech recognition, but has been used in many other pattern recognition applications that employ HMMs.

HTK is designed for research in automatic speech recognition and has been used in many commercial and academic research groups for many years. It is based on Hidden Markov Models, which are used to model any time-series. HTK enables software engineers to develop systems capable of recognising and analysing speech time-series.

Spoken language is a series of symbols, which when constructed in (a correct) sequence can convey meaning. HTK reverses the process. It takes spoken phrases and sentences, and breaks them back down into a series of symbols converted from the wave-form representation of the spoken sentence.

As a speech recogniser, HTK seeks to find matches in these series of symbols between the test takers' responses and a database of anticipated responses.

HTK is patented software. More information is available from Professor Stephen Young, based at Cambridge University. See <http://htk.eng.cam.ac.uk/>

The Linguistic Model

Once the test taker's response to a test item has been encoded, the Pearson Versant system runs a number of comparative analyses. This includes three components:

1. A language model
2. A lexicon
3. Segment map

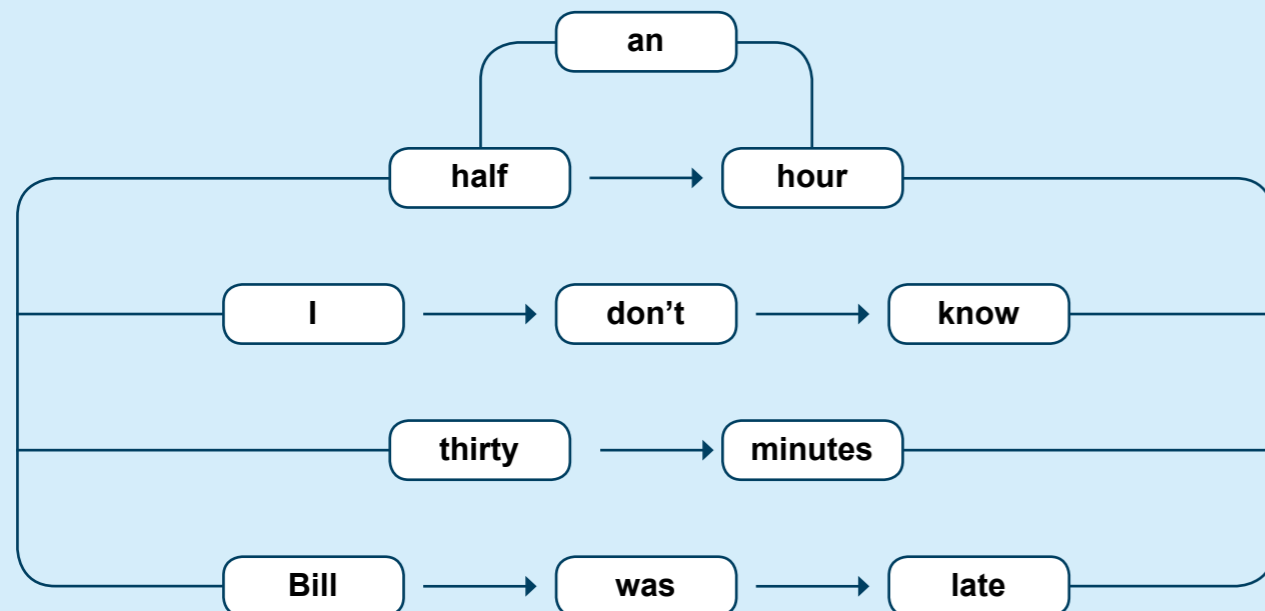
The language model provides a map of the sequence of the words that test takers are predicted to give. For example, if the test taker heard "Bill arrived at 1:30pm for his 1:00pm meeting. How late was Bill?" the system might predict any of the following responses:

- Half hour

- Half an hour
- 30 minutes
- Bill was late
- I don't know

A language model for this selection of predicted responses is given in the diagram below. These language models not only contain the most likely strings of words that test-takers are expected to say, but as mentioned above it also contains the types of mistakes and disfluencies that non-native English speakers are most likely to make.

The system also uses a lexicon, or dictionary, which lists the common pronunciations for each word that the system should recognise.



Scoring system

Using the acoustic models, dictionary, and language models, the speech recognition system employs statistical methods to identify the string of words that best matches the respondent's speech. The hypothesis of what the respondent said is then compared to the words in the item. Models based on Item Response Theory (IRT) use the 'correctness' of the content of individual responses in addition to the item's difficulty level to produce estimates of the test taker's abilities across six skill areas:

1. Pronunciation
2. Structure
3. Vocabulary
4. Fluency
5. Comprehension
6. Interaction

For example, in creating a score for Interaction, the system will score the test taker's responses in terms of the *Immediacy*, *Appropriateness* and *Information Conveyed* in the response. Other information is also extracted from the respondent's utterance such as speaking time, rate of speech, and mean pause duration. These and other paralinguistic parameters are then input into non-linear models that are optimized to predict how human listeners would judge the responses.

Pearson conducts ongoing validation studies on the Versant tests. The studies consistently show that the automated scoring algorithms evaluate test-takers' responses virtually the same way as trained, expert human evaluators. The correlation between machine scores and expert human judgments on the Versant tests is typically .97 (where 0 represents no relationship and 1 represents a perfect relationship) (Pearson, 2008).

Benefits

The major benefits of Pearson's Versant tests of spoken language are that they provide a practical, cost effective, and standardised means of testing key aspects of oral language proficiency. Human interview-based assessments are expensive, logistically difficult to schedule and administer, and difficult to scale. By contrast the Versant system operates on demand 24/7, administration and scoring is automatic and so trained examiners are not required, and test scores are available within minutes of taking the test. The test is available world-wide, meaning that test-takers in different locations can be administered a test and be confident that all are assessed using the same standardised scoring and reporting system, without the bias or error introduced by human examiners. Further, the tests are secure: each test is unique because it is made up of items drawn on-the-fly and semi-randomly from large item pools. This makes for an assessment solution for contexts where practicality, reliability and validity are key considerations. Pearson has accumulated a significant repertoire of assessments across a number of

languages and in different user contexts – citizenship, college entry requirements and professional competency testing, for example.

The Versant system has also designed an interesting solution to scoring spoken responses. The combination of pattern matching technology with a massive database of proficient responses is an approach that might well provide solutions in other assessment domains where the range of acceptable responses to an assessment item is very large.

LISC

Contact details

Alison Fowler
Address: LISC, University of Kent, Canterbury, Kent, CT2 7NZ
Email: A.M.L.Fowler@kent.ac.uk

Brief details

The second application features assessments of students' translation of sentences from English to other modern languages. The University of Kent provides such an assessment, called Language Independent Sequence Comparison (or LISC). LISC uses an automatic scoring approach to provide tutors with periodic assessments of students' linguistic understanding and skills. However, the system is fundamentally designed to provide the student taking the test with a marked-up error analysis of the first attempt, and permits the student to re-submit a second response. In doing this, LISC is designed to provide a formative element by giving students information about their errors and opportunity to improve. The original drive for LISC arose in the context of Spanish language courses at the university. However, the LISC approach is suitable for any language that carries its grammatical information at the end of words.

What was the problem?

The LISC assessments developed as a result of one person's drive and commitment. Alison Fowler was concerned that existing automatic scoring systems were based on natural language processing. While this approach works well in scoring good responses from students, it cannot provide analysis of weaker responses. LISC was therefore based on a string-matching approach, which can then provide test takers with a classification of the errors they have made in their translation. An additional specific drive for the LISC system arose from one professor in the Spanish Department at the University of Kent. At the time, he was teaching a group of 90 students and assigned them 2 translation exercises each week. This created a significant marking workload, which LISC was in part designed to alleviate.

The solution

Most computer-based foreign language testing and assessment programmes are based on short answer, gap-fill and multiple-choice types of question. These provide limited feedback to learners and rarely include error-detection.

Computer-Aided Language Learning (CALL) systems which are based on learners inputting whole sentences can provide a more authentic learning and assessment experience for the student. Existing systems which support this type of input typically have two weaknesses:

- The linguistic analysis is not capable of dealing with highly inaccurate responses from students, especially responses which are less than 50% accurate.
- The software has limited ability to detect and reward equivalent but not anticipated responses.

When development activity started on LISC, this was undertaken predominantly in spare time and during evenings. The requirements were to design a system which would:

- enable students to translate sentences from English to Spanish on screen;
- provide detailed marked-ups of student's errors;
- still cope with inaccurate and incomplete responses from students.

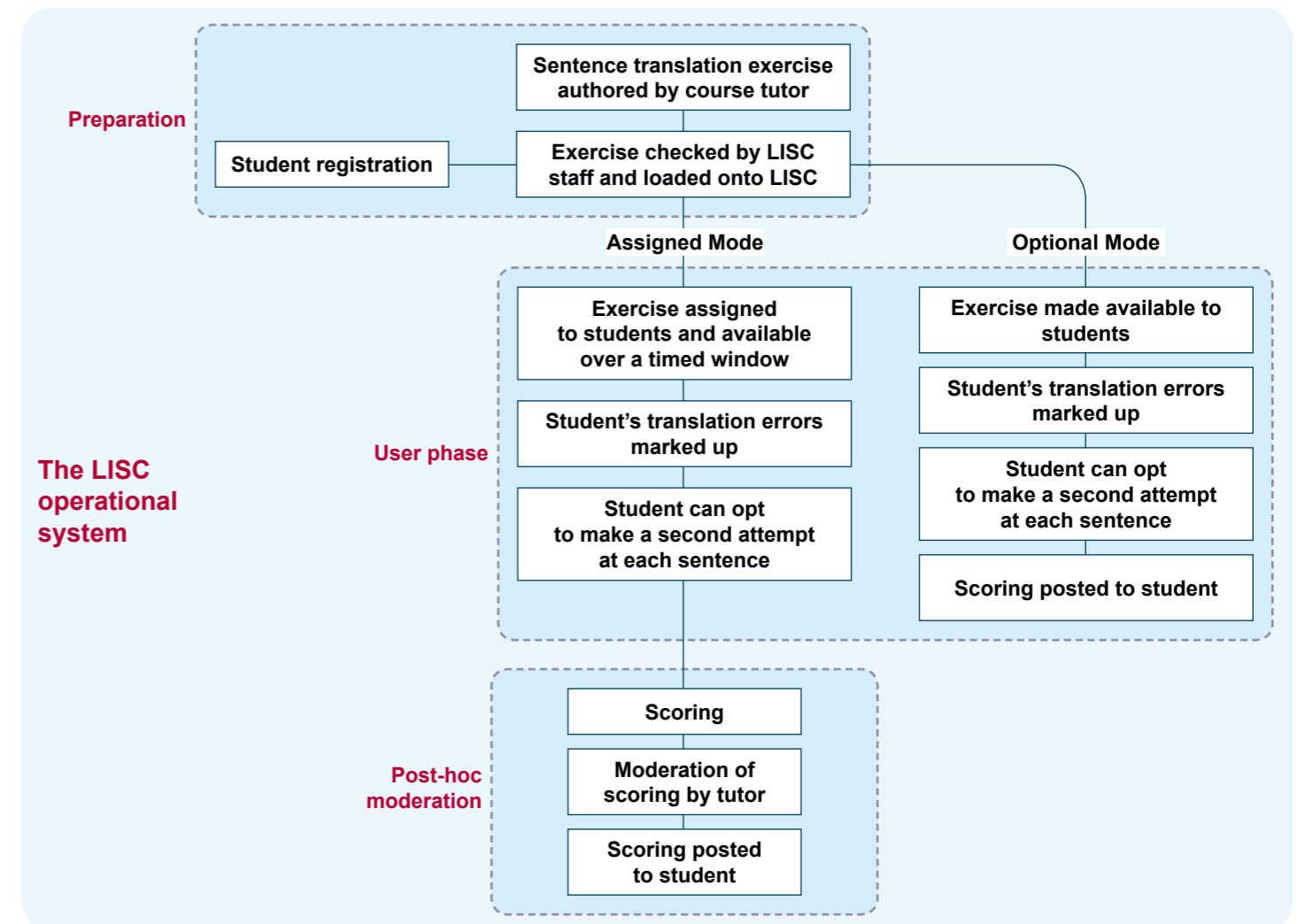
It should be noted that although the context of the initial requirement from the Spanish Department was for a system to support assessment of students' linguistic knowledge through sentence translation exercises, the design of the LISC system means that it can support a range of other types of exercise - provided there is a definitive set of acceptable answers - including:

- dictation
- rewording (e.g. active to passive)
- reading comprehension
- vocabulary testing

The LISC system has been designed to be language-independent – ie, it can operate in many different languages. Indeed, the only major constraint is that it supports only those languages which carry most of the grammatical information at the ends of words. But this covers a very wide range of European and other languages. In addition to being applicable to university-based language course, it could readily be applied to school-base courses such as GCSE and A-level. It is suitable for use in self-study and remote learning.

LISC – How it works.

The LISC system has been built, designed and is now managed by Alison Fowler. The following diagram summarises the main operational phases of the system.



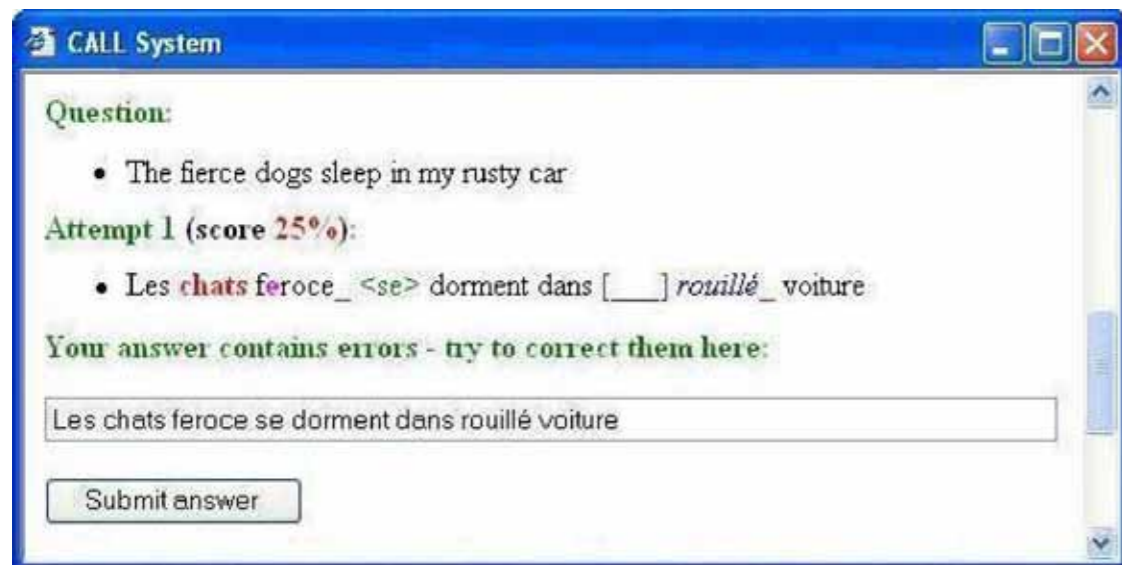
During preparation, tutor and student identities and permissions are set. Language tutors submit sentence translation exercises, and these are checked, prepared and loaded in the LISC system. This can take around 30 minutes for each 20-sentence exercise.

There are two modes of administration. Tutors can specify that they wish particular exercises to be completed by students within a specified period of time – perhaps over a week or two. In this **Assigned Mode**, students can log on whenever they wish, as many times as they wish and for as long as they wish to complete the translations. In **Optional Mode**, availability of the translation exercises is not constrained to a narrow window of time.

Once students have logged onto a LISC session, they are presented with sentences to translate, one at a time. For example, the following screen shot shows how the sentence “The fierce dogs sleep in my rusty car” would be presented to students.



Once the student has completed and submitted their first attempt at translation, the LISC system immediately sends back to the student a marked up copy of the first attempt. This mark-up uses a classification system (see below) to help students understand the nature of their errors. Note that the mark up does not provide hints or guidance as to the correct translation beyond marking the errors.



The system identifies and marks up the following categories of error:

- Incorrect words are shown in **bold red** font, e.g.:
– Je m'**ai** levé à dix heures (*I got up at ten o'clock*)
- Missing letters and erroneous spelling/conjugation in otherwise correct words are shown in **bold pink** font, e.g.:
– Ils ont admiré le **_**rideaux blanc**hes** (*They admired the white curtains*)
- Words which are correct but incorrectly placed in the answer are shown in **blue italic** font (or blue and pink if the word is also spelt/conjugated incorrectly), e.g.:
– Une petite **franciase** fille (*A little French girl*)
- Words omitted from the answer are denoted by pairs of brown square brackets: **[]**, e.g.:
– Elle pense que **[]** chiens sont féroces (*She thinks that dogs are fierce*)
- Words which are superfluous in an answer are shown in **<green angle brackets>**, e.g.:
– J'attendais **< pour >** Pierre (*I was waiting for Pierre*)

Under the bonnet of the LISC scoring system

LISC was developed using PROLOG over two years. PROLOG is used to host the database of translation sentences and to host the list of acceptable translations of those sentences. PROLOG also conducts the *string matching* of student's responses to those acceptable answers.

String matching is the process of identifying commonalities and differences in medium-length sequences of information (in the case of LISC, the information consists of the words, letters and spaces keyed by the student). Although PROLOG requires lengthy and detailed programming in order to analyse and score sentences, LISC has adopted PROLOG in preference to the main alternative approach, Natural Language Processing (NLP) of text. NLP approaches look for syntactical meaning in sentences, whereas the LISC approach in using PROLOG is to pattern match students' sentences against a set of pre-determined acceptable answers.

LISC's approach is based on a classification of student's translation errors:

- Incorrect word
- Errors within words, such as spelling mistakes
- Words that are in the wrong place in the sentence
- Missed words
- Unnecessary words

LISC begins by tokenising student responses. This involves breaking the sentence into a series of logical units, the most straightforward of which would be a word, separated from other words in the sentence by spaces. In this way a sentence comprises an array of tokens.

LISC then creates a first estimate of total possible score, by matching the tokens to the correct answer(s). The system then commences a series of back-tracks, through which it searches a tree of possible scores. Each potential branch is evaluated against the first estimated score, and if the branch yields a lower score it would be rejected. LISC completes a series of these comparisons, in each case staying with the higher of the possible scores, until all branches have been explored and evaluated. LISC then has produced the student's score for the sentence translation.

Evaluation of LISC: reliability and benefits

LISC was developed to meet specific local needs (defined by a Spanish tutor at the University of Kent). The original sponsor was clear that the sought-for benefit was to make savings in marking time. The system demonstrably achieves this. However, achievement of those savings requires an up-front investment of time. This is of course true in many developments and innovations – investment in development and refinement time is needed before the benefits can be realised. For an e-assessment product like LISC to be widely adopted, two of the essential pre-conditions are: that potential users (tutors) should see the benefits; and that those users should be sufficiently dis-satisfied with the current (hand-marking) arrangement to be motivated to invest time and effort.

The experience of LISC demonstrates how finely balanced these considerations can be. The system is no longer in use in the Spanish department, following three or four changes in staffing. At the same time, Alison Fowler has presented the system to a number of MFL departments in secondary schools. Two schools have adopted the system. Others have not done so, in Alison Fowler's view, because staff have not been convinced that they want to invest the time of providing sentence translation exercises to LISC.

LISC's experience provides the strongest of messages about the effects of human motivation and perceptions on the adoption of technology.

There can be little doubt that the LISC marking engine brings benefits. Its approach to scoring has been designed to ensure the closest of matches between a human expert's scoring of student's work and the evaluation provided by LISC. In addition to mirroring expert marking, LISC has a number of identified benefits:

Provides drill in grammar

In many technical aspects of learning, a major challenge for tutors is to identify students' misunderstandings and address them effectively. An unintended consequence of traditional drill exercises can be that a student's misunderstandings can be compounded throughout such exercises. By contrast, the LISC system provides a detailed mark-up of errors, helping students to learn rules of grammar in order to improve each translation. The system encourages the student to think about the error immediately and not repeat it. LISC's analysis of students' responses in translation exercises shows clear evidence of improvements in scores achieved, even over short 20-sentence exercises.

Students make good use of the software

Tutors and teachers using the LISC system with their students have indicated that students enjoy the on-screen approach. Students tend to be more diligent in completing assigned exercises, and there is evidence that students make greater use of optional exercises than would be the case with traditional paper-based approaches. In a questionnaire survey conducted by LISC, students identified a number of positive features of the system, including the immediacy of feedback and the helpful nature of the analytical feedback and the marked-up sentences.

Where next?

Alison Fowler has plans to enhance LISC in the future. These plans include the following:

- Creating an interface for tutors to upload sentences. Currently all translation exercises must be routed through Alison Fowler in order that they can be checked and uploaded. A tutor interface would be designed to improve the efficiency of uploading and to provide tutors with greater control. It could be that this would improve other tutors willingness to make the upfront investment to create translation exercises onto the system.
- The current student interface was developed using PERL. This works well for most of the time, but does not cope well if there is any interruption to the Internet connection during a student's translation session. In order to remedy this flaw, and in order to improve the interface generally, Alison Fowler is considering MOODLE.
- The third area of focus for the future is to extend the take-up of the system. Achieving this would require time and a marketing approach. It would also require the design of scaled up technical and human-support systems to cope with widening usage.

Further reading

Hincks, Rebecca, *Using speech recognition to evaluate skills in spoken English*. Lund University, Dept. of Linguistics, Working Papers 49 (2001), 58–61.

Available from [http://conference.sol.lu.se/fonetik2001/proceedings/bidrag15 .pdf](http://conference.sol.lu.se/fonetik2001/proceedings/bidrag15.pdf)

Jennifer Balogh, Jared Bernstein, Masa Suzuki, Pradeep Subbarayan, Matthew Lennig: *Automatically Scored Spoken Language Tests for Air Traffic Controllers and Pilots*. Harcourt Assessment, Menlo Park, California, USA.

Available from <http://www.icao.int/trainair/meetings/gtc10/pivp5.pdf>

For a full information on HTK and the Hidden Markov Model, see <http://htk.eng.cam.ac.uk/>

Steve Young, Dan Kershaw, Julian Odell, Dave Ollason, Valtcho Valtchev, Phil Woodland, *The HTK Handbook*. Available from <http://nesl.ee.ucla.edu/projects/ibadge/docs/ASR/htk/htkbook.pdf>

LISC Homepage <http://www.cs.kent.ac.uk/people/staff/amf/CALL/HTML/menu.html>

Aliy Fowler, *LISC: web-based CALL for Language Teaching*, University of Kent

Robert Blake, Nicole Wilson, Christina Pardo-Ballestar, Maria Cetto. Measuring oral proficiency in distance, face-to-face, and blended classrooms. *Language Learning and Technology*, 12, (2008), 114-127.

Available from: <http://llt.msu.edu/vol12num3/blakeetal.pdf>

Pearson. Versant English: Test Description and Validation Manual.

Available from: <https://www.ordinate.com/technology/validation.jsp>



Short answer marking engines