


The promise of NLP and speech processing technologies in language assessment

Language Testing
27(3) 301–315
© The Author(s) 2010
Reprints and permission:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/0265532210364405
<http://ltj.sagepub.com>


Carol A. Chapelle and Yoo-Ree Chung

Iowa State University, USA

Abstract

Advances in natural language processing (NLP) and automatic speech recognition and processing technologies offer new opportunities for language testing. Despite their potential uses on a range of language test item types, relatively little work has been done in this area, and it is therefore not well understood by test developers, researchers or users in language assessment. This paper introduces NLP for language assessment as an area of inquiry and practice by describing the historical roots coming from computational linguistics, statistical NLP, speech recognition and processing technologies, language assessment, and computer-assisted language learning. It outlines uses of NLP and speech recognition and processing technologies in language assessment through illustrations of current testing projects, and identifies areas in need of further development.

Keywords

automated scoring, language processing technology, computer technology, natural language processing, speech recognition

As a rule, technical advances in modern life first see the day of light in an academic research laboratory; only later are they released to the public by a commercial enterprise. Software used in writing is an important exception. Most word processors, formatters, communications software, spelling checkers, or grammar checkers were developed commercially. (Dobrin, 1990, p. 67)

Dobrin's observation about technical advances in writing software could also be made about language processing software for English language assessment despite academics' recognition of the potential of language processing technologies for improving assessment (e.g., Alderson, 1987; Corbel, 1993; Stansfield, 1986). Because most of the activity in this area occurs in the commercial sector, few academics in language assessment know how language processing technologies work and what they can do. Moreover, little research investigating the validity of inferences and uses of automated scoring of L2 tests

Corresponding author:

Carol A. Chapelle, 339 Ross Hall, Ames, IA 50011-1201, USA
Email: carolc@iastate.edu

appears in journals in applied linguistics. As a consequence, the issues remain unfamiliar and mysterious to many professionals in language assessment who are skeptical of the validity of inferences and uses made from computer generated evaluations of examinees' constructed responses.

This special issue of *Language Testing* takes a step toward changing this state of affairs, but the fact remains that the development, evaluation and use of language processing in language assessment is a cross-disciplinary endeavor. Automated evaluation of human language for assessment relies on basic research in computational linguistics concerned with creating programs capable of recognition of human language (Tennant, 1981), but it also draws upon other areas from second language acquisition to educational measurement. Ideally, these cross-disciplinary contributions yield transdisciplinary knowledge as valuable new learning and assessment opportunities are created for second language learners. In view of the academic richness of this area, it is ironic that the development of such language evaluation systems actually takes place outside of academia, and that academic journals in language assessment are almost void of research on the effectiveness of various computer-assisted response analysis schemes for various assessment purposes.

This paper charts the research issues appearing in the transdisciplinary space at the intersection of language processing technologies and second language assessment. To begin to understand its scope, we define it as including the study of the range of occasions when the second language learner produces language in oral or written form of any length in response to any type of prompt, and a computer program is used to evaluate the linguistic response in order to return analysis-based feedback. We assume that readers understand the necessity of evaluating examinees' actual language production for many test purposes. Learner language can prompt feedback as a marked-up text, a feedback message, or a score; it may be returned immediately or delayed. The feedback may be used for making high-stakes decisions, low-stakes decisions, or for informing students about the correctness of their language. Automated essay scoring (AES) can be used for evaluation of written essays, automated speech recognition (ASR) technologies for analysis of spoken responses, automated scoring systems for short-answer tasks, intelligent response evaluation in computer-assisted language learning programs, and automated evaluation of users' text in word processing programs such as *Microsoft Word*. We use the phrase 'language processing technologies' to refer to all of these technologies, which process learners' language using sufficiently flexible procedures to return an analysis.

The Beginnings

The historical underpinnings for this complex endeavor can be traced from work in three areas. One is research and practice in the automated analysis of English essays. The important and complex issues that define this area today were hardly forecast by the simple aim of Page when he set out to develop an automated essay grader that would lighten the load of the overworked English teacher, who regularly spent enormous amounts of time grading essays. Wresch (1993) retells this story:

In 1965, Page presented a plan to the College Entrance Examination Board (CEEB) to have computers automatically evaluate student writing samples. He had already completed two years of research in which he found that computers could grade an essay as effectively as any human teacher – just by checking for such easily detectable attributes as sentence length, and the presence or absence of particular words. (p. 46)

Wresch goes on to explain that Page's idea was before its time because despite positive findings, the research was 'derided or ignored' (p. 46) by the would-be beneficiaries of Project Essay Grade (PEG). This first pioneering automated essay scoring system was published in 1966, after obtaining funding from the College Board (Page, 2003). The development and evaluation of PEG continued until 1968 (Kukich, 2000), when the project was recessed due to technical limits. The project was then revived in the mid-1980s (Phillips, 2007), making it the most long-lasting automated essay scoring system, but since that time many similar projects have been initiated for analysis of both first and second language writing (Bailin, 1989; Wresch, 1993).

From the 1960s, research in computer-assisted language learning (CALL) had explored a range of approaches to language processing in order to provide informative feedback to learners (Hart, 1981; Pusack, 1983), to construct models of learners' knowledge (e.g., Marty, 1981; Chanier et al., 1992), or to engage in intelligent-appearing dialogue with the learner (Underwood, 1984). Such systems could result in scores reported back to students and teachers, but the primary goal was to produce feedback that would increase learning. Hart (1981) described the authoring tools used by materials developers to produce item types including dictation, fill-in-the-blank, and sentence construction. Such tools conducted an analysis that would return error messages by underlining and providing symbols to indicate errors in word order, morphology, and spelling, for example. A number of projects in academic settings in the USA and Canada set out to develop automated essay analysis of second language learners' essays (Heift and Schulze, 2007), but in terms of the scope of product use today, we believe that most would agree that commercial giant *Microsoft Word* has provided the most used essay analyzer for second language learners.

Little work in computer-assisted language testing was concerned with analysis of learner's responses except for a few projects reported in papers in Stansfield's 1986 edited collection on *Technology in Language Testing*, which include Molholt and Presler's (1986) study on automated speech rating of the Test of Spoken English and Reid's (1986) investigation of the automated essay rating that utilizes NLP technology. The majority of the early research in computer-assisted second language assessment targeted the use of technology for developing computer-adaptive tests with a multiple-choice format for responses (e.g., Dunkel, 1991). Such tests were intended to test more efficiently the same constructs that had been assessed using paper and pencil tests with multiple-choice questions, typically for making decisions about placement or selection. The mandate for such tests was reported as coming from teachers: 'Consultation at BYU [Brigham Young University] with intensive-English directors confirmed that the instrument they needed was an efficient and accurate ESL proficiency test rather than a diagnostic test' (Madsen, 1991, p. 245). Accordingly, results from research on the test

included the number of items required for placement; mean number of items attempted on the computer-adaptive test (CAT); mean amount of time it took students to complete the test; students' affective responses to taking the test on the computer. Such efficiency-oriented projects kept the focus of research on computer-assisted language assessment relatively narrow, not encompassing the constructed response item types or assessments intended to provide feedback to learners about their language. Today, the concept of adaptivity has expanded to include 'assessments that are useful for the benefit of educators who wish to place and evaluate examinees as well as for learners who wish to better understand what they know and what they need to work' and therefore adaptivity no longer needs to be linked to multiple-choice items (Mislevy et al., 2008, p. 22).

These three strands contributed in complementary ways to the study of language processing technologies in second language assessment today. Work on automated essay evaluation revealed the importance of the community's (lack of) acceptance of the technology. Wresch's (1993) retrospective on this area notes the consistently skeptical community that disregarded and rejected research on automated essay scoring regardless of the results. Research in CALL attempting to include NLP technologies demonstrated the difficulty in identifying errors in learner language (Felshin, 1995). The many dollars that were spent in academic research and development projects yielded minimal results relative to what is found in commercial software such as *Microsoft Word*. Research in language assessment proved to be narrow, focusing on adaptive testing to test the same constructs and for the same purposes that multiple-choice tests were being used for at the time. Chapelle and Douglas (2006) pointed out that an efficiency-oriented agenda alone eclipses the potential promise of technology for expanding the uses and usefulness of computer-assisted language assessment. With this background, we can look at some of the major projects employing language processing technologies today.

Approaches to Automated Language Processing

A number of assessments for English as a second language writing and speaking serve as working examples of current language processing capabilities. All such systems incorporate both statistical and linguistic methods to accomplish scoring by identifying relevant linguistic evidence in the examinees' response and then analyzing and combining that evidence in a manner that optimizes prediction of scores obtained from human raters. Various systems approach these two steps differently, but the most important advances in this area have occurred through the combination of sophisticated language recognition and statistical procedures to create expert systems for scoring. An expert system is 'a computer system or program which incorporates one or more techniques of artificial intelligence to perform a family of activities that traditionally would have to be performed by a skilled or knowledgeable human' (Tanimoto, 1987, p. 461).

Increasing sophistication is evident if we contrast Page's invention in the 1960s with the systems currently in use and under development. As shown in Table 1, Page identified surface features of the examinee's written essays such as the number of words in the essay, the variation in word length, and the numbers of various parts of speech (e.g., prepositions). These three categories were intended to serve as evidence of the examinees' fluency, diction, and syntactic complexity, respectively. The values for each of these features were

Table 1. Examples of automated essay scoring with their constructs, observations, statistical approaches and correspondence indices with human raters

System	Target aspects of the writing construct	Example observations	Statistical approaches	Correspondence indices with human raters
PEG	Fluency Diction Syntactic complexity	Essay length Word length Part of speech	Regression	$r = 0.61\text{--}0.87$ (Page, 2003); $r = 0.77\text{--}0.89$ (Keith, 2003)
e-rater [®]	Grammar, usage, mechanics, style Organization and development Lexical complexity Topic-relevance	Article errors, subject-verb agreement errors Discourse markers and identifiers Vocabulary frequency; average word length Topic-specific vocabulary usage	Regression	Agreement 87%–97% (Burstein, 2003; Valenti et al., 2003; Kukich, 2000); Exact agreement 45%–59% ^{**} (Attali & Burnstein, 2006)
IEA	Content Style Mechanics	Similarity to source Coherence Spelling	Latent Semantic Analysis (LSA); Regression	$r = 0.70\text{--}0.85$ (Landauer et al., 2003); agreement 85%–91% (Valenti et al., 2003)
BETSY	Text classification (i.e., the probability that an essay belongs to a good or bad category)	Stemmed words; stopwords (e.g., function words, adjectives, and adverbs)	Bayesian classification	Agreement 80% (Rudner and Liang, 2002)
Intelli-Metric [™]	Cohesiveness and consistency in purpose and main idea Breadth of content and support for concepts advanced Logic of discourse Syntactic complexity and variety Accuracy	500 semantic, syntactic, and discourse features such as • part of speech • vocabulary • sentence structure • concept expression	CogniSearch [™] and Quantum Reasoning ^{™*}	Agreement 96%–98% (avg. $r = 0.833$) (Rudner et al., 2005) Exact agreement 65%–86%; adjacent agreement 99% (Vantage Learning, 2006)

*Systems were reportedly developed using Artificial intelligence (AI) techniques that combine statistical and computational language processing.

**These agreement rates are comparable to those between two human raters, which range between 43% and 59% for a 6-point score scale.

combined using multiple regression analysis in order to maximize prediction of holistic scores provided by human raters (Page, 2003; Valenti et al., 2003), obtaining results that were noted to be as good as those obtained in studies of agreement between two human raters. Based on the results of confirmatory factor analysis of PEG data collected through the Praxis test administered to applicants for teaching certification, Keith (2003) showed that PEG scores had a better correlation (0.92) with essay true scores than three different pairs of human judge scores (0.88, 0.89, and 0.89).

Despite these positive results, PEG scoring relies on surface features of writing and it might therefore be easily fooled if the examinee knows about the grading mechanism of the system. Moreover, as of the 2003 description, the large number (30–40) of predictor variables resulted in high multicollinearity, which conceals the predictive power of individual variables, precluding use of this approach for diagnostic purposes. In part due to these limitations, future projects have attempted more sophisticated methods of feature extraction and optimization as they have become available.

With the aim of incorporating a more expert-based analysis of examinees' essay responses, e-rater[®] draws upon work in natural language processing, 'the analysis and synthesis of natural language by computer' (Cullingford, 1986, p. 394). Developed by Educational Testing Service (ETS), e-rater[®] was operationalized as a second rater to complement human scoring for the Analytic Writing Assessment (AWA) portion of Graduate Management Admissions Test[®] (GMAT[®]) in 1999 and has been used as a second rater to score the Analytic Writing section of the Graduate Record Examination[®] (GRE[®]) and the independent writing task of the Test of English as a Foreign Language[®] (TOEFL[®]) iBT[™], and as the sole rater to provide writing scores and/or feedback for practice tests and products including the TOEFL[®] Practice On-line (TPO), and Criterion[®], among others (Burstein, 2003; Enright and Quinlan, 2010; Kukich, 2000; Catherine Trapani, personal communication, June 23, 2010; <http://www.ets.org/gre/general/scores/how>).

The response analysis begins with the part-of-speech (POS) tagger in the e-rater[®] syntactic parser, assigning POS labels to individual words. These POS-labeled words are then grouped together based on subcategorizations of verbs so as to form phrases and clauses building a syntactic tree. The e-rater[®] discourse analyzer is employed to capture the organization of ideas. The discourse module annotates essays using discourse identifiers such as parallel-relation markers (e.g., *first, second, third*, etc.), discourse indicators (e.g., *however, in addition*, etc.), and syntactic structures (e.g., complimentizer *that*) based on Quirk, Greenbaum, Leech, and Svartik's (1985) discourse classification schema. Topic analysis is operationalized through vocabulary usage in e-rater[®]. The measure of topical relevance in the e-rater[®] topical analysis module is based on the vector-space model, which enables the AES system to compare the content of test essays with those of training samples by computing word frequencies and converting them into weight vectors (Burstein, 2003). Numeric outputs separately produced by the three NLP modules are then stored and combined into features and assigned weights in a multiple regression model to produce the final score of the test essay, a procedure resulting in good agreement with human raters, as shown in Table 1.

Another approach to an expert-based measure of essay quality was developed for the Intelligent Essay Assessor (IEA), which incorporates an approach to information retrieval called Latent Semantic Analysis (LSA). LSA was developed on the basis of the

assumption that 'there is some underlying "latent" semantic structure in word usage data that is partially obscured by the variability of word choice' (Dumais et al., 1988, p. 282). Statistical techniques are used to estimate this latent structure (Deerwester et al., 1990), which can then be used for analysis of a text such as that produced on an essay test. IEA's analysis also draws upon other components to analyze mechanics and style, but according to Landauer et al. (2003), LSA-based content analysis contributes the most to IEA's prediction of essay quality, especially when the system is applied to classroom instruction. The utility of LSA is explained on the basis of the assumption that 'the meaning of a passage is contained in its words, and that all its words contribute to a passage's meaning' (Landauer et al., 2003, p. 88).

Bayesian Essay Test Scoring sYstem (BETSY), as its name indicates, approaches automated essay scoring by drawing upon Bayes' Theorem for classification of essays (Rudner and Liang, 2002). Put generally, 'in Bayesian classification ... we are given some observation and our job is to determine which of a set of classes it belongs to' (Jurafsky and Martin, 2000, p. 148). The specific techniques for classification have been developed through uses in document categorization such as identifying spam emails, and classifying job applicants' resumes. By analogy to these widespread uses, the goal of the BETSY project was 'to determine the most likely classification' of examinees' essays based on a set of optimally selected essay features (Rudner and Liang, 2002, p. 4). Essays were thus 'evaluated as the product of probabilities of the presence or absence' of each calibrated feature in the essay (Rudner et al., 2005, p. 3). With their focus on comparing the relative performance of two common Bayesian models, Rudner and Liang (2002) did not provide much information about the specific features of the essays, referring to them as stemmed words and stopwords. Despite some good agreement rates reported, Dikli (2006) points out that BETSY should be treated as a research tool due to the limited research on the system.

IntelliMetricTM by Vantage Learning is another expert-based automated essay scoring system which was commercially released in 1998 by Vantage Learning (Elliot, 2003). This system is theoretically grounded in cognitive linguistics and developed drawing upon cognitive processing, artificial intelligence (AI), natural language processing (NLP), computational linguistics, and statistical technologies (Elliot, 2003; Vantage Learning, 2006). IntelliMetricTM emulates human essay rating processes, using advanced AI technologies (Elliot, 2003; Dikli, 2006). According to Elliot (2003), the system models the characteristics of essay responses associated with each score using CogniSearchTM and Quantum ReasoningTM and applies this 'intelligence' in subsequent scoring (p. 71). In analyzing essays, IntelliMetricTM examines more than 500 semantic, syntactic, and discourse features, which are classified into five categories: focus and coherence, organization, development and elaboration, sentence structure, and mechanics and conventions. Texts are tagged in terms of parts of speech, vocabulary, sentence structure, and concept expression, and these data are coded for the computation of multiple mathematical models devised and patented by Vantage Learning. Each model then associates extracted features with the scores assigned in the training set. IntelliMetricTM produces a single score by integrating information generated from different models using a proprietary optimization technique (Rudner et al., 2005).

These examples of current work in automated essay scoring demonstrate the clear promise of a variety of expert-based approaches. Ultimately, it would be useful for researchers to explore the comparative utility of the various systems for particular testing purposes, but in the meantime, it seems clear that recent research has found some success. The work is less developed in automatic scoring of speech.

Automated Speech Scoring Systems

To date only a few notable attempts have been made to use speech recognition and processing technologies for automated speech rating of language learners' spoken language. Two examples of automated speech assessments demonstrate the current state of the art in language assessment: Versant Tests (also known as PhonePassTM) developed and commercially operated by Ordinate Corporation and SpeechRaterSM by Educational Testing Service (ETS).

Delivered over the telephone, Versant Tests apply an automated speech rating system to score most of the test tasks. The test was developed and operated by Ordinate Corporation in the late 1990s, when the test was named PhonePassTM. According to Bernstein (1999), the construct of PhonePassTM was 'facility in spoken English', which the author defined as 'the ability to understand spoken language and respond intelligibly at a conversation pace on everyday topics' (p. 2). As implied in this construct definition, the goal was to assess general listening and intelligibility of speaking in every day settings. Test tasks include sentence read-aloud, sentence repetition, saying opposite words, short responses to questions, and open-ended free responses to questions; only the latter is not scored. The overall score is computed by averaging five weighted subscores in listening vocabulary, repeat accuracy, reciting/pronunciation, reading fluency, and repeat fluency. The scores of the last two subareas are measures of suprasegmental features (e.g., timing, pause, rhythm, and concatenation of words) of speaking facility; however, the specifics of the scoring procedures are only briefly described in Bernstein (1999) and Townshend and Todic (1999).

The strengths of PhonePassTM include the efficiency of test administration, high reliability, and high predictability of examinees' performance in an ordinary setting. The test only lasts for approximately 10 minutes and yet results in high reliability (0.94 for PhonePassTM) and correlations with other norm-referenced speaking assessments ($r = 0.88$) such as Test of Spoken EnglishTM (Townshend and Todic, 1999). Nevertheless, most language teachers would point out that the tasks on PhonePassTM are limited in estimating examinees' speaking abilities as it does not take into account the multiple components of communicative competence (Xi et al., 2008). Bernstein (1999) for example points out that the test does not include 'social skills, higher cognitive functions, or world knowledge', nor high-level linguistic functions and features such as persuasiveness, coherence in discourse, understanding of culturally bound nuances (p. 4), most of which would be considered part of communicative competence.

SpeechRaterSM v1.0, an automated speech rating system developed by researchers at Educational Testing Service, is used in the Test of English as a Foreign LanguageTM (TOEFL[®]) Internet-based test (iBT) Speaking Practice Test. This project was driven by the increasing need for immediate response scoring and feedback generation for potential examinees in the online TOEFL[®] Speaking Practice Test (TPO) along with the launch of the TOEFL[®] iBT test that has incorporated speaking as a new component of the test.

The goal of SpeechRaterSM is to allow examinees to check their readiness to take the actual TOEFL[®] iBT. The SpeechRaterSM scoring method is intended to assess aspects of examinees' responses that provide evidence for the construct of communicative speaking in academic environments by following the scoring guidelines of actual TOEFL[®] iBT to produce scores comparable to those from the actual test (Xi et al., 2008).

The scoring process relies on a program that extracts potentially meaningful features of speech input to build a multiple regression scoring model (Xi et al., 2008). These 29 candidate features are trimmed to make the regression model technically sound and represent the construct of communicative, academic speaking ability to as full an extent as possible in the areas of fluency, vocabulary, grammar, and pronunciation. Examples of final candidate features include length of silences per word, mean of silence duration, speaking rate in words per second, unique words divided by duration of entire transcribed segment, average chunk length in words, and language model score. The correlation coefficients between the SpeechRaterSM model predicted scores and human scores are at a moderate range around 0.60–0.70, while the exact and adjacent agreement rates between human raters and SpeechRaterSM range between 95% and 99%. These correlations are one part of a more extensive validity argument that was developed to support the use of automated scores in online TOEFL[®] Speaking Practice Tests.

These examples help to illustrate what has been accomplished so far in technology-assisted language assessment and what can be further accomplished in terms of the design of automated assessment models and the adoption of technologies. They also support Dobrin's (1990) observation that most of the development has occurred in commercial products. In this setting, where the necessary resources have been assembled, the research tends to target optimizing results and supporting the intended uses. However, language processing technologies in assessment are of some consequence to test score meaning, test score use, and possibly to learning, which are central to the academic concerns in language testing.

Validity and Language Processing Technologies

In evaluating automated scoring of essays on high-stakes ESL tests, writing experts typically focus on the need to demonstrate that the scores awarded to examinees by the computer correlate with those that human raters would provide. Similarly, teachers discussing software tools such as *Microsoft Word* raise concerns about the accuracy and appropriateness of the feedback that the system provides to learners. In both of these cases, the concerns may provide a good starting point for considering the validity of the use of the assessment, but both cases also illustrate the need to consider the validity of assessment use more broadly. Referring to cases such as the first one, Clauser, Kane and Swanson (2002) point out the following: 'the validity argument for scores produced by automated systems must go beyond demonstration of the correspondence between these scores and ratings ... Closer consideration of this type of validity evidence suggests that this is not always necessary and is generally not sufficient' (p. 420).

They go on to explain that expert judgment is problematic as a sole criterion for assessing the validity of machine-scored responses because experts differ in their judgments. Moreover, different experts apply criteria differently, and the same experts may apply criteria differently from one time to the next. The fact that interrater and

intrarater reliability estimates from human raters are not perfect demonstrates the need to look beyond such judgments in building a validity argument concerning test score use. This has been pointed out elsewhere by Bennett and Bejar (1998), and in language testing by Chapelle (2003), and Alderson, Percsich, and Szabo (2000).

Clauser, Kane, and Swanson (2002) also outline other factors to consider in developing a validity argument for an assessment with automated response scoring. They focus on how development and implementation of the scoring model are tied to the types of inferences that one might want to draw based on test scores. The inferences are those presented by Kane, Crooks, and Cohen (1999) and Kane (2006), which are illustrated in language testing by Chapelle, Enright, and Jamieson (2008). Within the many inferences that underlie score interpretation and use, Clauser, Kane, and Swanson identify many areas of investigation which may strengthen the validity of inferences. For example, evaluation inferences might be supported by research demonstrating the consistency by which scoring rules are applied to test responses, and generalization inferences would need to be supported by demonstrating generalization over tasks, raters and conditions.

Explanation inferences can be supported by demonstrating the correspondence of the scoring model with the construct that is intended to explain performance. As Chapelle (2003) points out, however, the efficiency-oriented research that is the bread-and-butter of commercial organizations is not aimed at better understanding the detailed construct definitions that serve as explanation for language tests. Instead, a more scientifically oriented program is needed to investigate construct representation (Embretson, 1983) in second language testing. An important first step toward this goal is to undertake validation of test interpretations and uses that extend beyond correlations between scores obtained by human raters and automated procedures. This broader approach to validation is evident in the validity argument for the use of SpeechRaterSM (Xi, 2008; Xi et al., 2008).

This work on SpeechRaterSM demonstrates the significant role played in this process by the purpose of the assessment, and therefore provides a way of looking at the validity of uses of automated assessment other than those in high-stakes assessment. A look at the validity argument in the context of classroom assessment is useful for considering the validity of the use of automated analysis and feedback from *Microsoft Word* by writing teachers as a means of providing automated feedback for learners. Table 2 outlines a simple interpretative argument one might wish to make in support of the use of feedback from *Microsoft Word* in an ESL writing class. The conclusion or claim would be that students are able to use such grammar feedback after the class ends and students are writing in English on their own. The interpretive argument shown in Table 2 would form the basis of research, which would be needed to support a validity argument for the use of such writing assessment. The example assumptions should be recognizable as statements that could form the basis for empirical research.

An interpretive argument for the use of such an assessment tool helps to illustrate an expanded potential for assessments in instruction. Such an argument, which has a different type of conclusion than what would appear in an interpretive argument for a high-stakes test, helps to illustrate how language processing technologies might help to expand the uses of language assessments. Current research makes such an expansion to low-stakes classroom applications seem realistic to achieve the goal of providing diagnosis and specific feedback to help students learn. Cotos and Pendar's (2008) analysis is particularly optimistic: 'Obtaining detailed profiles of learner written

Table 2. An example of an interpretive argument and some example assumptions underlying the use of automated feedback in a university ESL writing class

Inferences	Warrants	Example assumptions
Utilization	Students will develop consciousness of their grammatical performance and use electronic tools to focus on the correctness of specific grammatical points in their writing.	Students will find repeated use of helpful grammatical feedback from <i>Microsoft Word</i> useful.
Extrapolation	Mark-up on students' essays in the ESL course will be the same as mark-up they will receive on their other writing.	<i>Microsoft Word's</i> language processing procedures produce consistent results across types and topics of ESL writing outside ESL classes.
Generalization	Mark-up on a student's essay is similar to what would appear on other essays with the same errors.	<i>Microsoft Word's</i> language processing procedures produce consistent results across academic essay types and topics.
Evaluation	Mark-up on a student essay accurately and consistently identifies areas worthy of the student's reconsideration.	<i>Microsoft Word's</i> language recognition procedures are sensitive to the types of errors ESL writers make.

performance across various components of the construct for diagnosing writing ability appears to be possible if NLP-based automated scoring is employed by [computer assisted language testing]' (p. 69).

Conclusion

Most research has been reported on the development of automated scoring of essay responses; however, if we consider the range of assessment purposes, it is evident that these technologies hold the potential for improving language assessment in many different ways. The scope of the issues should be apparent in view of the approaches that can be taken to language processing for language assessments and the need for appropriate validity research to support various assessment uses.

1. What approaches to automatic language processing are needed to provide adequate assessment results for various purposes? A variety of automated essay scoring (AES) systems have been developed and employed in actual language assessment since 1966 due to the increasing need for efficient, as well as reliable and valid, massive rating of essays from high-stakes tests. As an attempt to support the validity of these AES systems, human-rated scores were usually used as referential norms in building the scoring model (as a training set of the model) of an AES system and were compared with machine-generated scores based on correlation coefficients. However, language processing technologies used in other types of tests, particularly to extend beyond what human raters can do, will require an expanded set of approaches for validation.
2. How does the language processing technology affect the research that goes into the validity argument? As Clauser, Kane, and Swanson (2002) pointed out, the automated language processing employed in response scoring is relevant to every

type of inference that is made in an interpretive argument, and therefore developing a validity argument requires multiple types of research intended to provide evidence supporting the assumptions about the appropriateness of the methods.

3. How can validation research help to inform future applied work in language processing? The type of interpretive argument outlined above hints at the way in which validation research might be designed in a way that can inform practice. Moreover, when an interpretive argument includes an explanation inference, as one would expect to find in an interpretive argument for high-stakes testing, research needs to demonstrate evidence that the scoring procedures are consistent with the construct intended to form the basis for that inference. Such research ties together work attempting to develop learner models, theories of second language acquisition, language assessment, and computational linguistics all for the purpose of improving the validity argument for a particular test use.
4. Are the language technologies and assessment practices used for English as a second language available for languages other than English? The Versant tests are commercially available in Spanish and English, and a variety of test sets are available to suit examinees' different purposes for language learning or test users' testing purposes. Moreover, *Microsoft Word* provides help with language correctness and spelling for many different languages, thus providing tools that offer the type of immediate feedback on learners' writing that has been the goal of research in CALL for many years.

These broad questions have begun to be addressed through the work in this special issue of *Language Testing* and will undoubtedly be taken up in both commercial and academic research projects in the future. In view of the complexity of the issues in language processing technologies and second language assessment real progress will undoubtedly require cross-disciplinary teams, as Zock (1996) pointed out, to construct systems, conduct research and provide feedback to future research and development.

References

- Alderson JC (1987). *Innovation in language testing: Can the microcomputer help? Language Testing Update Special Report*, No. 1. Lancaster: Centre for Research in Language Education, Lancaster University.
- Alderson JC, Percsich R, and Szabo G (2000). Sequencing as an item type. *Language Testing*, 17(4), 423–447.
- Attali Y, Burnstein J (2006). Automated essay scoring with e-rater® V.2.0. *Journal of Technology, Learning, and Assessment*, 4(3). Retrieved on June 22, 2010 from <http://www.jtla.org>.
- Bailin A (ed.) (1989). *Computers and the humanities (Special Issue on Intelligent Computer Assisted Language Instruction)*, 23, 59–70.
- Bennett RE (1993). On the meanings of constructed responses. In RE Bennett and WC Ward (eds.), *Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment* (pp. 1–27). Hillsdale, NJ: Lawrence Erlbaum.
- Bennett RE, Bejar II (1998). Validity and automated scoring: It's not only the scoring. *Educational Measurement: Issues and Practice*, 17, 9–17.
- Bernstein J (1999). *PhonePass™ testing: Structure and construct*. Menlo Park, CA: Ordinate.

- Burstein J (2003). The *e-rater*[®] scoring engine: Automated essay scoring with natural language processing. In MD Shermis, and JC Burstein (eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 113–121). Mahwah, NJ: Lawrence Erlbaum.
- Chanier T, Pengelly M, Twidale M, and Self J (1992). Conceptual modelling in error analysis in computer-assisted language learning systems. In ML Swartz, and M Yazdani (eds.), *Intelligent tutoring systems for foreign language learning* (pp. 125–150). Berlin: Springer-Verlag.
- Chapelle CA (2003). *English language learning and technology: Lectures on applied linguistics in the age of information and communication technology*. Amsterdam: John Benjamins.
- Chapelle CA, Douglas D (2006). *Assessing language through computer technology*. Cambridge: Cambridge University Press.
- Chapelle CA, Enright MK, and Jamieson JM (2008). *Building a validity argument for the Test of English as a Foreign Language*[™]. New York/London: Routledge.
- Clauser BE, Kane MT, and Swanson DB (2002). Validity issues for performance-based tests scored with computer-automated scoring systems. *Applied Measurement in Education*, 15(4), 413–432.
- Corbel C (1993). *Computer-enhanced language assessment*. Sydney, Australia: National Centre for English Language Teaching and Research, Macquarie University.
- Cotos E, Pendar N (2008). Automated diagnostic writing test: Why? How? In CA Chapelle, Y-R Chung, and J Xu (eds.), *Towards adaptive CALL: Natural language processing for diagnostic language assessment* (pp. 65–81). Ames, IA: Iowa State University.
- Cullingford RE (1986). *Natural language processing: A knowledge-engineering approach*. Totowa, NJ: Rowman & Littlefield.
- Deerwester S, Dumais S, Landauer T, Furnas G, and Harshman R (1990). Indexing by latent semantic analysis. *Journal of the American Society of Information Science*, 41(6), 391–407.
- Dikli S (2006). An overview of automated scoring of essays. *The Journal of Technology, Learning, and Assessment*, 5(1), 3–35.
- Dobrin DN (1990). A limitation on the use of computers in composition. In DH. Holdstein and CL Selfe (eds.), *Computers and writing: Theory, research, practice* (pp. 40–57). New York: Modern Language Association of America.
- Dumais ST, Furnas GW, Landauer TK, Deerwester S, and Harshman R (1988). Using Latent Semantic Analysis to improve access to textual information. In *Human factors in computing systems, CHI'88 conference proceedings (Washington, D.C.)*, May (pp. 281–285). New York: ACM.
- Dunkel A (1991). Research on the effectiveness of computer-assisted instruction and computer-assisted language learning. In P Dunkel (ed.), *Computer-assisted language learning and testing* (pp. 5–36). New York: Newbury House.
- Elliot S (2003). IntelliMetric[™]: From here to validity. In MD Shermis and JC Burstein (eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 71–86). Mahwah, NJ: Lawrence Erlbaum.
- Embretson SE (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin*, 93, 179–197.
- Enright M, Quianlan T (2010). Complementing human judgment of essays written by English language learners with e-rater[®] scoring *Language Testing*, 27(3): 317–334.
- Felshin S (1995). The Athena language learning project NLP system: A multilingual system for conversation-based language learning. In V. M. Holland, J. Kaplan, and M. Sams (eds.), *Intelligent language tutors: Theory shaping technology* (pp. 257–272). Hillsdale, NJ: Lawrence Erlbaum.

- Hart RS (1981). Language study and the PLATO system. *Studies in Language Learning*, 3, 1–24.
- Heift T, Schulze M (2007). *Errors and intelligence in computer-assisted language learning: Parsers and pedagogues*. New York/London: Routledge.
- Jurafsky D, Martin JH (2000). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. Upper Saddle River, NJ: Prentice Hall.
- Kane MT (2006). Validation. In RL Brennan (ed.), *Educational measurement* (4th ed., pp. 18–64). Westport, CT: American Council on Education/Praeger.
- Kane MT, Crooks T, and Cohen A (1999). Validating measures of performance. *Educational Measurement: Issues and Practice*, 18(2), 5–17.
- Keith TZ (2003). Validity of automated essay scoring systems. In MD Shermis, and JC Burstein (eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 147–207). Mahwah, NJ: Lawrence Erlbaum.
- Kukich K (2000). Beyond automated essay scoring. *IEEE Intelligent Systems*, 15(5), 22–27.
- Landauer TK, Laham D, and Foltz PW (2002). The Intelligent Essay Assessor. *IEEE Intelligent Systems*, 15(5), 27–31.
- Landauer TK, Laham D, and Foltz PW (2003). Automated scoring and annotation of essays with the Intelligent Essay AssessorTM. In MD Shermis and JC Burstein (eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 87–112). Mahwah, NJ: Lawrence Erlbaum.
- Madsen HS (1991). Computer adaptive testing of listening and reading comprehension: The Brigham Young University approach. In P Dunkel (ed.), *Computer assisted language learning and testing: Research issues and practice* (pp. 237–257). New York: Newbury House.
- Marty F (1981). Reflections on the use of computers in second-language acquisition. *Studies in Language Learning*, 3, 25–53.
- Mislevy R, Chapelle CA, Chung YR, and Xu J (2008). Options for adaptivity in computer-assisted language learning and assessment. In CA Chapelle, Y-R Chung, and J Xu (eds.), *Towards adaptive CALL: Natural language processing for diagnostic language assessment* (pp. 9–24). Ames, IA: Iowa State University.
- Molholt G, Presler A (1986). Correlation between human and machine ratings of Test of Spoken English reading passages. In C Stansfield (ed.), *Technology and language testing* (pp. 111–128). Washington, DC: TESOL Publications.
- Page EB (2003). Project Essay Grade: PEG. In MD Shermis, and JC Burstein (eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 43–54). Mahwah, NJ: Lawrence Erlbaum.
- Phillips SM (2007). Automated essay scoring: A literature review. Available from www.sace.ca/pdfs/036.pdf.
- Pusack JP (1983). Answer-processing and error correction in foreign language CAI. *System*, 11(1), 53–64.
- Quirk R, Greenbaum S, Leech S, and Svartik J (1985). *A comprehensive grammar of the English language*. New York: Longman.
- Reid J (1986). Using the Writer's Workbench in composition teaching and testing. In CW Stansfield (ed.), *Technology and language testing* (pp. 167–186). Washington, DC: TESOL Publications.
- Rudnerz L, Garcia V, and Welch C (2005). *An evaluation of IntellimetricTM essay scoring system using responses to GMAT[®] AWA prompts* (GMAC research report number RR-05-08). Retrieved December 27, 2008, from www.vantagelearning.com/docs/intellimetric/IM_ReseachSummary_IntelliMetric_Accuracy_Across_Genre_and_Grade_Levels.pdf.

- Rudner LM, Liang T (2002). Automated essay scoring using Bayes' Theorem. *The Journal of Technology, Learning, and Assessment*, 1(2), 3–21.
- Stansfield C (ed.) (1986). *Technology and language testing*. Washington, DC: TESOL Publications.
- Tanimoto SL (1987). *The elements of artificial intelligence*. Rockville, MD: Computer Science Press.
- Tennant H (1981). *Natural language processing: An introduction to an emerging technology*. New York: Petrocelli Books.
- Townshend B, Todic O (1999). *Comparison of PhonePassTM Testing with the Educational Testing Service Test[®] of Spoken EnglishTM (TSE[®])*. Menlo Park, CA: Ordinate.
- Underwood J (1984). *Linguistics, computers, and the language teacher*. Rowley, MA: Newbury House.
- Valenti S, Nitko A, and Cucchiarelli A. (2003). An overview of current research on automated essay grading. *Journal of Information Technology Education*, 2, 319–329.
- Vantage Learning (2006). *Research summary: IntelliMetricTM scoring accuracy across genres and grade levels*. Retrieved December 27, 2008 from www.vantagelearning.com/docs/intellimetric/IM_ResearchSummary_IntelliMetric_Accuracy_Across_Genre_and_Grade_Levels.pdf.
- Wresch W (1993). The imminence of grading essays by computer – 25 years later. *Computers and Composition*, 10(2), 45–58.
- Xi X (2008). What and how much evidence do we need? Critical considerations in validating an automated scoring system. In CA Chapelle, YR Chung, and J Xu (eds.), *Towards adaptive CALL: Natural language processing for diagnostic language assessment* (pp. 102–114). Ames, IA: Iowa State University.
- Xi X, Higgins D, Zechner K, and Williamson DM (2008). *Automated scoring of spontaneous speech using SpeechRater v1.0* (ETS Research Rep. No. RR-08-62). Princeton, NJ: ETS.
- Zock M (1996). Computational Linguistics and its use in real world: The case of computer assisted language learning. In *Proceedings of the 16th International Conference on Computational Linguistics (COLING)* (pp. 1002–1004).

Copyright of Language Testing is the property of Sage Publications, Ltd. and its content may not be copied or emailed to multiple sites or posted to a listserv without the copyright holder's express written permission. However, users may print, download, or email articles for individual use.