JLTA Newsletter No. 30
日本言語テスト学会
The Japan Language Testing Association

JLTA Newsletter No. 30　　発行代表者:浪田克之介　　2010 年 (平成 22 年) 12 月 15 日発行
発行所:日本言語テスト学会 (JLTA) 事務局
〒389-0813 長野県千曲市若宮 758 TEL 026-275-1964 FAX 026-275-1970
e-mail: youichi@avis.ne.jp URL: http://jlta.ac

＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊

JLTA 2010 annual conference

Yo In'nami
(*Toyohashi University of Technology*)

I would like to take this opportunity to thank you all for joining the JLTA annual conference held at Toyohashi University of Technology. The keynote speech by John Read started by critically reviewing Coxhead's (2000) Academic Word List. Although this well-established list provides guidelines for constructing a vocabulary test, Read argued that adopting a wider approach considering collocations and formulaic expressions is needed to complement the Academic Word List. The symposium by Yasuyo Sawaki, Hideki Sakai, Rie Koizumi, Tomoko Ishii, and John Read reflected the conference theme, "diagnostic testing in language teaching"; this is a flourishing area in language testing and learning, since the need to link testing with learning has increased over the years. The conference was overall very well received, as shown by the 19 presentations (12 of which were in English, while 8 were delivered by overseas scholars from as far as Finland) and a relatively large number of attendees (86).

Nevertheless, there is room for improvement. First, although five of the presentations ran simultaneously owing to time constraints, the number should have been restricted to just three or four at most. Many participants complained that they had difficulty selecting from the high quality of presentations delivered simultaneously. Second, although we posted information about our annual conference on several mailing lists, we were able to attract only two testing companies for commercial exhibition. Given the important roles tests play in society as well as in tertiary education, we need to consider how to better appeal to testing publishers and companies in order to obtain better knowledge about the latest test products and to enable publishers and companies to learn from the cutting-edge research in the field.

第 14 回 日本言語テスト学会
全国研究大会　研究発表
於：北海学園大学
平成 22 年 9 月 11 日（土）

研究発表:

Using word frequency lists to investigate the vocabulary used in a pilot version of a new university entrance exam

Jamie Dunlea
(*Society for Testing English Proficiency*)

The purpose of this study was to investigate the validity of a pilot version of TEAP (Test of English for Academic Purposes) in comparison with the reading and vocabulary sections (from Grade Pre-2 to Grade 1) of EIKEN by analyzing the vocabulary used in the texts of these tests. TEAP, developed by STEP and Sophia University, is meant to measure the English ability of high school students who are planning to study at a university level. The analysis tools for this research were the following two versions of Range by P. Nation, the GSL (General Service List) / AWL (Academic Word List) version and the BNC version. The BNC version was developed in order to analyze the words beyond the first 3000 word level because the GSL/AWL version does not classify these words.

The materials analyzed were 1. Pilot version of TEAP reading section and 2. Vocabulary and reading sections of EIKEN tests (from Grade Pre-2 to Grade 1), and both 1. & 2. were administered in 2009.
The main findings of this research are as follows; 1) Vocabulary profiles of the reading sections of the EIKEN tests analyzed showed clear differences across levels, with higher percentage of more frequent words in the lower level tests. 2) Vocabulary profiles for the TEAP were generally closest to the Pre-1 test of EIKEN, but fell between Pre-1 and Grade 2 for some levels. 3) The number of words needed to cover 95% of running words in the tests increased in a consistent way across EIKEN levels. 4) The 4000 word family level may be a reasonable guideline to comprehend the pilot TEAP, based on the 95% benchmark.

Some future implications, such as investigation of the relationship between vocabulary, comprehension and test item difficulty, are also discussed in the end.

Reported by Yukie Koyama
(*Nagoya Institute of Technology*)

Modality and context effects in estimating the lexical knowledge of the tertiary-level Japanese learners of English with the yes/no test format

Toshihiko Shiotsu
(*Kurume University*)
John Read
(*University of Auckland*)

In this presentation, Dr. Shiotsu and Dr. Read reported on the extended version of the Yes/No test. Their research was very innovative in three ways. First, it presents the target words both in spoken and written form. Secondly, it explores how the addition of two types of sentence-based context influences performance on the Yes/No task. Thirdly, it investigates whether reaction time adds a significant dimension to the measurement of vocabulary knowledge with this test format. They developed two forms of an English Yes/No test, based on samples of items from the BNC word frequency lists. Thus, twelve Yes/No test versions varying in form (A vs. B), mode (oral vs. written) and context condition (none vs. syntactic vs. semantic) were produced.

With the sample of 346 Japanese university students, the Yes/No test proved to be a reliable measure in both the oral and written modes, and the oral and written versions of the same test tended to have similar difficulties. In terms of the context effect, the results indicated that providing semantic context might make the same Yes/No test more difficult, especially in the oral mode. The reaction time data showed that the participants were consistently

slowest in responding to the syntactic contexts, whereas the latencies for the no-context and semantic conditions were quite similar, and did not account for reading and listening comprehension test performances in any useful way.

It was pointed out that it is necessary to establish the criteria on which a non-word can be made effectively because it might affect reaction time of learners.

Reported by Katsumasa Shimada
(*Momoyama Gakuin University*)

Investigating the construct of lexico-grammatical knowledge in an academic ESL writing test

Yasuhiro Imao
(*University of California, Los Angeles*)

Assessing grammatical ability in writing skills is a challenging task, in that a number of components seem to be interwoven in the construct of grammatical ability. Based on the definitions of lexico-grammatical knowledge by Pupura (2004) and Rimmer (2006), this paper examines the relationship between the construct of lexico-grammatical knowledge and human rating data patterns. 200 ESL writing samples were assessed by ESL-trained teachers using three assessment criteria (accuracy, appropriateness, and range). The Confirmatory factor analysis of human rating data patterns showed that the accuracy and range factors were moderately correlated, suggesting that these two factors are statistically within the same construct. However, appropriateness was not correlated with the two, suggesting that appropriateness is not included in the same construct of accuracy and range. Further sophisticated statistical analyses show that the three assessment criteria seemed to be differentiated. This paper gives an insight into investigation of the construct of lexico-grammatical ability and issues of human raters in productive skills. This study also clearly showed a possibility of exploring a theoretical framework (i.e. construct) of lexico-grammatical ability. After the presentation, a few important questions from the floor facilitated a deeper and lively discussion.

Reported by Tomoyasu Akiyama
(*Bunkyo University*)

Diagnosing reading and writing in a second or foreign language

Ari Huhta
(*University of Jyväskylä*)
J. Charles Alderson
(*Lancaster University*)

The past decade has witnessed a growing interest in diagnostic language testing in the field of language test development and research. In this presentation, the authors claim that despite a number of attempts to develop such tests, they have rarely been based on a theory of language learning or a theory of diagnosis. The present paper reported on an international 4-year (2010-2013) research project into the diagnosis of reading and writing abilities in L2. The project brings together scholars from various relevant fields, including applied linguistics, psychology and assessment. The purpose of the project (DIALUKI – diagnosing reading and writing in a second or foreign language) is to identify the cognitive and linguistic features which predict a learner's strengths and weaknesses in those areas. The paper described the three main studies presenting a number of instruments that had been developed for the learners leaning English as a foreign language and Finnish as a second language in Finland. Besides the theories that underlie a range of diagnostic testing methods, a number of useful and interesting tools were presented and demonstrated, including a range of cognitive and psycholinguistic measures in informants' L1 as well as diagnostic tools for detecting L1 dyslexia, in order to examine their applicability of L2 diagnosis. The presentation continued for 40 full minutes and ran out time for Q & A session. Though the presentation was quite

informative and insightful, one of the things which was lacking involved the diagnostic feedback. That is, it seems remaining to be explored what feedback should be given to the student in what way, so it may help him or her to promote his or her learning processes.

Reported by Yoshinori Watanabe
(*Sophia University*)

Diagnosing the English speaking ability of college students in China--Development and validation of the College English Diagnostic Speaking Test

Zhongbao Zhao
(*Shanghai Jiao Tong University*)

This presentation was canceled.

A comparison of four types of spelling tests among Japanese EFL learners: Focusing on sound-letter correspondences

Sachiyo Takanami
(*Graduate School, University of Tsukuba*)

The purpose of this study is to compare and to determine the difficulties of four types of the written form spelling tests among Japanese learners of English. Relationships between English spellings and sounds (i.e., pronunciations) are quite difficult for learners of English to acquire due to their complicated orthographic system. Yet spelling and reading are important aspects of literacy that cannot be disregarded in language competence. Nortyby (1936) compared five forms of spelling tests−namely, story form (i.e., fill in the appropriate words in a given passage), timed dictation (i.e., write whole sentences in a limited time), list form (i.e., write the words as pronounced), multiple-choice (choose the correct spelling from among several choices), and oral form (spell the word orally)−for diagnostic purposes and investigated the variability of student performance and found that the multiple-choice form was the easiest while the timed dictation form was the most difficult. Focusing on the written form, Moore (1937) determined that the multiple-choice form was useful for measuring learners' spelling ability. Yet it remains unclear which type of spelling test is the most appropriate for EFL learners. Twenty words were used in both Northby's and Moore's study (i.e., ninety, succeeded, similar, nickel, admission, carnival, appearance, excitement, planned, finally, orchestra, occasionally, convenience, exhibition, schedule, arriving, various, humor, Saturday, and island).

The current study compares the four types of written spelling tests (i.e., story form, timed dictation, list form, and multiple-choice) among 80 mid-level high school Japanese learners of English. Participants completed a pre-test of approximately 90 regular words (i.e., with sound-letter correspondences; Mori, 2007) to determine their knowledge of English spelling rules. Participants subsequently completed the four types of spelling tests (randomly administered); answers were carefully analyzed using for example Cook's (1997) categorization for L1 and L2 spelling errors (i.e., insertion, omission, substitution, transposition, grapheme substitution, and other) to classify participants' misspellings. The materials which were used in this study were based on the previous findings (Northby, 1936; Moore, 1937). However, to determine the levels of the target words, Gakken's corpus data and JACET 8000's vocabulary lists were considered as a standard. The spelling tests are effective tools to examine learners' level of proficiency in writing mechanics. Thus, the results of this research will be beneficial and helpful for Japanese teachers of English to decide which form to use in his or her English lessons.

Learner adapted testing: An individualistic approach to language assessment

Parviz Alavinia
(*Urmia University*)

This presentation was canceled.

The National English Ability Test of Korea

Oryang Kwon
(*Seoul National University*)

For several years, The Republic of Korea has been in the vanguard among Asian nations coordinating expertise in the field of language testing and foreign language education with government initiatives in order to enhance the English language abilities of its human resources in an age of globalization.

In 2007, a project was initiated to develop a comprehensive testing program of all four English language skills – Speaking, Listening, Reading, and Writing. Its objectives are two-fold: 1) to promote communicative competence in English as an objective for teaching and learning in its high schools and universities, and 2) to better inform stakeholders in the contexts of Korea's educational, governmental and civilian institutions than the variety of instruments that are presently being used.

The National English Ability Test, or NEAT, is to have three different versions, each with content corresponding to the target language use domains for three distinct groups of stakeholders and contexts. Level 1 will be designed to measure the English language proficiency of its university students and graduates. Level 2 is to evaluate both proficiency and achievement of high school students in grades 11 and 12. Level 3 will evaluate both proficiency and achievement of high school students in grade 10. The test content for levels 2 & 3 will be tied closely to the high school curriculum for English language ability in all four skills proscribed by the Ministry of Education; hence the inclusion of 'achievement' within the construct.

Dr. Kwon presented a detailed overview of test development for Levels 2 and 3 of the NEAT, with a target date of 2013 for general use. Piloting sessions of prototypes for the NEAT began in 2009, and by the end of 2011 a grand total of 180,000 students will have participated in piloting at various stages of development.

The NEAT Levels 2 & 3 are to be used for gate-keeping by universities, with scores on the Level 3 exam pertaining to those departments where students will need a satisfactory amount of practical English language skills to study. Departments that require not only practical English language skills but also more advanced ability in Academic English will use scores on Level 2. The distinctions correspond roughly to those of abilities in BICS and CALP (Cummins, 1980). In 2012, the NEAT will be administered at schools nation-wide and validation procedures on the outcome will be conducted to decide on the efficacy of eventually replacing the CSAT presently being used by universities for this purpose.

The NEAT Level 2&3 team is addressing the great number of practical and theoretical issues that a project of this magnitude entails. Because development of the NEAT is still in the midst of on-going validation procedure of hypothesis testing, the content of Dr. Kwon's presentation focused to those areas for which concrete measures to remedy perceived challenges could be reported.

Concerning practical challenges, there is the matter of managing the material and human resources for writing and rating the various items/tasks across administrations, and administering the exam itself, while safeguarding both security and consistency.

An estimated 600,000 high-school students are projected take the NEAT Levels 2 & 3 via the Internet within a certain time period during the school year. With a projected capacity to administer the exam to 50,000 candidates for a single 145-minute sitting, it would take 12 administrations within 3 days to accomplish. If a decision were made to implement 2

sittings rather than one, the result would be a two-fold increase in needed resources.

The need for equating scores across the various forms used for administrations over time presents another challenge; the use of IRT and anchor items to equate the different forms over time would be ideal, but the practical limitations are formidable.
In the area of quality control and security for item writing and rating, the plan is to create a pool of qualified secondary school teachers via a 60-hour regimen of training and on-going assessment for certification. Eighty teachers were trained and certified in 2010. The goal of the program is to have obtained a pool of trained, certified 960 teachers by the year 2012.

Several issues involving test content for Levels 2 & 3 and their respective theoretical constructs remain areas for deliberation. On one hand, the methods for assessment are to complement the high school curricula. On the other hand, this match between testing and teaching on a nation-wide scale is intended to result in an increased level of communicative competence in English that can be verified by actual performance at universities and eventually, in society at large.

To match testing and teaching at high-schools, how should Levels 2 & 3 be defined in terms of operational objectives to be taught and assessed? Dr. Kwon presented a theoretical model based on BICS & CALP and showed how it could be incorporated in components both in the high-school curricula and in the test blueprint of test items and tasks for the NEAT.

Governments and educators throughout Asia share many of the same kinds of problems in their endeavors to improve the ability of its people in the use of English Language and achieving a positive and meaningful washback from reforms in the face of adverse social and economic paradigms. However, the comprehensive and careful approach as presented by Dr. Kwon leads me to believe that the initiative to develop the NEAT in Korea has a very good chance of success, and I look forward to reports on progress and results in the future.

Reported by Jeff Hubbell
(*Hosei University*)

Validation of the listening comprehension component of the Centre Test in Japan: Listening in the real world, in the Course of Study, and in the Centre Test

Kozo Yanagawa
(*Graduate School, University of Bedfordshire*)

The presentation aimed to identify important differences between the presenter's definition of real life listening, listening in the Courses of Study curriculum guidelines, and the listening component of the Center Test. The definition of real life listening was derived from a comprehensive framework of contextual parameters developed by the presenter working within Weir's socio-cognitive framework of test validation.

Yanagawa described 10 important differences between real life listening and listening as taught in Japanese high schools. Results from a survey of teachers and students concerning which of these elements should be better reflected in the Center Test showed strongest support among the teachers for increasing the varieties of English (American, British, etc.) and better reflecting adjustments to pronunciation in connected speech (Sandhi variation).

Yanagawa's concern for improving listening instruction in Japanese high school EFL classes was clearly evident. He stressed that one of the aims of introducing a listening component to the Center Test was positive washback but suggested this potential had not been realized. The presentation was followed by a lively and informative discussion with the audience. It is worth noting that the study concentrated on two of the five kinds of validity described in Weir's framework: Cognitive

and Contextual Validity. The concern with positive washback would suggest that any appeal to change in the Center Test would benefit from also explicitly addressing Consequential Validity. Indeed, any attempt at an integrated, evaluative validation of such a high stakes test would benefit from a balanced consideration of all of the elements of validity in the socio-cognitive framework proposed by Weir, including Scoring Validity.

Yanagawa's work to define a framework of contextual parameters relevant to EFL listening comprehension tests will be a welcome contribution to language testing, and the presenter suggested the full framework would be made available in the future.

Reported by Jamie Dunlea
(*Society for Testing English Proficiency*)

Effects of note-taking strategy training in listening comprehension tests

Junghyun Kim
(*Sookmyung Women's University*)

This study investigated how students' test performance would be influenced after receiving the training of note-taking skills in TOEFL listening tests. Three different groups of Korean university students were compared: (a) a previously untrained group that was not allowed to take notes during the listening tests, (b) a previously untrained group that was allowed to take notes during the listening tests, (c) a previously trained group (the experimental group) that was allowed to take notes during the listening tests. Based on the contents of test preparation book by Lee (2006), the experimental group was given two lectures on effective note-taking strategies of iBT TOEFL listening comprehension tests.

The results showed that the groups that were allowed to take notes performed better in listening comprehension tests than the group that was not allowed, and that even after receiving note-taking strategies training, students did not show significant improvement of test performance. Furthermore, surveys and interviews after the tests showed that some memories or habits of note-taking training impaired their test scores. Even though many participants admitted the benefits of note-taking strategies, lack of practice time of note-taking skills or insufficient time for takings notes while testing was the main cause for the absence of its positive effects. The speaker and the floor shared the same view that more systemic training and more training time should be needed.

Reported by Iimura Hideki
(*Tokiwa University*)

Comparison of Japanese and native English-speaking raters' perspectives on oral English performance

Takanori Sato
(*Graduate School, Sophia University*)

Although various studies have suggested the similarities and the differences in language performance ratings between native speakers (NSs) and non-native speakers (NNSs) of a particular language, there is only a limited number of research which conducted direct comparisons between NS and NNS raters. In addition, there were variations among the results in the previous studies. Thus, the present study investigated the differences between NSs' and NNSs' evaluation of Japanese EFL learners' oral English performance.

A total of 30 university students performed monologues on three topics. Then, 4 NS raters and 4 NNS raters assessed the learners' performance based on the following criteria: Overall Communicative Effectiveness (OCE), Grammatical Accuracy, Fluency, Vocabulary Range, Pronunciation, and Content Elaboration/Development. The raters first assessed OCE and then the other analytical measures. A 6-point scale was used for each measurement. The result demonstrated that although the significant difference was not found between the mean

scores of OCE, the Japanese NNS raters evaluated accuracy and pronunciation higher than the NS raters. It was suggested that whereas the raters shared the common views of the communicative effectiveness, Japanese NNS raters' leniency toward accuracy and pronunciation was induced by sharing the common native language with the students. It was also revealed that while all of analytical measures of the NS raters significantly predicted OCE, only fluency and content elaboration were found to be the significant predictors in the NNS raters' case. This difference may be due to recent pedagogical approaches in Japan, such as meaning-focused, communicative language teaching.

In the 10-minute discussion, the following topics were presented: reasons or ideas for Japanese raters' leniency on grammar, the difference between the lengths of teaching experiences of the NS and NNS raters, and the necessity of comparison between the NS and NNS raters.

Reported by Katsuyuki Konno
(*Graduate School, University of Tsukuba*)

Native and non-native raters' judgment of English pronunciation at a placement test

Ji Eun Lee
(*University of Illinois at Urbana Champaign*)

This qualitative research examined an aspect of validity of a placement test for the ESL program at the University of Illinois at Urbana-Champaign. The ESL instructors judge international incoming students' intelligibility of pronunciation on a group discussion task. The placement decision is based partly on the results of this test. The researcher interviewed eight instructors with varied teaching experience and different native languages. The results showed disagreement among the instructors on the definition of "intelligibility." Native speaking instructors rated students' intelligibility more severely compared to non-native instructors. The instructors were also found to have different opinions about the use of group discussion task for the placement test. One of the questions that were raised from the audience was the issue of rater training. The researcher explained that the lack of rater training could be one of the causes of the rater disagreement. Another issue addressed the reason for using pronunciation for placement decision. The researcher indicated that the instructors have currently started discussing this issue. It seems odd to use this aspect of student oral performance for placement decision, although it may have a great advantage of practicality of administrating the test.

Hidetoshi Saito
(*Ibaraki University*)

Oral reading fluency as an assessment instrument

Masanori Suzuki
(*Pearson Knowledge Technologies*)

This presentation showed that oral reading can also be an effective assessment instrument. Reading passages out loud fluently with proper expression involves the reader's ability to recognize whole words rapidly and effortlessly and to express the meaning of the passage using appropriate pausing, intonation, and phrasing. The presenter presented three measurable aspects of oral reading ability — reading rate, reading accuracy, and expressiveness — which provide a barometer of a student's literacy.

This presentation described a validation and usability study which involved the test development team together with a US local school district and then demonstrates how the assessment tool saved teacher time, gave feedback and encouragement to students, created a digital portfolio of students' reading performances, and could potentially involve parents in the process of reading together with their children. It also reported on the relation between (ORF Oral Reading Fluency) and language learning progress. The presentation

concluded with suggestions for using such an automated ORF test in the context of Japan to create a standard for English education in Japan.

Reported by SHIOKAWA Haruhiko
(*Teikyo Kagaku Daigaku*)

Do test practice and keyword list help oral summary test performance?

Hidetoshi Saito
(*Ibaraki University*)

Saito reported on a study which compared participants' performance on an oral summary task (story retelling based on reading a passage) in four conditions: (a) practiced with a keyword list, (b) practiced without a keyword list, (c) only a keyword list, and (d) no practice and no keyword list (impromptu). He tested the viability of predictions derived from two hypotheses: The Test Practice Effect Hypothesis—test practice facilitates better performance on performance test tasks than no test practice, and the Robust Test Context Hypothesis—test takers' performance is unaffected by immediate task planning and other manipulations of task aspects (e.g., concreteness of information) in testing contexts. The results show that some predictions derived from each of the two hypotheses and that neither of the hypotheses is superior in explaining all the results. One notable finding in his study is that the use of keyword list positively influences performance, only when accompanied by practice. Without practice, keyword list does not seem to benefit the oral summary test performance; it merely seems to provide psychological security. The findings of this study are helpful in determining how teachers can help learners better prepare for an oral summary test, which in turn would enhance the development of learners' abilities for oral summary.

Reported by Atsushi Mizumoto
(*Kansai University*)

新旧 TOEIC®テストの比較検証--４テスト・セット 800 問の受験データ分析

法月健（*静岡産業大学*）
伊藤彰浩（*西南学院大学*）
島谷浩（*熊本大学*）

法月・伊藤・島谷・木下（2009）(以下、分析 1)では新旧 TOEIC®の難易度を分析し、新旧テスト間の難易度に有意差があり、表面上は従来通りの問題形式でも項目やテキストの特性に顕著な差が存在する可能性を示唆した。しかし、分析 1 は新旧１テストずつの比較であるため個々のテストの特性の差の可能性とテスト実施の順番が旧→新の順であったため新テストに練習効果の可能性が指摘されていた。

この指摘に基づき、分析 1 で使用したテストを含め、新旧それぞれ２テスト・セット（計４テスト・セットの 800 問）を使い、日本の３大学の学生 86 名を 2 グループに分け、グループ間で新旧テストの受験順序を変えてテストを実施した。

素点に基づいて新たに分析に加えた新旧２テスト･セットの平均値を 2 グループで比較したところ、グループ間に有意差はなかったが、読解部門のみ新テストが有意に高かった。一方、等化した 4 テスト･セットのラッシュ項目難易度の平均値の比較においては、分析 1 に使用した新テストの聴解項目のみ他の部門の項目よりも有意に低くくなる結果が示された。以上から、新旧テスト間には若干の難易度差があるが、これは新旧テストの恒常的な違いというよりは個々のテストの特性の要因である可能性が高く、今回の分析において新旧テストのいずれかに有利な練習効果が生じた可能性は低いことが分かった。

さらに、等化した４テスト・セット８００問を難易度別に内容分析を行った。その結果、例えば、難易度の高い聴解項目は①音声中の状況になじみがない、②音声中の表現と正答選択肢の内容との

関係が間接的、③誤答選択肢の表現が音声中に頻繁に使用される、などの特徴が挙げられた。

報告者　谷　誠司
（*常葉学園大学*）

EFL ライティング・ルーブリックの信頼性と妥当性の検証

大年順子　（*岡山大学*）
金志佳代子（*兵庫県立大学*）
久留友紀子（*愛知医科大学*）
正木美知子（*大阪国際大学*）
山西博之（*関西外国語大学*）

発表者の先生方が開発されたライティング採点表(「ルーブリック　２００９」=analytic scale)の信頼性と妥当性を詳細に検証した研究である。３０人の大学生の作文を日本人教員５名、ネイティブ教員４名が採点し、上記採点表を使用した日本人教員の結果については、５つの項目（「内容・展開」、「構成」、「文法」、「語彙」、「綴り・句読点」）と全体スコアの相関によりかなり高い信頼性が「綴り・句読点」を除く４項目について得られた。ネイティブ教員は総合評価（holistic scoring）を行い、その結果と日本人教員の採点結果の相関を基に併存的妥当性の検証がなされたが、そこでは「綴り・句読点」に加えて「構成」についての相関が有意ではなかった。更なる妥当性の検証のためにＫＪ法による教員のコメントの詳細な分析がなされ、日本人教員とネイティブ教員が重要視するポイントについて共通点や相違点が明らかになった。イントロダクションにおけるトピック・センテンスの導入の仕方などの「構成」面の違いが見られたので、今後は問題のある部分の再検討や共通理解のための説明書きの作成を進めていく。

会場からは、「文法」や「語彙」という評価点の中に適切さの概念が含まれていないのは不十分ではないかという指摘があったが、まずルーブリックによる作文の評価を広めていくことを第一と考えており、詳細な説明は別途付けたいという回答があった。ライティングの採点表の信頼性と妥当性を統計的、質的両面から分析した野心的な取組みであり、今後の展開に期待できるが、９人の教員の感覚を代表的と考えていいものか、そしてネイティブ教員の総合評価（同じ採点表を使ったものではなく）を基に併存的妥当性を見るのが適切かどうかについて少々疑問が残った。

報告者　松本佳穂子
（*東海大学*）

英語エッセイの評価に求められる教師の特性とは―指導経験，環境，及び言語能力を背景に―

長橋雅俊
（*筑波大学大学院生*）

本発表は特性の異なる評価者が英語作文を評価し、その採点の一貫性と配点の厳しさを分析した研究である。参加者として，教育経験などが多岐にわたる現職教師 10 名、および教職課程（英語）を履修中の学部生 10 名と、英語教育学を専攻し教職免許を有する大学院生 4 名より協力を得た。評価者の変数には、指導経験年数、勤務校での指導の実態、普段指導している学習者の習熟度レベル、そして言語熟達度／言語判断力が用いられた。評価者 24 名は、作文 25 編ずつを 6 段階の分析的評価法（内容展開、構成、正確さ、全体）で 2 回採点し、これらの間にはウェブサイト上での訓練タスクが課せられた。この訓練タスクは、誤文訂正、パラグラフ整合性などを問う課題からなり、その中で各評価者の言語判断力が測定された。以上の採取データより，教師が勤務している高校の生徒習得度、作文の採点の厳しさや判断の一貫性、評価者の言語能力をそれぞれに多相ラッシュ・モデルで計測、および分析した。

訓練前（1 回目）の採点では、一貫性に欠ける（ミスフィット値がつけられた）教師は 10 人中 2 人（20%）であったのに対し、学生は 14 人中 5 人（36%）に上り、指導経験が一貫した判断に貢献していたことが窺われる。評価者の言語判断力に関しても、教師は学生に比べ明らかに高かったが、この測定値が低かった教師は該当の学生同様、はずれ値が多くみられた。訓練後（2 回目）の採点では、一貫性に欠けると判断された評価者の人数は減少し、トレーニング効果が見られた。また、教師が勤務している学校のレベル差や評価者の言語判断力が採点の厳しさに影響している可能性も指摘された。

報告者　村上京子
（*名古屋大学*）

英語 Can-Do 調査分析に基づく TUFS 言語フレームワーク構築の試み

長沼君主（*東京外国語大学*）
工藤洋路（*東京学国語大学*）
吉冨朝子（*東京外国語大学大学院*）

東京外国語大学（TUFS）における「英語自律学習支援プログラム」は，2008 年度に設立された東京外国語大学英語学習支援センター（ELC）により提供されており，学生への総合的な学習サポートを目的としている。このプログラムは，①Speaking Corner（ネイティブスピーカーの英語講師とのスピーキングセッション。テーマを決めて，気軽に英語を話すことを楽しむ），②Writing Corner（ウェブベースの自動採点プログラム(Criterion)を利用したテーマ別ライティング。アドバイザーとのセッションで対面による指導も受けられる），③e-Learning Programs（TUFS e-Learning System を利用した速読・多聴学習支援プログラムなどを通して，リーディングやリスニングのトレーニングを行う。主専攻及び副専攻英語科目においては学習課題が設定されており，単位認定条件となる），④English Library（Graded Readers などを取り揃え，多読を通した英語学習を推進。スピーキング・セッションのモデルダイアログなどの多聴用コンテンツもポッドキャスト等で配信），の 4 つから構成されている。このように，インプット学習だけでなく，アウトプット学習機会も設けられており，4 技能を総合的に網羅するプログラムとなっている。英語を主専攻語または副専攻語として履修している学習者には，一定量の授業外学習課題をレベルに応じて課され，これらの学習の成果として，ポートフォリオ評価により，CEFR に準拠した「TUFS 言語パスポート」が発行される。CEFR に準拠しつつも，東京外国語大学独自の言語フレームワーク（TUFS 言語フレームワーク）を開発するため，その基礎資料として，「英語アカデミック Can-Do 調査（長沼，宮嶋，2006）」の結果を，IRT を用いて困難度の分析をし，示唆を得た。現在は Can-do 評価タスクを開発中である。フロアーからは多くの質問があり，大変活発な議論がなされた。

報告者　髙波幸代
（*筑波大学大学院*）

インハウス CAT の設計手法－実施済みテストデータを利用したシミュレーション－

秋山實
（*東北大学大学院教育情報学教育部／株式会社ｅラーニングサービス*）

アダプティブテスト（Computerized Adaptive Test；CAT）は、受検者の能力水準に応じた難易度の問題項目を提示して、短時間で効率良く高い精度で能力を測定できる。すでに大規模テストの CAT 化は実現しているが、本研究は比較的小さい特定の組織、例えば大学、企業、団体等が構築した項目プールを念頭に置き、その項目プールを用いたときの最

適アルゴリズムと実施条件を探るためのシミュレーションシステムを開発することが目的であった。当日の発表では、CAT の特徴を詳細に説明した後、インハウス CAT の定義とその必要性、さらに CAT の仕組みが報告された。また、その設計の難しさも丁寧に説明されたので、本研究の意義がフロアにいた参加者にも正確に伝わった。CAT の最適実施条件は項目プールの大きさと受験者特性によって異なるので、本研究のシステムは、主に初期条件（能力値の初期値、初期テストレット）、項目選択（Owen の方法とフィッシャー情報量）、能力推定モデル（1・2 母数ロジスティックモデル）、終了条件（能力推定値の標準誤差 SE、SE の変化値、受験項目数）などを制御できるように工夫されている。このシステムを用いた実際のシミュレーション実験の報告に対し、フロアから難易度別の項目提示率に関する質問があったが、質疑応答に十分な時間を取ることができず、個別に対応することとなった。なお、実際の CAT では受験者の心理的な側面も考慮することも必要であると思われるが、本研究はシミュレーションシステムの開発が目的であるから、そこまで射程に入れることができない。この点は CAT 化を実現した組織が検討すればよいことであり、システムの問題とは言えない。本研究のシミュレーションシステムは教育機関に公開される予定であると報告されていたので、今後、CAT の設計へ本研究のシステムが活用されることを期待したい。

報告者　服部　環
（*筑波大学*）

日本人 EFL 学習者におけるワーキングメモリ容量と記憶表象との関係：動詞分類課題による検証

高木修一
（*筑波大学大学院生*）

読解プロセスとは、テキストの記述に関する心的な表象、つまりイメージを作り上げることである。読解力はこの心的表象、つまりイメージ作りができるかどうかで決まる。本研究の目的は、心的表象の構築プロセスに関する学習者要因として、ワーキングメモリ（WM）容量について検証することである。心的表象はテキスト記述のみから作られる表層構造、テキストベース、そしてテキスト記述と読み手のスキーマなどの相互作用から作られる状況モデルの３つに区分される（e.g., Van Dijk & Kintsch, 1983）。本研究においては、EFL 学習者（24 名）において WM 容量が大きい学習者のほうが、より精緻な表象が構築されるのではないかという仮説を設定し、イベント索引化モデルの枠内で使用されている動詞分類課題を使用し、その関係性を検証した。結果として、読み手の WM が記憶表象に与える影響が明らかになった。

報告後、「L1 で認知レベルが低い学習者の場合読解力を伸ばすにはどの次元を強調すべきか」、や「内容理解能力との相関はあるか」、などの質問がなされた。ワーキングメモリの容量の大きい学習者は読解の処理能力も高いということは、メモリ容量が大きく処理能力が早いコンピュータを彷彿し、PC にはない人間だけが可能にする「想像力」や「創造力」との関係性も調査すると興味深いかもしれない。ワーキングメモリは、学習者の読解や理解という複雑な作業を行う処理能力である。しかし、これを測定するのは容易でない。報告者が指摘したように、動詞分類課題という手法に限界が感じられたが、迅速な情報処理とはどう意味をもつのかを示唆する貴重な研究であると言えよう。

報告者　李洙任
（*龍谷大学*）

スピーチにおける自己評価の妥当性:質的観点から

深澤 真
（*茨城県立竹園高等学校*）

生徒は自己評価を行うことについてどのように感じているのか。自己評価は生徒にどのような利益をもたらすのか。これらは自己評価に対して我々が持つ疑問であるが、本発表からこれらの答えとなるヒントを得ることができた。

本研究では高校生がスピーチを行った後に自己のパフォーマンスに関しての評価を行ったが、その評価に対して「客観的な評価が難しかった」や「スピーチで緊張して正確な評価ができなかった」等の意見が得られた。これらの意見から自己評価に対しての生徒自身の不安をうかがうことができる。生徒は普段から自己評価を行っていた訳ではなく、自分自身を評価するということに慣れていなかった。もし自己評価の練習を重ねればこれらの不安が徐々に取り除かれるのではないかという意見が出されたが、その時間を捻出するという課題が残されているようだ。しかし、このような自己評価の客観性に関する不安がある中でも、生徒は自身の発表を振り返る機会としてや自分の英語力を反省する機会として自己評価を見ていた。評価項目（発音、文法、流暢さ、態度など）については普段から指導が行われていたが、それらに注意を向けさせるという点においても自己評価には良い波及効果があると考えられる。

自己評価が成績に考慮され得るか、もしくはされるべきかという議論の余地は十分にあるが、動機付けという点において自己評価は有益な方法であることは認められた。課題はうまく自身を評価できない生徒への対処や不安を取り除く指導であると思われる。

報告者　佐藤敬典
（*メルボルン大学院*）

ワークショップ：分散分析―3 元配置デザインを中心に

講師　　平井　明代（*筑波大学*）
伊藤尚子（*筑波大学大学院生*）

This year’s two-hour workshop conducted by Professor Akiyo Hirai of University of Tsukuba and her student Naoko Ito concerned the use of a three-way analysis of variance (ANOVA) for analyzing language testing and learning data. The example of the three-way design data demonstrated at the workshop included three variables (rater, strategy, and proficiency), wherein it was investigated whether raters (trained/untrained) were able to assess learners’ speaking performance without being influenced by learners’ test-taking strategy (more/less) and proficiency level (high/mid/low). The rater was a between-design variable, whereas the strategy and the proficiency level were repeated-design variables. The data were analyzed using the SPSS advanced model and found to show a three-way interaction. The data were then collapsed into two-way designs and analyzed for each variable. Although this can be done via pull-down menus, the lecturers stressed the advantage of programming syntax, especially when analyzing complex interactions.

The audience posed several questions in response to which Professor Hirai and Ms. Ito stated the following: It is technically possible to enter as many variables as possible into an ANOVA design, but including less than three variables is recommended because of difficulty in interpreting interactions; it is suggested that reporting an ANOVA table along with a graph showing an interactional effect will be effective, because visual information greatly helps us interpret such an effect. Further, researchers are encouraged to report not only *p* values but also effect sizes (e.g., [partial] eta squared), because effect sizes indicate the proportion of variance accounted for by a variable. Overall, the presentation was very well received, and

the question-and-answer session was truly insightful.

Reported by Yo In'nami
(*Toyohashi University of Technology*)

講　演：

Issues in the assessment of academic vocabulary

John Read
(*University of Auckland*)

学術語彙知識の測定に関しての講演であった。

最初に，Nation (2001)をもとに，学術語彙知識の構成概念（construct）をどのように定義するのかという話から始まった。学術語彙の特徴の一つとして，幅広い学術文において見られ，学術文以外ではあまりみられないことがあげられる。

次に，学術語彙リスト(The Academic Word List [AWL])についての説明がなされた。AWL は，ビクトリア大学で学部生向けに使われた学術文 350 万語のコーパスに基づいて，頻度や使用される学問領域の幅に従って抽出されている。AWL は，学術語彙分野においてコーパスに基づいて体系的に作られた最初のリストである。また，広く頒布されており，無料である。

そして，AWL は学術語彙分野の診断テストとしても役に立つことが，主に 3 つの点から説明された。一つ目は，AWL は体系的な方法でもって学術語彙の領域（domain）を規定したこと。二つ目は，テストされるべき単語を抽出する際の基礎ができたこと。三つ目は，テスト結果から学習者が知っている AWL 単語の割合を解釈できること。

また，AWL は，学術語彙テストの作成以外に教育でも役立つことが紹介された。たとえば，AWL で頻繁に使われる単語，コロケーションの提示である。

AWL は，必ずしもあらゆる学問分野のテキストを集めているわけではなく理系分野のテキストが少ない，また，word family が単位の基礎となっており一つの word family でも複数の意味を持っている場合テスト作成時には気をつけなければならない，といった AWL の限界も同時に紹介された。

報告者　片桐一彦
（*専修大学*）

シンポジウム：

Diagnostic testing in language teaching

Coordinator & Panelist
Yasuyo Sawaki
(*Waseda University*)

Panelist
Hideki Sakai
(*Shinshu University*)
Rie Koizumi
(*Tokiwa University*)
Tomoko Ishii
(Rikkyo University)

Discussant
John Read
(*University of Auckland*)

The symposium entitled "Diagnostic testing in language teaching", which also was the conference theme, was really fulfilling. Prof. Sawaki overviewed current research on diagnosing second language ability. As an example of a diagnostic test DIALANG was explained. As a 2nd approach, extracting detailed information about learner performance from existing L2 assessments was explained. She talked about group level feedback and

individualized feedback with examples and their limitations.

Prof. Ishii gave a presentation on her investigation into a systematic and integrative way to look at vocabulary size and depth. She described vocabulary tests addressing four different aspects: vocabulary size, multiple meanings, derivatives and lexical choice. She suggested that by accumulating and using the data on these four aspects, diagnostic judgments can be made.

Prof. Sakai and Koizumi introduced the ELPA English Diagnostic Test of Grammar. The test is targeted at Japanese secondary school students, and has two characteristics: each test item belongs to one of the five NP groups classified according to their internal complexity; distracters can indicate test-takers' error patterns, or tendencies for misunderstanding. After a detailed description of the test, they concluded that teachers can use students' diagnostic information from the test and start remedial teaching based on the information.

After three presentations, Dr. Read commented on and summarized their studies, emphasizing the need to use and integrate high-stakes proficiency and achievement tests (introduced by Sawaki) and simple tests of aspects of language knowledge for classroom use (Sakai, Koizumi and Ishi) for diagnostic assessment.

Several questions were posed by the audience. Some responses from the

panelists were: Edit Grammar is one of many possible ways to receive diagnostic information and that integrated tests can also be used; tests of more complex vocabulary construct can be designed for diagnostic testing; we should reconsider the value of integrative measures such as cloze tests, dictation and elicit imitation. The symposium was really insightful for researchers and educators.

Rintaro Sato
(*Nara University of Education*)

事務局よりお知らせ

学会賞発表：

総会において，本年度から始まった The JLTA Best Paper of the Year の受賞者発表がおこなわれました。本年度の受賞者は，佐藤敬典氏（メルボルン大学大学院生）。後日，渡部良典・最優秀論文表彰委員会委員長より，受賞者の佐藤孝典さんに学会賞（最優秀論文）が授与されました。

◆ The JLTA office would be grateful if you could update us on your recent achievements relevant to the field of language testing and evaluation. Any information on your presentations, publications, awards, and so forth would be greatly appreciated. The relevance of the information will be evaluated by the office and given in the newsletter in due course.

日本言語テスト学会事務局
〒389-0813 長野県埴科郡戸倉町芝原 758
TEL 026-275-1964 FAX 026-275-1970
e-mail: youichi@avis.ne.jp
URL: http://www.avis.ne.jp/~youichi/JLTA

編集： 広報委員会
委員長 片桐一彦（専修大学），副委員長 齋藤英敏（茨城大学）
委員 秋山實（東北大学大学院／株式会社ｅラーニングサービス），
長沼君主（東京外国語大学），佐藤臨太郎（奈良教育大学）