Student achievement and French sentence repetition test scores

Deryle Lonsdale and Benjamin Millard
Brigham Young University
Provo, Utah, USA 84602
lonz@byu.edu, millbenjoe@hotmail.com

Abstract

Sentence repetition (SR) tests are one way of probing a language learner's oral proficiency. Test-takers listen to a set of carefully engineered sentences of varying complexity one-by-one, and then try to repeat them back as exactly as possible. In this paper we explore how well an SR test that we have developed for French corresponds with the test-taker's achievement levels, represented by proficiency interview scores and by college class enrollment. We describe how we developed our SR test items using various language resources, and present pertinent facts about the test administration. The responses were scored by humans and also by a specially designed automatic speech recognition (ASR) engine; we sketch both scoring approaches. Results are evaluated in several ways: correlations between human and ASR scores, item response analysis to quantify the relative difficulty of the items, and criterion-referenced analysis setting thresholds of consistency across proficiency levels. We discuss several observations and conclusions prompted by the analyses, and suggestions for future work.

Keywords: sentence repetition test, French, speech recognition, scoring methods

1. Introduction

Sentence repetition (SR) is a cost-effective testing method for assessing language learners' oral proficiency at a particular level of granularity. Test-takers hear, and then repeat, sentences of varying length and complexity. Items are carefully designed to reflect various levels of difficulty based mainly on vocabulary frequency bands, grammatical difficulty, and sentence length.

SR was first used in the early 1970's to assess native-speaking English children's development (Slobin and Welsh, 1971) and then for learners of French (Naiman, 1974). A few years later a study (Hood and Lightbrown, 1978) called into question the validity and reliability of the method based on practices and assumptions that required further attention at the time, but which helped shape the future of SR usage.

By the mid 1990's researchers had begun using SR to estimate global proficiency in second-language learners (Bley-Vroman and Chaudron, 1994). Since then work has followed in validating the SR approach and developing guidelines for its effective use (Erlam, 2009).

The basic underlying assumption of SR is that when a sentence is elicited from learners, several systems are involved: (1) the speech comprehension system, (2) the representation, (3) memory and (4) the speech production system. The learner must first process the incoming sentence through their speech comprehension system, and then form a representation in short term memory (STM). As the learner reproduces the sentence the representation must pass through the speech production system. The core idea is that once a certain item length threshold is reached, the learner is no longer be able to repeat the sentence by pure rote imitation, but would have to pass the sentence through the above-mentioned systems during the repetition process. The stage of development of these systems will constrain the response.

SR test responses can be scored by humans or by automatic speech recognition (ASR) tools (Graham et al., 2008). This is due to the highly constrained nature of the responses, permitting forced alignment techniques for scoring (Moreno et al., 1998).

We have developed and evaluated SR tests for several languages. In this paper we discuss results from administering a French SR test to college-level language learners. Elsewhere we have shown that human and ASR scoring techniques for this French SR test correlate well with each other (Millard and Lonsdale, in print). In this paper we present an analysis of how well human and ASR scoring correlate with external measures of student achievement, in particular class level and Oral Proficiency Interview (OPI) scores (Liskin-Gasparro, 1982).

The OPI is an interactive test in which the interviewer assesses the oral proficiency of the examinee. This interview technique has its strong and weak points. First, as a personal interview, it is able to more realistically duplicate a communicative event so that actual communicative proficiency is gauged. Furthermore, the interviewer is able, through probing, to determine the linguistic ceiling of the examinee in dynamic fashion. Validation studies of most other testing methods typically attempt to show a high correlation between their test results and the results of the same speakers on the OPI or another accepted oral proficiency measure (Bernstein et al., 2010; Radloff, 1991).

Another type of test is the automated proficiency test. These are typically administered via a computer program, and speech is either elicited via questions and tasks, or examinees are asked to repeat sentences. One major problem with automated testing is that it relies on ASR technology, which is not always 100% reliable. It is also not as accurate as the other testing methods, partly because proficiency is inferred by correlation rather than directly measured. This method is thus normally only used as a screening method or in low-stakes situations. The positives of this method may outweigh the problems in certain situations. The ability to instantly test a high quantity of speakers and provide quick results at low cost is very attractive.

In this paper we focus on how well student scores (human and ASR) correspond to their achievement levels, judged by their class level in college and by their OPI scores. We first discuss the data and methods used, and then present an analysis of the results.

2. Data and methods

We have created our own set of SR test items for French. SR items must be carefully engineered to assure that they are neither too simple or too complicated. To make natural and informative French SR items, we employed several language resources; following is a brief summary:

- lexical information on pronunciation, syllabification, orthography, and morphology derived from two lexical databases, BDLex (De Calmès and Pérennou, 1998) and Lexique (New et al., 2004)
- lexical frequency information from a corpus-based frequency dictionary (Lonsdale and LeBras, 2009)
- OPI guidelines on testing criteria for oral proficiency interviews (Lowe, 1982)
- part-of-speech tags provided by TreeTagger (Schmid, 1994)
- treebank parses from the French GigaWord corpus (Mendonça et al., 2009) that were generated by the Berkeley Parser for French via the Bonsai platform (Candito and Crabbé, 2009)

The parsed sentences are then analyzed; useful ones are stored in a database for later use in SR test design. In total we have thus collected an annotated corpus of some 600,000 sentences of between 5 and 20 words in length, which are most suitable for SR tests.
Figure 1 sketches the flow of information between these resources; further technical details are available elsewhere (Millard, 2011).

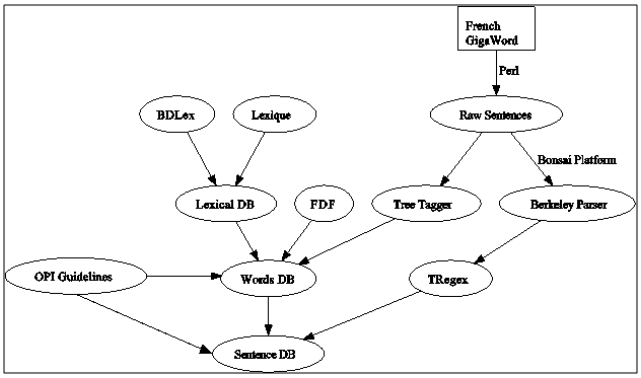

Figure 1: French SR item design dataflow

We administered an 82-item French SR test to 94 students from the French and Italian Department at Brigham Young University. Participants came from a variety of classes and proficiency levels, from French 101 (the entry level course) to graduate students and native speakers. Three of the students were absolute beginners who were recruited to ensure that there were sentences which could distinguish between the lowest-level speakers—sentences that could not be correctly imitated by memorization alone.
21 of the 94 students were also given an OPI test at about the same time; Table 1 lists the distribution of the OPI scores, and Table 2 lists the class level of all 94 participants.

ACTFL OPI Score	Participants
Novice Low (assumed)	3
Novice Mid	0
Novice High	0
Intermediate Low	3
Intermediate Mid	1
Intermediate High	4
Advanced Low	3
Advanced Mid	4
Advanced High	2
Superior	5
Total	25

Table 1: Participants with OPI scores

Class Level	Participants
Absolute beginner	3
101 (first semester)	20
102 (second semester)	26
200 (second year)	21
300 (third year)	11
400 (fourth year)	8
500 (graduate)	2
Native	3
Total	94

Table 2: All participants by class level

Responses were recorded and underwent post-processing to improve audio quality. Then they were graded, both by humans and by ASR methods. In particular, 4 trained human raters (one native and 3 non-native French speakers) scored each elicited response using an automated syllable scoring interface. The ASR engine we used for scoring was the Sphinx4 engine (Walker et al., 2004) with French language and acoustic models (Deléglise et al., 2009).
We consider three different SR scoring methods that have been discussed previously:

- The 4-score scalar method subtracts one point for each error in a sentence until 0 is reached, giving a score from 0 to 4.
- The binary method simply assigns a 1 to a perfect response sentence, otherwise a 0.
- The percentage method is given according to what percentage of the syllables (or words) in the response sentence are correct.

In ASR scoring, the speech recognizer reads each test response audio file and performs standard speech-to-text transcription on the contents. For the work reported here, we initially tested several ASR configurations to determine which worked best with the right amount of speed and accuracy. We determined that the flatLinguist, a simple configuration that uses only the acoustic model and a grammar, would be sufficient for SR scoring. This application is very fast but normally less accurate than the core engine. It works relatively well with SR, however, because the exact sequence of expected words is known by the system.

Different finite-state language grammars can be incorporated into the flatLinguist to tell the recognizer exactly which set of symbols to expect in which order. In SR testing, two main categories of grammars are typically used: traditional word-based grammars and syllable-based grammars.

Word-based grammars have been the norm for scoring SR items with ASR. This is the easiest approach since a closed set of predefined words are expected in the SR test. Comparison of the input is made against all possible word sequences, and the ASR engine looks to match the exact expected sentence. Any deviation from the expected order will cause recognition to fail. Hence ASR scoring often has a binary flavor—for normal grammars the outcome is all-or-nothing. Disfluencies such as restarts are acceptable in such grammars as long as the sentence is uttered in its entirety at some point. More involved corrections, repairs and other speech disfluencies are not handled as well with this type of narrow grammar. A valid sequence for a word-based grammar might be the sentence:

nous avons travaillé

meaning "we have worked".

Word-based grammars can be generalized to an extent, creating a Kleene star grammar. This type of grammar has more freedom in recognizing normal speech phenomena like corrections, restarts, stutters, pauses, filled pauses, and so on. It is much more forgiving than the normal word-based grammar, seeking to match 0 or more occurrences of any of the listed words. Sometimes this flexibility causes the recognizer to be less accurate, but on the other hand it allows each word to be a start and end point. This allows the system to start on any word, skip any word, end on any word, and process any word as many times as necessary. It does, however, often overgeneralize and overcompensate, making it, at times, inaccurate. A Kleene star version of our sample sequence would be:

nous* avons* travaillé*

which would admit 0 or more instances of the word nous followed by 0 or more instances of the word avons, followed by 0 or more instances of the word travaillé.

Syllable-based grammars are very similar to the word-based grammars except that each word is broken into its constituent syllables. In previous SR testing for English, we have only approximated syllable scoring by breaking up the words into other words that sounded similar to the desired syllables. To specify syllable grammars for French, we syllabified the SR item words, and then the syllables that were not homophonous with actual words were added to the system dictionary as pseudo-words. About two hundred syllables were added to the dictionary in this fashion. Our sample sentence encoded in a syllable grammar would look like this:

nous a vons beau coup tra vai llé

and a Kleene star syllable-based sentence would look like this:

nous* a* vons* beau* coup* tra* vai* llé*.

Using a syllable grammar afforded the ASR engine a closer similarity to the human scorer as human scoring is also done on a syllable basis.

3. Results

In this section we report on how well the ASR scoring compares with human scoring for the participants' tests. Three evaluation methods are performed: correlations, item response analysis, and criterion-referenced analysis.

3.1. Correlations

Much prior work on scoring sentence repetition tests involves exploring correlations of various sorts:

- *multiple human scorers rating the same test items*: Generally these correlations are high no matter which of the scoring methods mentioned above is used (Lonsdale et al., 2009). Even non-native speakers of English are able to carefully grade English items without adversely affecting interrater reliability.

- *sentence repetition scores versus other tests*: A close correlation with more traditional "gold standard" methods (oral interviews, simulated interviews, etc.) is desirable for establishing the effectiveness and validity of sentence repetition tests.

- *human scoring versus scoring by automatic computerized methods*: As ASR methods have developed and are being used more commonly in test scoring, automatically derived results are compared against the "gold standard" of expert evaluations (Cook et al., 2011).

For our French test, we first correlated test scores from each scoring method against the OPI (see Table 3). As expected, the OPI correlations with human evaluations are all high. Though slightly lower than the human rating correlations, our ASR results correlate very well with OPI results, particularly the ASR syllable binary scores. The 4-score values underperform substantially. This casts some doubt on the usefulness of the 4-score for ASR scoring of sentence repetition tests.

3.2. Item analysis

We also performed an analysis of the test results to gauge item (stimulus sentence) difficulty and ability to discriminate between participants at different proficiency levels. Item response theory (IRT) has been used in previous studies to determine the ability of test items to distinguish between learner levels (Lord, 1980). Once the best-discriminating SR items are found, the test can be shortened, re-calibrated, and re-tested (Grimes, 1992). We have

	ACTFL OPI Results
class level	0.913
human 4-score	0.912
human binary score	0.878
human percent	0.905
ASR word binary	0.877
ASR word 4-score	0.670
ASR word percent	0.814
ASR syllable binary	0.883
ASR syllable 4-score	0.669
ASR syllable percent	0.822

Table 3: Pearson correlations: ASR scores vs. OPI results

used IRT analysis in the past to identify top-performing English SR items (Graham et al., 2008).
Using the Winsteps program[1], we assessed French item relative difficulty. Figure 2 plots the measures for SR test sentences administered to our group of participants, based on human 4-score measures.

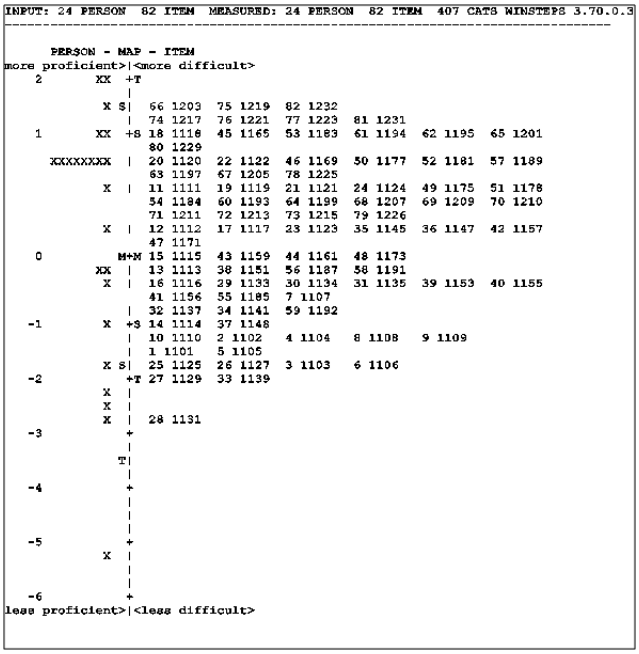

Figure 2: IRT analysis: SR items (human 4-score measures)

As expected, the most difficult items are for the most part those associated with higher proficiencies. Many of the items do discriminate well: the participants (shown on the left-hand side) spread out across the levels. Near the upper levels, one large cluster is formed: it consists of highly proficient students including native speakers. Item difficulty (on the right-hand side) was mostly normally distributed. However, there is noticeable skewing at the top: too many items were considered difficult, even for this comparatively advanced group of learners.

[1]http://www.winsteps.com/winsteps.htm

Compare this with Figure 3. Using syllable binary ASR scores for all 92 of the participants (except for two whose audio was corrupted and hence excluded), the picture changes somewhat. This analysis shows that many of the same items are listed (right-hand side) as the most difficult, but it spreads them out across a greater distribution through the proficiency levels.
The participants, many of whom are intermediate speakers, now cluster (left-hand side) around the intermediate level (near -2). This analysis shows that ASR is able to distinguish between the participants at this level to a high degree. In addition, the participant scores are much more normally distributed.

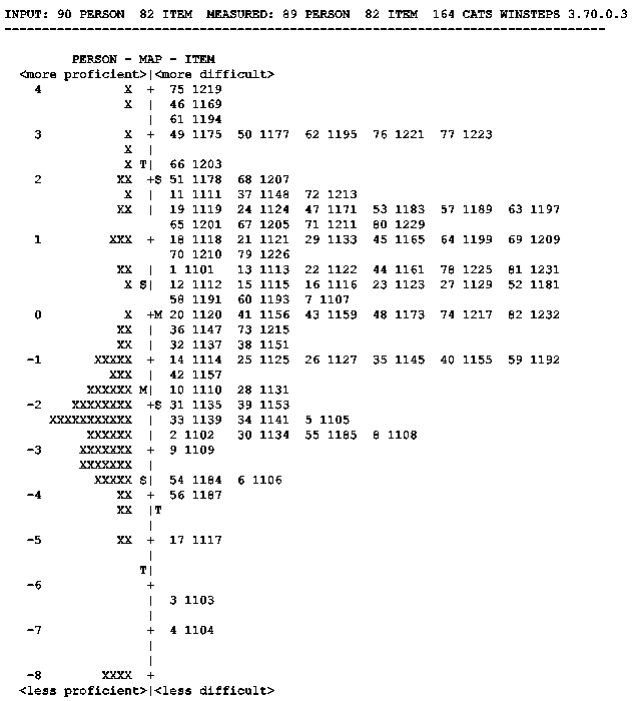

Figure 3: IRT analysis: SR items (ASR syllable binary scores)

Finally, we then took only the top 61 items according to the IRT and reran the correlation measures; the results are shown in Table 4. There is very little variation in the correlations between this reduced set and the full set, indicating that we could reduce the number of test items accordingly without loss of test effectiveness.

3.3. Criterion-referenced analysis

Another method that is emerging in the field of language assessment is criterion-referenced analysis (Brown and Hudson, 2002). This type of assessment has not traditionally been used to analyze SR tests because of the lack of incorporating criteria into proficiency testing. Partly because of this lack, norm-referenced interpretations have been more pervasive. The aid of natural language processing techniques, however, is changing the ability of researchers to include criteria in test development and allow for criterion-referenced interpretations.
For this test in particular, participants with OPI scores can be evaluated on their performance at each proficiency level

	ASR syllable binary	Human 4-score	Human percentage
ACTFL OPI results	0.886	0.902	0.919
ASR syllable binary	1	0.893	0.874
Human 4-score		1	0.973
Human binary			0.886
Human percentage			1

Table 4: Pearson correlations: human & ASR scores after IRT

and a threshold of consistency can be established between the participant scores at the two levels. This type of analysis is useful for the calibration of an SR test and for future applications like adaptive computerized testing.

We performed a criterion-referenced analysis for our SR test. During the test design, each stimulus sentence was associated to an ILR proficiency level between 1 and 3 based on OPI testing features (Lowe, 1982) to enable criterion-referenced analysis of student results. For example, items thus associated to level 1 (i.e. ACTFL Intermediate 4-6) are correlated to OPI scores and ASR scores of various types (binary and percentage scores for both syllable and word-based grammars); see Table 5. We then plot the OPI results against human and ASR scores and use linear regression to determine the best fit line. Cutoff thresholds are set at points of maximal separation between proficiency levels. Outliers fall into the top-left or bottom-right quadrants.

ACTFL	ILR
Novice	0
Intermediate	1
Advanced	2
Superior	3

Table 5: Proficiency level correspondences

Following is a summary of some of the highlights from pertinent results; an exhaustive examination is beyond the scope of this paper.

Levels 0 and 1:

Students were assigned to level 0 on the sole basis of their inability to perform consistently at level 1. All scoring methods showed a clear separation between the Intermediate Low participants (level 4 on the ACTFL scale) and the absolute beginners who were tested and listed as level 1. For example, Figure 4 shows the analysis for ASR word percentage scoring of Level 1 students. Note the clear separation from the three Level 0 novices at the bottom.

Level 2:

The level 2 ASR analyses are probably the most informative. They largely succeed in not placing any first-year and second-year in this advanced category. Only one participant score falls outside of the expected thresholds for the binary ASR method, giving it a 95% accuracy rate for this level (see Figure 5).

On the other hand, Level 2 items were very problematic for human scorers. ASR has a greater ability to distinguish between the intermediate and advanced speakers. For example, by almost all human scoring regimes the Intermediate High participants were placed too high. There are probably

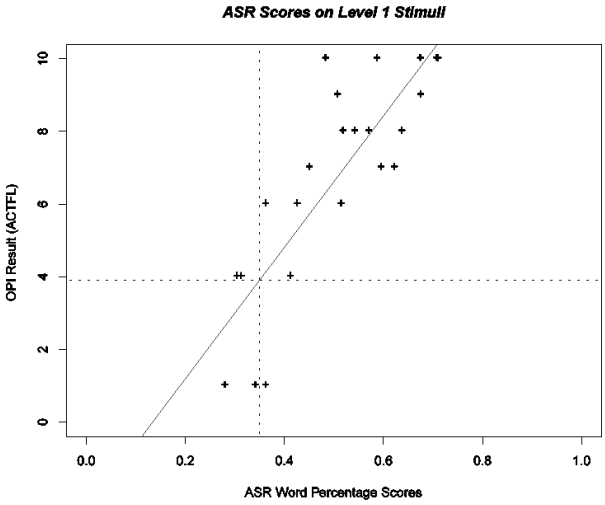

Figure 4: ASR syllable 4-scores on level 1 items

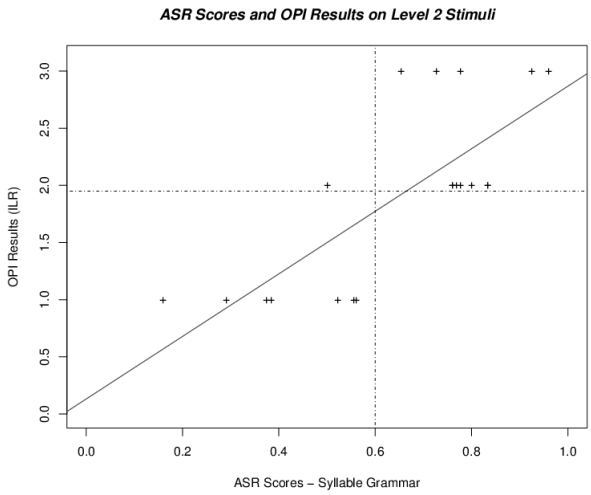

Figure 5: ASR syllable grammar scores on level 2 items

items at Level 2 that did not discriminate well enough; this may be solved in time as the poorly performing items are eventually culled out.

Level 3:

In the analysis of level 3 sentences, the most interesting result is the ability of the items to draw a sharp distinction between native superiors and non-native superiors (though all have a score of 10). The two non-native superiors are

grouped much more closely with the advanced speakers (scores 7-9), even by using Level 3 sentences. Figure 6 illustrates this point, where the 3 natives show clear separation from the others in ASR Kleene word scoring.

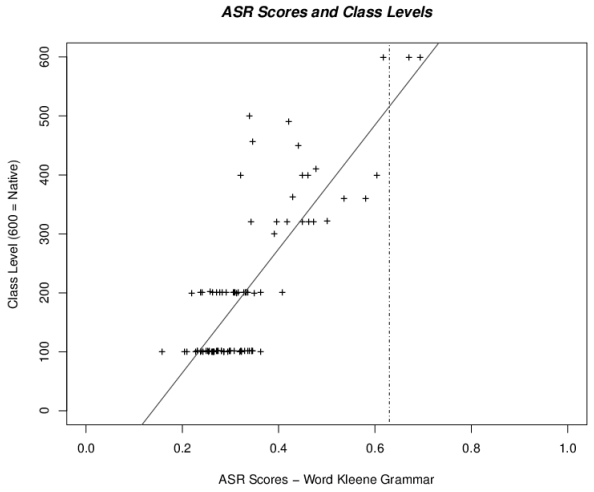

Figure 6: ASR Kleene word scoring on Level 3 items

General ASR performance:
The ASR percentage scoring is too generous at low proficiency levels. This is not uncommon in speech recognition, where the engine is attempting as best it can to accommodate spoken input with respect to the models specified. As we have seen, ASR scoring also exhibits inaccuracies for superior or native speakers.

Item difficulty:
In analyzing the items we observed floor and ceiling effects (items that are too easy or hard). There may be a need for easier stimulus sentences to further separate intermediate and novice speakers. More work also seems necessary to better distinguish natives from non-native superior speakers. In almost all of the thresholds set for Level 3 speakers (superiors), the non-native superiors consistently fell behind and would have been classed with their advanced speaker counterparts.

Methods comparison:
4-score methods generally performed worst in the correlations and appear to perform the worst at each of the proficiency levels. Since this scoring method has very little tolerance for error, it does not provide for a strong separation between the levels. However, syllable binary ASR scoring rendered the best results in every area of analysis, and should be the default scoring approach, at least for French ASR.

The inclusion of testing-feature-based criteria in test development has greatly enhanced discriminating ability. With a proficiency association to each test item, we can now easily and accurately distinguish between the 4 major proficiency groups—novice (0), intermediate (1), advanced (2) and superior (3)—by setting automatically computed thresholds between them.

4. Conclusions and future work

In this paper we have sketched how we used various language resources to develop French SR test items, and then administered that test to almost 100 French language learners. In our analyses of the results we have shown that SR testing can accurately estimate oral proficiency in French speakers, even when scored by ASR. High correlations are obtained at each level using most of the ASR scoring techniques.

In our IRT analysis of the items we identified the most effective items, and in a post-hoc analysis showed how we can obtain similar result from using only 61 items (versus the original 84 items).

We also carried out a criterion-referenced analysis of the data, associating items with levels of achievement. This led to numerous interesting observations about how well various scoring techniques distinguish students at different proficiency levels.

We see several possible directions for future related work. In this effort we used publicly available off-the-shelf ASR language and acoustic models trained on native speaker data. However, we are testing non-native learners of the language whose language by definition deviates greatly from native speech. We, as well as others, have incorporated learner errors into language models to improve the performance of ASR grading for SR items (Han et al., 2010; Lonsdale and Matsushita, 2013). These techniques should transfer to French testing, given enough relevant raw learner data.

Two core areas of oral language proficiency are accuracy and fluency (Housen and Kuiken, 2009). SR tests evaluate the former—how well a participant can accurately reproduce a stimulus sentence. Work in using fluency measures as a part of oral proficiency testing has increased greatly in the last 3 years. Fluency measures are typically based on ASR features from spontaneous speech or prompted conversations. Through principled combination of automatically computed SR and fluency measures, more exact and comprehensive computerized assessment of oral proficiency is possible (Lonsdale and Christensen, 2014). This direction could be pursued for French.

Automatic SR scoring opens up another possibility: adaptive language testing. With real-time scoring results and items of varying complexity, a test could be calibrated online based on the responses it receives. This helps render the test more tractable to the participant and more effective to the evaluator. Though we have not yet implemented such a system, simulations run on prior data shows that an English test we developed could be reduced in length by about two-thirds without loss of scoring precision (Lonsdale and Christensen, 2011). Similar results could probably be obtained for French. Our use of criterion-referenced analysis would be especially helpful in informing an adaptive system on which items are most appropriate for students at a given level of achievement.

5. References

Bernstein, J., Moere, A. V., and Cheng, J. (2010). Validating automated speaking tests. *Language Testing*, 27(3):355–377.

Bley-Vroman, R. and Chaudron, C. (1994). Elicited imitation as a measure of second-language competence. In Tarone, E., Gass, S., and Cohen, A., editors, *Research methodology in second-language acquisition*, pages 245–61. Lawrence Erlbaum, Northvale NJ.

Brown, J. D. and Hudson, T. (2002). *Criterion-referenced language testing*. Cambridge University Press.

Candito, M.-H. and Crabbé, B. (2009). Improving generative statistical parsing with semi-supervised word clustering. In *Proceedings of the International Workshop on Parsing Technologies (IWPT)*, pages 169–172. Association for Computational Linguistics.

Cook, K., McGhee, J., and Lonsdale, D. (2011). Elicited imitation for automatic prediction of OPI test scores. In *Proceedings of the Sixth Workshop on Innovative Uses of NLP for Building Educational Applications*, pages 30–37. Association for Computational Linguistics.

De Calmès, M. and Pérennou, G. (1998). BDLex: a lexicon for spoken and written French. In *Proceedings of the First International Conference on Language Resources and Evaluation (LREC '98)*.

Deléglise, P., Estève, Y., Meignier, S., and Merlin, T. (2009). Improvements to the LIUM French ASR system based on CMU Sphinx: What helps to significantly reduce the word error rate? In *Proceedings of the 10th Annual Conference of the International Speech Communication Association (INTERSPEECH 2009)*, pages 2123–2126.

Erlam, R. (2009). The elicited imitation test as a measure of implicit knowledge. In Ellis, R., editor, *Implicit and explicit knowledge in second language learning, testing and teaching*. Multilingual Matters, Bristol, UK and Buffalo, NY.

Graham, C. R., Lonsdale, D., Kennington, C., Johnson, A., and McGhee, J. (2008). Elicited Imitation as an Oral Proficiency Measure with ASR Scoring. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2008)*, pages 1604–1610.

Grimes, J. E. (1992). Calibrating sentence repetition tests. In Casad, E., editor, *Windows on Bilingualism*, pages 73–85. Summer Institute of Linguistics and the University of Texas at Arlington, Dallas.

Han, N.-R., Tetreault, J., Lee, S.-H., and Ha, J.-Y. (2010). Using an Error-Annotated Learner Corpus to Develop an ESL/EFL Error Correction System. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC '10)*.

Hood, L. and Lightbrown, P. (1978). What children do when asked to 'say what I say': does elicited imitation measure linguistic knowledge? *Reprints from Allied Health and Behavioral Sciences*, 1:195–220.

Housen, A. and Kuiken, F. (2009). Complexity, accuracy, and fluency in second language acquisition. *Applied Linguistics*, 30:461–473.

Liskin-Gasparro, J. E. (1982). *Educational Testing Service Oral Proficiency Testing Manual*. Educational Testing Service, Princeton, NJ.

Lonsdale, D. and Christensen, C. (2011). Automating the scoring of elicited imitation tests. In *Proceedings of the ACL-HLT/ICML/ISCA Joint Symposium on Machine Learning in Speech and Language Processing*.

Lonsdale, D. and Christensen, C. (2014). Combining sentence repetition and fluency features for oral proficiency measurement. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC '14)*.

Lonsdale, D. and LeBras, Y. (2009). *A Frequency Dictionary of French: Core Vocabulary for Learners*. Routledge, New York, NY.

Lonsdale, D. and Matsushita, H. (2013). Modeling speech errors by analogy. In West, R. L. and Stewart, T. C., editors, *Proceedings of the 12th International Conference on Cognitive Modeling (ICCM 2013)*, pages 17–22. Cognitive Science Society.

Lonsdale, D., Dewey, D. P., McGhee, J., Johnson, A., and Hendrickson, R. (2009). Methods of scoring elicited imitation items: An empirical study. Presentation to the American Association of Applied Linguistics (AAAL).

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Lawrence Erlbaum Associates, Inc., Mahwah, NJ.

Lowe, P. (1982). *ILR Handbook on Oral Interview Testing*.

Mendonça, A., Graff, D., and DiPersio, D. (2009). French Gigaword Second Edition.

Millard, B. and Lonsdale, D. (in print). French oral proficiency assessment: Elicited imitation with speech recognition. In Côté, M.-H. and Mathieu, E., editors, *Proceedings of the 41st Linguistic Symposium on Romance Linguistics (LSRL 2011)*.

Millard, B. (2011). Oral proficiency assessment of french using an elicited imitation test and automatic speech recognition. Master's thesis, Brigham Young University.

Moreno, P., Joerg, C., Thong, J. V., and Glickman, O. (1998). A recursive algorithm for the forced alignment of very long audio segments. In *Proceedings of the International Conference on Spoken Language (ICSLP-8)*.

Naiman, N. (1974). The use of elicited imitation in second language acquisition research. *Working Papers on Bilingualism*, 2(1).

New, B., Pallier, C., Brysbaert, M., and Ferrand, L. (2004). Lexique 2: A new French lexical database. *Behavior Research Methods, Instruments, & Computers*, 36(3):516–534.

Radloff, C. F. (1991). *Sentence repetition testing for studies of community bilingualism*. Dallas Summer Institute of Linguistics and the University of Texas at Arlington.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*.

Slobin, D. I. and Welsh, C. A. (1971). Elicited imitation as a research tool in developmental psycholinguistics. In Lavatelli, C. B., editor, *Language training in early childhood education*. University of Illinois Press, Urbana, IL.

Walker, W., Lamere, P., Kwok, P., Raj, B., Singh, R., Gouvea, E., Wolf, P., and Woelfel, J. (2004). Sphinx-4: A flexible open source framework for speech recognition. Technical Report SMLI TR2004-0811, Sun Microsystems Inc.