A User Study: Technology to Increase Teachers' Linguistic Awareness to Improve Instructional Language Support for English Language Learners

Jill Burstein, John Sabatini, Jane Shore, Brad Moulder, and Jennifer Lentini

Educational Testing Service
666 Rosedale Road, Princeton, New Jersey 08541
{jburstein, jsabatini, jshore, bmoulder, jlentini}@ets.org

Abstract

This paper discusses user study outcomes with teachers who used *Language Muse*[SM] a web-based teacher professional development (TPD) application designed to enhance teachers' linguistic awareness, and support teachers in the development of language-based instructional scaffolding (support) for their English language learners (ELL). System development was grounded in literature that supports the notion that instruction incorporating language support for ELLs can improve their accessibility to content-area classroom texts –in terms of access to content, and improvement of language skills. Measurement outcomes of user piloting with teachers in a TPD setting indicated that application use increased teachers' linguistic knowledge and awareness, and their ability to develop appropriate language-based instruction for ELLs. Instruction developed during the pilot was informed by the application's linguistic analysis feedback, provided by natural language processing capabilities in *Language Muse*.

1 Introduction

Statistics show that between 1997 and 2009 the number of ELLs enrolled in U.S. public schools has increased by 51% (National Clearinghouse for Language Acquisition, 2011). ELLs who have *lower literacy skills*, and who are reading *below* grade level may be mainstreamed into regular content-area classrooms, and may not receive supplemental English language instruction. Unfortunately, K-12 content-area teachers[1] are *less likely to be trained* to adapt their instructional approaches to accommodate the diverse cultural and linguistic backgrounds of students with varying levels of English proficiency (Adger, Snow, & Christian, 2002; Calderón, August, Slavin, Cheun, Durán, & Madden, 2005; Rivera, Moughamian, Lesaux, & Francis, 2008; Walqui & Heritage, 2012). This situation motivated the development of *Language Muse*[SM], a web-based application designed to offer teacher professional development (TPD) for content-area teachers to support their understanding of potential sources of linguistic unfamiliarity that may obscure text content for ELLs, and their ability to develop relevant language-based instructional scaffolding. We reasoned that prerequisite to effectively planning or implementing instructional supports for ELLs, teachers first needed to be able to *recognize* potential sources of linguistic difficulty. Further, teachers might need training about the specific linguistic structures that might be unfamiliar to learners, and which might lead to learners' inaccessibility to core content in text.

The motivation for *Language Muse*, thus, grew from the need to provide teachers with training about linguistic features in texts that may be unfamiliar to learners. In complement to training videos and reading resources, *Language Muse* contains a module that provides automated and explicit linguistic feedback for texts, and is intend-

[1] These are Kindergarten-12th grade teachers of subject areas, including math, science, social studies, and English language arts.

Proceedings of the 2th Workshop of Natural Language Processing for Improving Textual Accessibility (NLP4ITA), pages 1–10,
Atlanta, Georgia, 14 June 2013.

ed to support teachers in the development of lesson plans with language-based instructional activities and assessments to support reading and content comprehension of texts. The linguistic feedback module uses various natural language processing methods to provide feedback at the *vocabulary*, *phrasal*, *sentential*, and *discourse* levels. Another motivation of application was efficiency. Even with a strong linguistic awareness, manual identification of linguistic features would be a very time-consuming process.

Outcomes from pre-post teacher assessments delivered through user piloting with teachers indicated that teachers who used *Language Muse* showed gains in linguistic knowledge. Outcomes also indicated that *Language Muse* use supported teachers in the ability to develop appropriate language-based instruction for ELLs, informed by the application's linguistic analysis feedback.

2 Related Work

In a brief literature review, we address the language demands for ELLs in reading content-area texts, and the need for relevant teacher training for content-area teachers (Section 2.1). We also discuss NLP-related applications that support the linguistic analysis of texts -- typically in the context of developing *readability measures* -- which continues to be a prominent area of research; other research supports student tools allowing direct interaction with language forms (Section 2.2).

2.1 Language Demands on ELLs, and Teacher Training

Language Demands on ELLs. The English Language Arts Common Core State Standards[2] (*Standards*) (NGA Center & CCSSO, 2010) has now been adopted by 46 states and is a trend-setter in U.S. education. The *Standards* emphasize the need for all learners (including ELLs[3]) to read progressively more complex texts across multiple genres in the content areas, preparing learners for college and careers. To accomplish this, learners must have familiarity with numerous linguistic features related to vocabulary, English language structures, and a variety of text structures (discourse).

In terms of **vocabulary demands**, research reports on investigations of academic vocabulary and the *Tier* word system (Beck, McKeown, & Kucan, 2008; Calderón, 2007). Specifically, *Tier 1* words are those used in everyday conversation; *Tier 2* words are general academic words; and *Tier 3* words are found in specific domains (Beck et al, 2008; Coleman & Pimental, 2011a). All three *Tiers* are necessary to academic content learning. *Key content-area terms* in any text would include the vocabulary that students are expected to learn regardless of the *Tier*. However, there are many other vocabulary terms in the same text that may or may not be key content, but may still pose difficulties for an ELL reader. For instance, the phrase "*rock star*" is a figurative term whose meaning is not obvious from knowing the various meanings of "*rock"* or "*star"*. A deficit in *morphological awareness* can be a source of reading comprehension difficulties among native speakers of English, (Berninger, Abbott, Nagy, & Carlisle, 2009; Nagy, Berninger, & Abbot, 2006), but even more so among ELLs (Carlo, August, McLaughlin, Snow, Dressler, Lippmann, & White, 2004; Kieffer & Lesaux, 2008). Teaching morphological structure has been shown to be effective with ELLs (Lesaux, Kieffer, Faller, & Kelley, 2010; Proctor, Dalton, Uccelli, Biancarosa, Snow, & Neugebauer, 2011). *Native language support* can also aid students in learning text-based content (Francis, August, Goldenberg, & Shanahan, 2004). Specifically, lessons that incorporate *cognates* (e.g., *individual* (English) and *individuo* (Spanish)) have been found to be effective in expanding English vocabulary development and aiding in comprehension (August, 2003; Proctor, Dalton, & Grisham, 2007). *Polysemous words* can contribute to overall text difficulty. Papamihiel, Lake & Rice (2005) specifically discuss difficulties of content-specific, polysemous words, where the more common meaning may lead to a misconception when using that meaning to infer the more specific content meaning (e.g., *prime* in *prime numbers*). Unfamiliar cultural references (e.g., *He's a member of the Senate.*), when reading an unfamiliar language to learn unfamiliar content, imposes a triple cognitive load for ELLs (Goldenberg, 2008).

With regard to **sentence-level demands**, long, multi-clause sentences can present frustrating

[2] http://www.corestandards.org/

[3] For details and about *Standards* and ELLs, see: http://ell.stanford.edu/.

complexities. Readers need to analyze sentence clauses to understand and encode key information in working memory as they build a coherent mental model of the meaning of a text (Kintsch, 1998). Different subject areas often have sentential and phrasal structures that are unique to that subject, resulting in comprehension breakdowns, e.g., the noun phrases in math texts "*a number which can be divided by itself …*" (Schleppegrell, 2007; Schleppegrell & de Oliveira, 2006).

Regarding **discourse structure demands,** content-area texts may represent varying discourse relationships. Discourse relations such as, compare-contrast, cause-effect can all be intermingled within a single passage (Goldman & Rakestraw, 2000; Meyer, 2003). Teachers need to learn how to identify discourse-level information and develop scaffolding to support students' ability to navigate discourse elements in texts. Students may also be challenged in keeping track of and resolving referential (anaphoric) relationships. *Pronomial reference* can be a challenge for ELLs in texts with multiple characters or agents (Kral, 2004). An equal challenge concerns the resolution of referential relations among nouns, phrases, or ideas - a common occurrence in expository texts- whether the category of reference is pronominal, synonymy, paraphrase, or determiner, e.g., *this*, *that,* or *those* (Pretorius, 2005). Also critical to learning new content is understanding *connector words* functions (e.g., *because*, *therefore*) for building text cohesion (Goldman & Murray, 1992; Graesser, McNamara, & Louwerse, 2003).

Teacher Training. Teachers need to become *linguistically aware* of aspects of the English language that present potential obstacles to content access for ELLs. Yet, teachers often lack training in the identification of features of English that may challenge diverse groups of ELLs (Adger et al., 2002; Calderón et al., 2005; Rivera et al , 2008; Walqui & Heritage, 2012), and in the implementation of strategies to help ELLs academic language and vocabulary acquisition (Flinspach, Scott, Miller, Samway, & Vevea, 2008). Further, the number of teachers trained in effective instructional strategies to meet the range of needs of ELLs has not increased consistently with the rate of the ELL population (Gándara, Maxwell-Jolly, & Driscoll, 2005; Green, Foote, Walker & Shuman, 2010). Studies suggest that teachers with specialized training have a positive impact on student performance (Darling-Hammond, 2000; Peske & Haycock, 2006).

2.2 Text Accessibility and NLP

Considerable research in NLP and text accessibility has focussed on linguistic properties of text that render a text relatively more or less accessible (comprehensible). This research stream has often fed into applications offering *readability measures* – specifically, measures that predict the grade level, or grade range of a text (e.g., elementary, middle or high-school). Foundational research in this area examined the effect of morphological and syntactic text properties. Flesch (1948) reported that text features such as syllable counts of words, and sentence length were predictors of text difficulty. Newer research in this area has included increasingly more NLP-based investigations (Collins-Thompson & Callan, 2004; Schwarm & Ostendorf, 2005; Miltsakaki, 2009). Some research examines text quality in terms of discourse coherence of well-formed texts (Barzilay & Lapata, 2008; Pitler & Nenkova, 2008; Graesser, McNamara, & Kulikowich, 2011).

Human evaluation of text complexity in curriculum materials development (i.e., adaptation and scaffolding of reading texts, and the creation of activities and assessments) is a time-consuming, and typically intuitive process. Determining text complexity is also not a clear and objective measure. For example, what is complex for a native English speaker reading *on* grade level may vary from what is complex (or unfamiliar) for an ELL reading *below* grade level. This area of research continues to grow as is evidenced by NLP shared tasks (Mihalcea, Sinha & McCarthy, 2010), in the research and educational measurement communities (Burstein, Sabatini, and Shore, in press; Nelson, Perfetti, Liben & Liben, 2012).

The REAP system uses statistical language modeling to assign readability measures to Web documents (Collins-Thompson & Callan, 2004). This system is used in college-level ESL classrooms for higher level ESL students. It is designed to support automatic selection and delivery of appropriate and authentic texts to students in an instructional setting (Heilman, Zhao, Pino, & Eskenazi, 2008). Teacher users can set a number of constraints (e.g., reading level, text length, and

target vocabulary) to direct the text search. The system then automatically performs the text selection. The system also has tools that allow English learners to work with the text, including dictionary definition access and vocabulary practice exercises. In pilot studies with high-intermediate learners in a university setting, a post-test showed promising learning outcomes (Heilman et al, 2008).

WERTi (Working with English Real Texts interactively) (Meurers et al., 2010) is an innovative Computer-Assisted Language Learning (CALL) tool that allows learners to interact directly with NLP outputs related to specific linguistic forms. In the context of a standard search environment, learners can select texts from the web. NLP methods are applied to identify linguistic forms that are often problematic for ELLs, including, use of determiners and prepositions, *wh*-question formation, and phrasal verbs in the texts. Meurers et al. point out that this CALL method is intended to draw learners' attention to specific properties of a language (Rutherford and Sharwood Smith , 1985). ELLs' direct interaction with different linguistic forms could support them in language skills development, and content accessibility.

To our knowledge, *Language Muse* is unique from other NLP applications in that it is designed as a teacher professional development (TPD) application intended to enhance teachers' linguistic awareness, and as a result, aid teachers in the development of language-based scaffolding to support learners' content accessibility, and language skills development. Key text complexity drivers *cannot* be communicated to teachers through numerical aggregate readability measures which appear to be the predominant approach to analysis of text difficulty described in the literature. ***Language Muse* fills a critical TPD gap.** The application is an innovative resource designed to help teachers understand the specific linguistic features that may contribute to text difficulty and ELLs' inaccessibility to text content; linguistic feedback features in SYSTEM are grounded in the literature about ELL language demands (Section 2.1).

3 *Language Muse*

Language Muse is a web-based application for enhancing teachers' linguistic awareness and supporting the development of language-based instruction for ELLs. It uses NLP methods to provide explicit linguistic feedback that is grounded in the literature discussing ELL language demands and needs (Section 2.1).

We will discuss (a) the system's specific lesson planning components, and (b) a text exploration tool that provides automated linguistic feedback.

The *lesson planning component* has three modules that support the creation of lesson plans, and related activities and assessments. To create a lesson plan, teachers complete a lesson plan template (provided by the system) with five sections commonly found in lesson plans: (a) standards and objectives, (b) formative and summative assessments, (c) engaging student interest/connecting to student background knowledge, (d) modeling and guided practice, and (e) independent practice. Teachers use system functionality to link specific texts to a lesson plan. Texts have typically been analyzed, first, using the feedback tool. Feedback is then used to *inform* lesson plan development. Activities and assessments may also be created for a specific lesson plan and will also be linked to the plan. Teachers are instructed to use linguistic feedback from the tool to develop language-focused activities and assessments that can be used to support the language objectives proposed in the lesson plan. The *Text Explorer & Adapter (TEA-Tool)* feedback module uses NLP methods for automatic summarization (Marcu, 1999); English-to-Spanish machine translation (SDL n.d.); and, linguistic feedback. A text[4], or a webpage with the relevant text is uploaded, or accessed, respectively, into the *TEA-Tool* module. The summarization capability may be used to reduce the amount of text that learners are exposed to reduce cognitive load. The machine translation capability can be used to offer native language support to learners with little English proficiency. The primary focus in this section, however, will center around the linguistic feedback that supports the *core goal* of building teachers' awareness of specific linguistic features in texts**.** The linguistic feedback includes specific information about vocabulary, phrasal and sentence complexity, and discourse relations. For *vocabulary*[5], categories of feedback include: academic words, cognates, collocations and figurative words and terms, cultural

[4] Microsoft Word, PDF, and Plain text files may be used.

[5] For academic words, cognates, cultural references, and homonyms, customized word lists are used. No NLP is used in these cases.

references, morphological analysis, homonyms (e.g., *their*, *there*, and *they're*), key content words, and similes[6]. For *phrasal and sentential complexity*, complex verb and noun phrases, sentences with one or more dependent clauses, and passive sentences. For *discourse*, cause-effect, compare-contrast, evidence and details, opinion, persuasion, and summary relations.

The remainder of this section describes features in the *TEA-Tool* module that use NLP to generate linguistic feedback. Providing individual evaluation descriptions for each NLP feature is beyond the scope of this paper[7], intended to focus on user study outcomes associated with *Language Muse* use (Section 4).

The specific *vocabulary* (lexical) features that use NLP methods or resources include these options[8]: *basic* and *challenge synonyms*, *complex and irregular word forms*, *variant word forms*, and *multiple word expressions*. As discussed earlier, unfamiliar vocabulary is recognized as a big contributor to text inaccessibility. The *Basic Synonym* and *Challenge Synonym* features support the vocabulary comprehension and vocabulary building aspects, respectively. To generate the greatest breadth of synonyms, the tool uses a distributional thesaurus (Lin, 1998), WordNet (Miller, 1995) and a paraphrase generation tool (Dorr and Madnani, to appear). Previous research has evaluated using these combined resources with relevant constraints to prevent too many false positives (Burstein and Pedersen, 2010). An additional slider feature allows users to adjust the number of words for which the tool will return synonyms for existing words in the text. Outputs are based on word frequency. Frequencies are determined using a standard frequency index (Breland, Jones, and Jenkins, 1994). If users want synonyms for a larger number of words across a broader frequency range that includes lower (more rare words) and higher (more common words) frequency words, then they move the slider further to the right. To retrieve synonyms for fewer and rarer words, the slider is moved to the left. For all words in the text that are within the range of word frequencies at the particular point on the slider, the tool returns synonyms. If users select *Basic Synonyms*, the tool returns all words with equivalent or higher frequencies than the word in the text. In theory, these words should be more common words that support basic comprehension. If users select *Challenge Synonyms*, then the tool returns all words with equivalent or lower frequencies than the word in the text. In this case, the teacher might want to work on vocabulary building skills to help the learner with new vocabulary. If the user selects both the *Basic Synonyms* and *Challenge Synonyms* features, then the tool will output the full list of basic (more familiar), and challenge (less familiar) synonyms for words in the text. The teacher can use these synonyms to modify the text directly, or to develop instructional activities to support word learning. The *Complex and Irregular Word Forms and Variant Word Forms* feature offers feedback related to morphological form. A morphological analyzer originally evaluated for an automated short-answer scoring system (Leacock & Chodorow, 2003) is used. This analyzer handles derivational and inflectional morphology. Feedback can be used for instructional scaffolding that includes discussion and activities related to morphological structure is an effective method to build ELLs' vocabulary. There are two features that identify words with morphological complexity, specifically, words with prefixes or suffixes: (1) *Complex and Irregular Word Forms* and (2) *Variant Word Forms*. For (1), the morphological analyzer identifies words that are morphologically complex. A rollover is available for these words. Users can place their cursor over the highlighted word, and the word stem is shown (e.g., *lost* ⇒ *stem: lose*). For (2), the system underlines words with the same stem that have different parts of speech, such as *poles* and *polar*. Teachers can build instruction related to this kind of morphological variation and teach students about variation and relationships to parts of speech.

Multiple word expressions (MWE) may include idioms (e.g., *body and soul*), phrasal verbs (e.g., *reach into*), and MWEs that are not necessarily idiomatic, but typically appear together (collocations) to express a single meaningful concept (e.g., *heart disease*). All of these MWE types may be unfamiliar terms to ELLs, and so they may interfere with content comprehension. Teachers can get feedback identifying MWEs to design relevant scaffolding for a text. To identify MWEs, two resources are used. The WordNet 3.0 compounds

[6] This new feature was not available during the pilot study.
[7] For details, see Burstein, Sabatini, Shore, Moulder, Holtzman & Pedersen (2012).
[8] These reflect the feature names in *TEA-Tool*.

list of approximately 65,000 collocational terms is used in combination with a collocation tool that was designed to identify collocations in test-taker essays (Futagi, Deane, Chodorow, & Tetreault, 2008). Some terms in the WordNet list are complementary to what is found by the collocation tool. We have found that both outputs are useful. Futagi et al.'s collocation tool identifies collocations in a text that occur in seven syntactic structures that are the most common structures for collocations in English based on The BBI Combinatory Dictionary of English (Benson, Benson, & Ilson, 1997). For instance, these include Noun of Noun (e.g., swarm of bees), and Adjective + Noun (e.g., strong tea), and Noun + Noun (e.g., house arrest). See Futagi et al. (2008) for further details.

Complex phrasal or sentential features can introduce potential difficulty in a text. A rule-based NLP module is used to identify all of these features using a shallow parser that had been previously evaluated for prepositional phrase and noun phrase detection (Leacock & Chodorow, 2003). The module to identify passive sentence construction had been previously evaluated for commercial use (Burstein, Chodorow, & Leacock, 2004). The following feedback features can be selected: *Long Prepositional Phrases*, which identifies sequences of two or more consecutive prepositional phrases (e.g., *He moved the dishes from the table to the sink in the kitchen*.); *Complex Noun Phrases*, which shows noun compounds composed of two or more nouns (e.g., emergency management agency) and noun phrases (e.g., shark-infested waters); *Passives*, which indicate passive sentence constructions (e.g., *The book was bought by the boy*.); *1+Clauses*, which highlights sentences with at least one dependent clause (e.g., *The newspaper indicated that there are no weather advisories*.); and *Complex Verbs*, which identifies verbs with multiple verbal constituents (e.g., *would have gone, will be leaving, had not eaten*).

With regard to *discourse transition features*, discourse-relevant cue words and terms are highlighted when the following discourse transitions features are identified, including: Evidence & Details, Compare-Contrast, Summary, Opinion, Persuasion, and Cause-Effect. A discourse analyzer previously evaluated for a commercial automated scoring application is used (Burstein, Kukich, Wolff, Lu, Chodorow, Braden-Harder, & Harris, 1998). The system identifies cue words and phrases in text that are being used as specific discourse (or rhetorical) contexts. For instance, "*because*" is typically associated with a cause-effect relation. However, some words need to appear in a specific syntactic construction to function as a discourse term. For instance, the word first functions as an adjective modifier and not a discourse term in a phrase, e.g., "the first piece of cake." When first is sentence-initial, as in, "First, she sliced a piece of cake," then it is more likely to be used as a discourse marker, indicating a sequence of events.

4 TPD Pilot

We report on *Language Muse* use as it was integrated into a Stanford University TPD program for in-service[9] teachers. The site agreed to integrate the application into their coursework to support coursework instruction, and instructional goals. This section describes a pilot study and outcomes with in-service teachers enrolled in the program.

4.1 Study Design

4.1.1 Site Description

Stanford University's courses are offered entirely online to teachers as part of a professional development program that awards the California State Cross-Cultural Language and Academic Development (CLAD) certificate through its California Teachers of English Learners (CTEL) certification process. By state law, all California teachers of ELLs must obtain a CLAD/CTEL or equivalent certification.

4.1.2 Teacher Participants

Responses to a background survey administered to teachers indicated a range of teaching experience from less than a year of teaching experience to as much as 37 years of teaching experience. Teachers taught across a broad range of content areas, including Art, Computers, Health, Language Arts, Math, Music, Physical Education, Science, and Social Studies, and grade levels from Kindergarten through 12th grade.

[9] This refers to teachers who have teaching credentials, and can be employed as a classroom teachers.

4.1.3 Pilot Instructional Activities[10],

After responding to the background survey, and the two pre-tests (Section 4.1.4), teachers completed the following TPD activities before moving on to post-tests (Section 4.1.4.) First, teachers read an article written by a teacher training expert on the team. The article describes best practices for developing language-based scaffolding for ELLs. The article also offers strategy descriptions as to how to use *Language Muse* to complete the lesson plan assignment (Section 4.1.4), in particular. Teachers then viewed three instructional videos that provided instruction about how to use the tool. Videos were created by a research team member, and included additional instruction about scaffolding strategies. Finally, teachers completed two practice activities with *Language Muse* which gave them an opportunity to use the different tool modules (*TEA-Tool* and lesson planning) before developing the final lesson plan assignment.

4.1.4 Measurement Instruments[11]

Teachers completed two surveys, one pre-survey, responding to questions about their professional background and school context, and a second post-survey responding to questions related to perceptions about *Language Muse* use. To evaluate teacher knowledge gains, pre- and post-test instruments were developed by the project team, and included: (a) a multiple-choice (MC) test that evaluated teachers' knowledge of linguistic structures at the Vocabulary, Sentence, and Discourse levels, and (b) a constructed- response[12] (CR) test t measured teachers' ability to identify linguistic features in a text[13] that were likely to interfere with content comprehension, and to suggest language-based instructional scaffolding to support comprehension. The pretests were administered prior to exposure to *Language Muse* (through the instructional activities (Section 4.1.3)), and the posttest after exposure. The same test was administered at pre- and post-.[14] The CR task was scored by two human raters on a 6-point scale (0 to 5, where 5=highest quality response). Inter-rater reliabilities[15] were 0.72 for Vocabulary; 0.75 for Sentences; and 0.71 for Discourse CR items. At *post-test only*, teachers developed a lesson plan using the *lesson planning* and *TEA-Tool*[16] modules in *Language Muse*. This occurred *after* teachers had completed the instructional activities included as part of *Language Muse* integration in the Stanford program. Lesson plans were evaluated by two human raters using two distinct rubrics: a) quality of *Language Skill* objectives or b) *ELL-specific Skills* objectives, i.e., unique challenges to ELLs such as, idioms or cultural references. Inter-rater reliabilities were 0.61 and 0.71 respectively. In addition, raters reviewed the linguistic feedback features that teachers had used to explore the lesson plan text, using *TEA-Tool*. The raters then examined the lesson plan and recorded the number of features explored that ended up informing the lesson plan. Inter-rater reliabilities were 0.69.

4.2 Study Results

Pre-Posttests, MC and CR. Analyses were conducted for 107 teacher participants for pre- and post-MC; 103 pre- and post-CR[17]. Paired-samples t-test showed statistically significant (p=0.02) increase in the MC Discourse score from pre-test (M =13.71, SD =2.22) to post- (M=14.20, SD =2.35; (p=0.02) increase in CR Vocabulary pre (M=2.79, SD=0.88) to post- (M=2.99, SD=0.86); in the CR Sentences score (p=0.02) from pre- (M=1.51, SD=1.23) to post- (M=1.91, SD=1.24); in the CR Total score (p=0.00) pre- (M=5.96, SD=2.35) to post- (M=6.76, SD=2.08). There were no statistically significant increases in the MC Vocabulary, Sentences, and Total scores, nor CR Discourse.

Lesson Plans. Of the 112 teachers who completed the *Lesson Plan assignment*, a significant

[10] Instructional activities are available on the *Language Muse* homepage. Teachers save all of their work in *Language Muse* so it can be viewed by course instructors and the research team, and accessed by users.

[11] For measurement instruments details, see Burstein et al, (2012).

[12] Constructed-response tasks require extended written responses.

[13] An 300-word, 8th grade Social Studies text about U.S. colonization was used.

[14] There was a lapse of approximately 8 weeks between the pre- and the post-test.

[15] Inter-rater reliabilities in this study reflect Pearson correlations.

[16] The TEA-Tool module is used to explore the linguistic features in the text; feedback features are then used to inform lesson plan development with regard to the creation of language-based scaffolding.

[17] Analyses are reported only for participants who responded to the pre- and post-.

correlation of 0.205 was found between the *Language Skills Score* and the *number of feedback features* used to inform the lesson plan.

5 Discussion and Conclusions

This paper discusses how *Language Muse*, an NLP-driven TPD application, supported K-12 teachers in understanding linguistic features in text that may be obstacles to content understanding during reading. Through the development of teachers' linguistic awareness, our original hypothesis was that teachers would become more knowledgeable about linguistic structures, and in turn, this would support them in the practice of creating lesson plans with greater coverage of text language and language objectives that would facilitate students' text and content understanding.

Study outcomes indicated that the teacher professional development package can be successfully implemented in the context of in-service, post-secondary course work. Through a study with a TPD program at Stanford University, results of the pre-post assessments administered in the study indicated at statistically-significant levels that teachers *did* improve their linguistic knowledge about vocabulary, sentences relations, and discourse relations, and that they also demonstrated and increased ability to offer language-based scaffolding strategies as evidenced by an gains pre-post total score on the CR. In the context of lesson plan development, as a secondary post-test evaluation, teachers who productively used the linguistic feedback to inform their lesson plans designed higher-quality plans (i.e., addressed language objectives that target development of new language skills), than those who did not.

The *Language Muse* TPD package is now being evaluated with nine middle-school teachers with high populations of ELLs in California, New Jersey, and Texas. After completion of the TPD, teachers will develop lesson units using *Language Muse*, and administer the lessons in their classrooms. Pre- and post-tests will be administered to students to evaluate the effectiveness of the lesson plans vis-à-vis language-based instruction.

Acknowledgments

Research presented in this paper was supported by the Institute of Education Science, U.S. Department of Education, Award No. R305A100105. Any opinions, findings, and conclusions or recommendations are those of the authors and do not necessarily reflect the IES's views. We are grateful to Steven Holtzman and Jennifer Minsky for statistical analysis support. We would like to thank Dr. Kenji Hakuta for supporting this work through his TPD program at Stanford University.

References

Adger, C. T., Snow, C., & Christian D. (2002). *What teachers need to know about language.* Washington, DC: Center for Applied Linguistics.

August, D. (2003). *Supporting the development of English literacy in English language learners: Key issues and promising practices* (Report No. 61). Baltimore, MD: Johns Hopkins University Center for Research on the Education of Students Placed at Risk.

Barzilay, Regina and Mirella Lapata (2008). 'Modeling Local Coherence: An Entity-Based Approach.' *Computational Linguistics*, 43(1): 1-34.

Beck, I. L., McKeown, M. G., & Kucan, L. (2008). *Creating robust vocabulary: Frequently asked questions and extended examples.* New York, NY: Guilford Press.

Benson, M., Benson, E., & Ilson, R. (Eds.). (1997). *The BBI Combinatory Dictionary of English: A Guide to Word Combinations.* Amsterdam & Philadelphia: John Benjamins Publishing Company.

Berninger, V., Abbot, R., Nagy, W., & Carlisle, J. (2009). Growth in phonological, orthographic, and morphological awareness in grades 1-6. *Journal of Psycholinguistic Research, 39,* 141-163.

Breland, H. Jones, R., and Jenkins, L (1994). The college board vocabulary study. Technical Report College

Burstein, J., Sabatini, J., & Shore, J. (in press). In Ruslan Mitkov (Ed.), Developing NLP Applications for Educational Problem Spaces, Oxford Handbook of Computational Linguistics. New York: Oxford University Press.

Burstein, J., Shore, J., Sabatini, J., Moulder, B., Holtzman, S., & Pedersen, T. (2012). *The Language Muse system: Linguistically focused instructional authoring* ETS RR-12-21. Princeton, NJ: ETS.

Burstein, J., and Pedersen, T. (2010). Towards Improving Synonym Options in a Text Modification Application. *University of Minnesota Supercomputing Institute Research Report Series*, UMSI 2010/165, November 2010.

Burstein, J., Chodorow, M., and Leacock, C. (2004). Automated Essay Evaluation: The Criterion Online Service, AI Magazine, 25(3), 27-36.

Burstein, J., Kukich, K., Wolff, S., Lu, C., Chodorow,

M., Braden-Harder, L., and Harris, M. D. (1998). *Automated Scoring Using A Hybrid Feature Identification Technique*. In the Proceedings of the Annual Meeting of the Association of Computational Linguistics, August, 1998. Montreal, Canada.

Calderón, M. (2007). *Teaching reading to English language learners, grades 6-12: A framework for improving achievement in the content areas.* Thousand Oaks, CA: Corwin Press.

Calderón, M., August, D., Slavin, R., Cheung, A., Durán, D., & Madden, N. (2005). Bringing words to life in classrooms with English language learners. In A. Hiebert & M. Kamil (Eds.), *Research and development on vocabulary*. Mahwah, NJ: Lawrence Erlbaum Associates.

Carlo, M. S., August, D., McLaughlin, B., Snow, C. E., Dressler, C., Lippman, D. N., & White, C. E. (2004). Closing the gap: Addressing the vocabulary needs of English language learners in bilingual and mainstream classrooms. *Reading Research Quarterly, 39*, 188-215.

Coleman, D., & Pimentel, S. (2011a). *Publishers' criteria for the Common Core State Standards in English Language Arts and Literacy, grades 3-12.* Washington, DC: National Governors Association Center for Best Practices and Council of Chief State School Officers.

Collins-Thompson, Kevyn and Jamie Callan (2004). 'A Language Modeling Approach to Predicting Reading Difficulty.' In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics.* Boston, MA: Association for Computational Linguistics, 193-200.

Darling-Hammond, L. (2000). Teacher quality and student achievement: A review of state policy evidence. *Education Policy Analysis Archives, 8.*

Flesch, R.. (1948). A new readability yardstick. *Journal of Applied Psychology, 32*, 221-233.

Flinspach, S. L., Scott, J. A., Samway, K. D., & Miller, T. (2008, March). *Developing cognate awareness to enhance literacy: Importante y necesario.* Paper presented at the Annual Meeting of the American Educational Research Association, New York, NY..

Francis, D., August, D. Goldenberg, C., & Shanahan, T. (2004). *Developing literacy skills in English language learners: Key issues and promising practices.* Retrieved June 11, 2007, from: www.cal.org/natl-lit-panel/reports/Executive_Summary.pdf

Futagi, Y., Deane, P., Chodorow, M., & Tetreault, J. (2008). A Computational Approach to Detecting Collocation Errors in the Writing of Non-native Speakers of English, *Computer Assisted Language Learning*, Vol. 21, pp. 353–367.

Gándara, P., Maxwell-Jolly, J., & Driscoll, A. (2005). *Listening to teachers of English language learners: A survey of California teachers' challenges, experiences, and professional development needs.* Sacramento, CA: The Regents of the University of California. Retrieved from http://www.cftl.org/documents/2005/listeningforweb.pdf.

Goldenberg, C. (2008). Teaching English language learners: What the research does—and does not—say. *American Educator, 32*, 8-21.

Goldman, S. R., & Rakestraw Jr., J. A. (2000). Structural aspects of constructing meaning from text. In M. L. Kamil, P. B. Mosenthal, P. D. Pearson, & R. Barr (Eds.), *Handbook of reading research* (Vol. III, pp. 311-335). Mahwah, NJ: Lawrence Erlbaum Associates.

Graesser, Arthur C., Danielle S. McNamara, and Jonna M. Kulikowich (2011). 'Coh-Metrix: Providing Multilevel Analyses of Text Characteristics.' *Educational Researcher,* 40(5): 223-234.

Green, C., Foote, M., Walker, C., & Shuman, C. (2010). From questions to answers: Education faculty members learn about English learners. In S. Szabo, M. B. Sampson, M. M. Foote, & F. Falk-Ross (Eds.), *Mentoring literacy professionals: Continuing the spirit of CRA/ALER after 50 years* (pp. 113-125). Commerce, TX: Texas A&M University Press.

Heilman, Michael, Lee Zhao, Juan Pinto, and Maxine Eskenazi (2008). 'Retrieval of Reading Materials for Vocabulary and Reading Practice.' In *Proceedings of the Third Workshop on Innovative Use of NLP for Building Educational Applications.* Columbus, OH: Association for Computational Linguistics, 80-88.

Kieffer, M. J. & Lesaux, N. K. (2008). The role of derivational morphology in the reading comprehension of Spanish-speaking English language learners. *Reading and Writing, 21*, 783-804.

Kintsch, W. (1998). *Comprehension: A paradigm for comprehension.* Cambridge, UK: Cambridge University Press.

Leacock, C. & Chodorow, M. (2003). C-rater: Scoring of Short-Answer Questions. *Computers and the Humanities*, Vol. 37, pp. 389–405.

Lesaux, N. K., Kieffer, M. J., Faller, S. E., & Kelley, J. G. (2010). The effectiveness and ease of implementation of an academic vocabulary intervention for linguistically diverse students in urban middle schools. *Reading Research Quarterly, 45*, 196-228.

Lin, Dekang (1998). 'Automatic Retrieval and Clustering of Similar Words.' In *"Proceedings of the 17th International Conference on Computational Linguistics and the 36th Annual Meeting of the Association for Computational Linguistics.* Montreal, Canada: 768-774.

Madnani, Nitin and Bonnie J. Dorr (in press). 'Generating Targeted Paraphrases for Improved Translation.'

ACM Transactions on Intelligent Language Muses and Technology: Special Issue on Paraphrasing.

Marcu, Daniel (1999). 'Discourse Trees Are Good Indicators of Importance in Text. In *Advances in Automatic Text Summarization,* eds. Inderjeet Mani and Mark T. Maybury. Cambridge, MA: MIT Press, 123-136.

Meurers, W. Detmar, Ramon Ziai, Luiz Amaral, Adriane Boyd, Aleksandar Dimitrov, Vanessa Metcalf, and Niels Ott (2010). 'Enhancing Authentic Web Pages for Language Learners.' In *Proceedings of the NAACL HLT 2010 Fifth International Workshop on Innovative Use of NLP for Building Educational Applications,* eds. Joel Tetreault, Jill Burstein, and Claudia Leacock. Los Angeles, CA: Association for Computational Linguistics, 10-18.

Meyer, B. J. F. (2003). Text coherence and readability. *Topics in Language Disorders, 23,* 204-221.

Mihalcea, Rada, Ravi Sinha, and Diana McCarthy (2010). 'SemEval-2010 Task 2: Cross-Lingual Lexical Substitution.' In *Proceedings of SemEval-2010: Fifth International Workshop on Semantic Evaluations.* Uppsala, Sweden: Association for Computational Linguistics, 9-14.

Miller, George A. (1990). 'An On-line Lexical Database.' *International Journal of Lexicography* 3(4): 235-312.

Miltsakaki, Eleni (2009). 'Matching Readers' Preferences and Reading Skills with Appropriate Web Texts.' In *Proceedings of the European Association for Computational Linguistics.* Athens, Greece: Association for Computational Linguistics, 49-52.

Nagy, W., Beringer, V., & Abbott, R. (2006). Contributions of morphology beyond phonology to literacy outcomes of upper elementary and middle school students. *Journal of Educational Psychology, 98,* 134-147.

National Clearinghouse for English Language Acquisition (2011). *The growing numbers of English learner students.* Washington, DC: Author. Retrieved from http://www.ncela.gwu.edu/files/uploads/9/growingLEP_0809.pdf.

National Governors Association Center for Best Practices and Council of Chief State School Officers (2010). *Common Core State Standards for English language Arts & Literacy in History/Social Studies, Science, and Technical Subjects. Appendix A: Research supporting key elements of the Standards.* Washington, DC: Author.

Nelson, Jessica, Charles Perfetti, David Liben, and Meredith Liben (2012). *Measures of Text Difficulty: Testing Their Predictive Value for Grade Levels and Student Performance.* Washington, DC: The Council of Chief State School Officers. Retrieved from http://www.ccsso.org/Documents/2012/Measures%20ofText%20Difficulty_final.2012.pdf.

Pappamihiel, N. E., Lake, V., & Rice, D. (2005). Adapting a Social Studies lesson to include English language learners. *Social Studies and the Young Learner, 17,* 4-7.

Peske, H. G., & Haycock, K. (2006). *Teaching inequality: How poor and minority students are shortchanged on teacher quality.* Washington, DC: The Education Trust. Retrieved from http://www.edtrust.org/sites/edtrust.org/files/publications/files/TQReportJune2006.pdf.

Pitler, Emily and Ani Nenkova (2008). 'Revisiting Readability: A Unified Framework for Predicting Text Quality.' In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing.* Honolulu, HI: Association for Computational Linguistics, 186-195.

Proctor, C. P., Dalton, D., Uccelli, P., Biancarosa, G., Mo, E., Snow, C. E., & Neugebauer, S. (2011). Improving comprehension online (ICON): Effects of deep vocabulary instruction with bilingual and monolingual fifth graders. *Reading and Writing: An Interdisciplinary Journal, 24,* 517-544.

Proctor, C. P., Dalton, B., & Grisham, D. (2007). Scaffolding English language learners and struggling readers in a multimedia hypertext environment with embedded strategy instruction and vocabulary support. *Journal of Literacy Research, 39,* 71-93.

Rivera, M. O., Moughamian, A. C., Lesaux, N. K., & Francis, D. J. (2008). *Language and reading interventions for English language learners and English language learners with disabilities.* Portsmouth, NJ: Research Corporation, Center on Instruction.

Rutherford William E. and Michael Sharwood Smith (1985). 'Consciousness-Raising and Universal Grammar.' *Applied Linguistics* 6(3): 274-282.

Schwarm, Sarah E. and Mari Ostendorf (2005). 'Reading Level Assessment Using Support Vector Machines and Statistical Language Models.' In *Proceedings of the Annual Meeting of the Association for Computational Linguistics.* Ann Arbor, MI: Association for Computational Linguistics, 523-530.

Schleppegrell, M. J. (2007). The linguistic challenges of mathematics teaching and learning: A research review. *Reading and Writing Quarterly, 23,* 139-159.

Schleppegrell, M. J., & de Oliveira, L. C. (2006). An integrated language and content approach for history teachers**.** *Journal of English for Academic Purposes, 5,* 254-268.

SDL. (n.d.). Automated translation. Retrieved from http://www.sdl.com/en/languagetechnology/products/automated-translation/

Walqui, A., & Heritage, M. (2012, January). *Instruction for diverse groups of ELLs.* Paper presented at the Understanding Language Conference, Stanford, CA.