Alignment of spoken narratives for automated neuropsychological assessment

Emily T. Prud'hommeaux, Brian Roark

Center for Spoken Language Understanding
Oregon Health and Science University
Portland, Oregon, USA
emilypx,roarkbr@gmail.com

***Abstract*—Narrative recall tasks are commonly included in neurological examinations, as deficits in narrative memory are associated with disorders such as Alzheimer's dementia. We explore methods for automatically scoring narrative retellings via alignment to a source narrative. Standard alignment methods, designed for large bilingual corpora for machine translation, yield high alignment error rates (AER) on our small monolingual corpora. We present modifications to these methods that obtain a decrease in AER, an increase in scoring accuracy, and diagnostic classification performance comparable to that of manual methods, thus demonstrating the utility of these techniques for this task and other tasks relying on monolingual alignments.**

I. Introduction

In this paper, we investigate the automation of the existing manual method for narrative recall scoring in the context of diagnosing Mild Cognitive Impairment (MCI), the earliest observable precursor to dementia. We focus on the Logical Memory Scale narrative recall subtest of the Wechsler Memory Scale [1], a widely used instrument for evaluating memory function in adults. In this task, the subject listens to a brief narrative and then verbally retells the narrative to the examiner. The examiner then scores the retelling by counting the number of recalled *story elements*, each of which typically corresponds to a short phrase in the narrative. The final score is simply the total number of items recalled by the subject. We propose to automatically determine which story elements were recalled by establishing an alignment between substrings in the retelling and substrings in the original narrative.

Although alignment methods have been widely explored for a range of NLP tasks, aligning retellings for clinical assessment is a novel task with unique obstacles. Existing word and phrase alignment-learning algorithms are typically developed to train on bilingual parallel corpora, a very different type of data from the small monolingual parallel corpora of narrative retellings analyzed here. In previous work, we showed that strong automatic scoring and diagnostic classification results could be obtained using an off-the-shelf word alignment package with no customization [2]. We now compare the output of this standard alignment approach with that achieved using two novel methods designed to take advantage of certain features of our corpus of retellings, and we evaluate whether these alignments provide sufficient accuracy for automated scoring. We then demonstrate the utility of these scores for performing diagnostic classification of MCI.

We show that, despite somewhat high alignment error rates, all of the automatically-generated alignments, in particular the alignments generated with our new alignment strategies, are acceptable substitutes for manual alignments for scoring purposes. In addition, scores derived from automatic alignments yield diagnostic classification comparable to that achieved using the manual scoring procedure.

The robustness of scoring and classification to poor alignment, however, should not obscure the need for further improvements of alignment in this domain in order to adequately measure common phenomena such as long-distance reordering, omission, embellishment, and repetition. Given that these phenomena are frequently observed in the retellings of impaired subjects, they are likely to provide significant discriminative power. The dual message of this work seems to be that scoring of narrative recall for diagnostic purposes is a very promising task for automation, but it is relatively poorly served by existing alignment algorithms. The simple but intuitive changes described here demonstrate that tailoring alignment for this task can yield improvements in both monolingual alignment itself as well as automatic scoring and diagnostic classification.

II. Background

Many neuropsychological protocols include a narrative recall or narrative generation task [3], [1], [4]. Poor performance on such tasks has been associated with of a number of cognitive and developmental problems such as language impairment [5], [6], autism [7], and dementia [8].

The earliest observable precursor of dementia, Mild Cognitive Impairment (MCI), is characterized by cognitive and memory symptoms that are clinically significant but do not necessarily interfere with daily living activities [9]. These symptoms are often difficult to detect using the common screening tests for dementia. As a result, MCI frequently goes undetected, delaying the introduction of appropriate treatment and remediation.

Most previous work in applying NLP-based analysis to spoken language and narratives for diagnostic purposes has focused not on evaluating coherence but rather on using language samples as a data source from which to extract speech and language features, such as pause duration or language complexity [10], [11]. Such features can be used to evaluate

Anna / Thompson / of South / Boston / employed / as a cook / in a school / cafeteria / reported / at the police / station / that she had been held up / on State Street / the night before / and robbed of / fifty-six dollars. / She had four / small children / the rent was due / and they hadn't eaten / for two days. / The police / touched by the woman's story / took up a collection / for her.

Fig. 1. Text of Wechsler narrative, segmented into 25 story elements

Ann Taylor worked in Boston as a cook. And she was robbed of sixty-seven dollars. Is that right? And she had four children and reported at the some kind of station. The fellow was sympathetic and made a collection for her so that she can feed the children.

Fig. 2. Sample retelling of the Wechsler narrative.

language development [12], [13], the validity of diagnostic instruments [14], and the presence of MCI [15] and language disorders [16].

Alignments of parallel texts are commonly used for NLP tasks such as paraphrase extraction [17], [18], word-sense disambiguation [19], and bilingual lexicon induction [20], but our work on leveraging alignments to automate scoring of a neuropsychological examination and to provide diagnostic information is the first of its kind to our knowledge.

III. Data

A. Subjects

The data used in this study was collected as part of a longitudinal study on brain aging at the Layton Aging and Alzheimer's Disease Center at the Oregon Health and Science University (OHSU). From the large number of participants in this study, we selected 72 subjects with MCI and 52 healthy seniors roughly matched for age and years of education.

Following [21], we defined our MCI and non-MCI groups according to the Clinical Dementia Rating (CDR) [22]. A diagnosis of MCI was defined as a CDR of 0.5, while the absence of MCI was defined as a CDR of zero. The CDR is calculated from a set of scores in key areas of cognitive function generated from clinical assessment via interview and the Neurobehavioral Cognitive Status Examination. Since we are investigating the utility of different methods of deriving information from a particular neuropsychological test, it is important to define MCI in ways that provide an independent unconfounded reference objective for evaluation. The CDR has been shown to have high expert inter-annotator reliability [22] and, importantly, is assigned independently of the neuropsychological test that we investigate in this paper. We refer readers to the above cited papers for a full definition.

B. Wechsler Logical Memory Test

The Wechsler Logical Memory task is one of several subtests of the Wechsler Memory Scale [1], a diagnostic instrument commonly used in assessment of memory and cognition in the adult population. In this test, the subject listens to a brief story, shown in Figure 1. The subject must then retell the story twice: once immediately after hearing it (Logical Memory I, LM-I) and again after a 30-minute delay (Logical Memory II, LM-II). The text of the narrative is divided into 25 *story elements*. The slashes in Figure 1 denote the boundaries between story elements. As the subject is speaking, the examiner notes which story elements were used. The final score reported, or summary score, is simply the raw number of story elements used. Story elements do not need to be produced in order or word-for-word. There are specific guidelines for each element that describe the permissible substitutions; for instance, the first story element, *Anna*, can be replaced in the retelling with *Annie* or *Ann*; the 16th story element, *fifty-six dollars*, can be replaced with any number of dollars between fifty and sixty.

Figure 2 shows an example LM-I retelling. The elements used in this retelling are *Anna, employed, Boston, as a cook, was robbed of, she had four, small children, reported, station, touched by the woman's story, took up a collection,* and *for her*, yielding a summary score of 12 according to published guidelines for manual scoring.

C. Training and testing data

Audio recordings of the Wechsler Logical Memory immediate and delayed retellings for the 124 subjects were transcribed at the word level, then tokenized and downcased. Partial words, punctuation, and pause-fillers were excluded from the transcriptions used for this study.

Most word alignment algorithms designed for machine translation take as input a parallel corpus aligned by sentence, in which a sentence on one side of the corpus is a translation of the corresponding sentence on the other side of the corpus. Thus, in order for us to be able to use these alignment algorithms to align retellings to the original narrative, we need to create parallel corpora consisting of retellings on one side and the original narrative on the other.

Due to the nature of these retellings, which contain omissions, reorderings, and embellishments, any part of the multi-sentence narrative can occur anywhere within the retelling; thus, the best coarse alignment would be the full original narrative to the full retelling. We therefore start by considering each entire retelling to be a single sentence corresponding to the single sentence of the entire original narrative. Recall that we currently have 72 subjects with MCI and 52 subjects without, for a total of 124 subjects. This means that for each subject, there are only 246 retellings (2 retellings for each of the remaining 123 subjects) available as training data. A word alignment algorithm is unlikely to learn many meaningful translation relationships from such impoverished data. In order to supplement the training data, we perform additional pre-processing to supply the alignment algorithms with more information from which to learn their models. In addition to full retelling-to-narrative parallel alignments, we provide phrase-to-phrase parallel alignments for phrases in the retelling that are associated with specific story elements, as shown in Figure 3. Each subject's retellings are aligned using a separate model built on these two kinds of high-level alignments derived from the retellings of the other 123 subjects.

Original Narrative	**Retelling**
anna	ann
employed	worked
boston	in boston
as a cook	as a cook
and robbed of	she was robbed of
she had four	she had four
reported	reported
station	station
small children	children
touched by the woman's story	was sympathetic
took up a collection	and made a collection
for her	for her

Fig. 3. Alignment of phrases to story elements.

[A anna$_1$] [B thompson$_2$] [C of$_3$ south$_4$] [D boston$_5$] [E employed$_6$] [F as$_7$ a$_8$ cook$_9$] [G in$_{10}$ a$_{11}$ school$_{12}$] [H cafeteria$_{13}$] [I reported$_{14}$] [J at$_{15}$ the$_{16}$ police$_{17}$] [K station$_{18}$] [L that$_{19}$ she$_{20}$ had$_{21}$ been$_{22}$ held$_{23}$ up$_{24}$] [M on$_{25}$ state$_{26}$ street$_{27}$] [N the$_{28}$ night$_{29}$ before$_{30}$] [O and$_{31}$ robbed$_{32}$ of$_{33}$] [P fifty-six$_{34}$ dollars$_{35}$] [Q she$_{36}$ had$_{37}$ four$_{38}$] [R small$_{39}$ children$_{40}$] [S the$_{41}$ rent$_{42}$ was$_{43}$ due$_{44}$] [T and$_{45}$ they$_{46}$ had$_{47}$ n't$_{48}$ eaten$_{49}$] [U for$_{50}$ two$_{51}$ days$_{52}$] [V the$_{53}$ police$_{54}$] [W touched$_{55}$ by$_{56}$ the$_{57}$ woman's$_{58}$ story$_{59}$] [X took$_{60}$ up$_{61}$ a$_{62}$ collection$_{63}$] [Y for$_{64}$ her$_{65}$]

Fig. 4. Text of Wechsler Logical Memory narrative with story-element labeled bracketing and word IDs.

Given the clear boundaries between story elements provided in the scoring guidelines, we know exactly which words in the original narrative together make up each story element. Figure 4 shows the original narrative labeled with element IDs ($A-Y$) and word IDs ($1-65$). Story element F, for instance, is composed of words 7, 8, and 9, *as a cook*. These mappings from words to elements will be used for automatic scoring, as described in Section VI.

In the next section, we present the details of the new alignment approaches we propose. In the subsequent sections, we present experiments conducted to examine the accuracy of several well-known methods and our novel methods for learning word alignment models for this task; the utility of these alignments for automated scoring; and the utility of automated scores for diagnostic classification.

IV. Methods

For this task of aligning a small monolingual parallel corpus, we explore two alignment strategies that exploit the fact that our corpus is monolingual. Specifically, we know that the identity alignment (i.e., when a word in the source is aligned with that same word in the target) – something that is typically ignored in standard word alignment algorithms – will play an important role in alignment of this data. Both of our approaches use an IBM Model 1 mixture model to derive initial translation probabilities for an HMM. We perform Viterbi decoding to extract the alignments using the translation and distortion probabilities estimated using the forward-backward algorithm in the HMM.

In our first approach, we include in our training data explicit parallel alignments of word identities to promote the likelihood that a word on one side of the corpus will be aligned to an instance of that same word on the other side of the corpus. Every word that appears in the original narrative and in every retelling other than the one being tested is included once in this additional training data. We train both the IBM Model 1 mixture model and the HMM with this additional data. The advantage of this method is that it can be incorporated trivially into existing word alignment frameworks.

In our second approach, we begin, as in the first approach, with the standard IBM Model 1 mixture model for estimating initial translation probabilities, but we do not train on word identities. Rather, we train only on the narrative- and phrase-level alignments, but we include a prior probability, λ, that an alignment is between two orthographically identical words. The value of this prior probability was estimated via relative frequency estimation (identical word alignments divided by total number of alignments) on a small held-out set of manual alignments. Given a source word, f_j, that is aligned with an target word, e_{a_j}, let

$$\omega(a_j, j) = \begin{cases} \lambda & \text{if } f_j \text{ and } e_{a_j} \text{are identical} \\ 1-\lambda & \text{otherwise} \end{cases}$$

In our held-out data, the value of λ was determined to be 0.69. The ω probabilities are used as a prior on translation probability during the expectation phase of parameter estimation, rendering the standard equation as the following:

$$Pr(f_1^J|e_1^I) = \prod_{j=1}^{J}\sum_{i=1}^{I}[p(i|j,I)\cdot p(f_j|e_i)\cdot\omega(i,j)]$$

The translation probabilities estimated in this first step are then used as the initial translation probabilities in the second step, an HMM.

As is typical in the HMM approach to word alignment, transition (distortion) probabilities are modeled as "jump widths" from position to position on the target side of the alignment. Since our data is likely to contain very large jumps in alignment (e.g., when a speaker remembers only that the police gave Anna some money), we model all jumps from 0 to the largest possible jump width individually rather than collapsing jump widths greater than 5 into a single parameter, as is sometimes done in bilingual alignment.

In both approaches, we build models independently in both directions (retelling to original and original to retelling), and extract alignments in both directions using Viterbi decoding. We then take the intersection of these two sets of alignments to produce the final set of alignments for each retelling.

In order to select the best translation probabilities to feed into the HMM step, we first compare the initial alignments generated using these two methods with alignments produced by the same procedure but without the word identity probability, ω, or the additional word identity training data. We then compare the output of our HMM trained on these initial probabilities with the output of two commonly used and publicly available alignment systems: 1) the Berkeley aligner [23] with default parameterizations (IBM Model 1 feeding

ann(1) : anna(1)
taylor(2) : thompson(2)
worked(3) : employed(6)
boston(5) : boston(5)
as(6) : as(7)
a(7) : a(8)
cook(8) : cook(9)
and(9) : and(31)
robbed(12) : robbed(32)
of(13) : of(33)
dollars(15) : fifty-six(34)
right(18) : due(44)
and(19) : and(45)
she(20) : she(36)
had(21) : had(37)
four(22) : four(38)
children(23) : small(39)
reported(25) : reported(14)
at(26) : at(15)
the(27) : the(16)
some(28) : on(25)
kind(29) : night(29)
station(31) : station(18)
sympathetic(35) : story(59)
made(37) : took(60)
a(38) : a(62)
collection(39) : collection(63)
for(40) : for(64)
her(41) : her(65)
that(43) : that(19)
she(44) : she(20)
can(45) : been(22)
feed(46) : eaten(49)

Fig. 5. Word-alignment generated by the Berkeley aligner with retelling words italicized.

into an HMM with joint training and posterior decoding); and 2) Giza++ [24] with default parameterizations. We test these two existing aligners both with the phrase- and sentence-level training data alone and with the full set of training data that includes word identity alignments.

V. Experiment 1: Word Alignment

Each word alignment algorithm yielded for each retelling a sequence of pairs of words, in which one word from the original narrative is paired with one word from the retelling. Figure 5 shows the results of aligning the retelling presented in Figure 2 using the Berkeley aligner. The alignments for the subset of subjects with available manual gold-standard word-level alignments were evaluated against the alignments produced. In the following tables we report the precision, recall, F1 measure, and word alignment error rate (AER) [24] for the various alignment algorithms.

Table I compares the results of standard IBM Model 1 alignment with that of our two strategies to encourage alignment of identical words: the model trained on explicit word identity training data and the model using the lexical identity weight, ω. We report the alignment results for the forward direction (aligning the retelling to the original: f-AER), the backward direction (aligning the original to the retelling: b-AER), and the intersection of the two directions (i-AER). Including additional training data consisting of word identity mappings does decrease word AER a great deal but primarily in the forward (retelling to original) direction. Weighting the translation probabilities during estimation improves both the forward and backwards alignments over the baseline, resulting in a large decrease in word alignment error rate of the intersection of the two directions.

Table II shows the performance of our HMM aligner under the three IBM Model 1 initialization conditions: 1) no word identity coercion (CSLU aligner); 2) word identity training data (CSLU+ID train); and 3) word identity weighting (CSLU+ID weight). As expected, the lexical identity weighting produces the lowest AER of the three CSLU aligners. In addition, Table II shows the performance of the Berkeley aligner and Giza trained with and without the word identity alignments. We observe that both the Berkeley aligner and Giza benefit from training on explicit identity alignments.

Finally, we take the intersection of outputs of the best CSLU aligner and the best existing aligner, which in this case is the Berkeley aligner trained on data including explicit word identities. This produces the lowest AER and highest F1 measure of all of the approaches.

The slight edge the Berkeley aligner holds over our aligner may be due to its use of joint training and posterior decoding, which the authors report are responsible for large improvements in AER over the independent training and Viterbi decoding used in the CSLU aligner [23]. We note, however, that the CSLU aligner occupies a different operating point on the precision/recall tradeoff continuum from that of either the Berkeley aligner or Giza++. This indicates that these two aligners are overgenerating alignments, which will have an impact on scoring accuracy.

There are a number of unexpected incorrect alignments in the sample Berkeley alignment shown in Figure 5. Some of these misalignments seem to have no obvious explanation, such as *kind(29)*–night(29). Others, such as *dollars(15)*–fifty-six(34) and *children(23)*–small(39), intuitively make more sense, given the way that our training data is structured. The aligner is presented with many alignment pairs in which the original narrative phrase is *small children* and the retelling phrase contains the word *children* but not necessarily the word *small*, which is frequently excluded in the retellings. It is thus not surprising that the aligner learns that there is some non-zero probability that it is the original word *small*, rather than the original word *children*, that aligns with the retelling word *children*. We have informally observed that this happens less frequently in the alignments produced using word identity coercion, and we expect that the preference for alignment of identical words could account for the superior AER observed

aligner	f-AER	b-AER	i-AER
IBM1	66.2	67.2	64.5
IBM1+ID train	55.4	64.5	51.9
IBM1+ID weight	54.8	56.8	47.2

TABLE I
IBM Model 1 initial word alignment results.

aligner	prec.	rec.	F1	AER
CSLU	77.3	46.0	57.7	42.3
CSLU+ID train	79.9	52.9	63.6	36.4
CSLU+ID weight	79.2	59.6	68.0	32.0
Berkeley	62.3	69.9	65.9	34.1
Berkeley+ID train	63.8	73.8	68.4	31.6
Giza	46.7	71.9	56.6	43.4
Giza+ID train	47.0	72.8	57.2	42.8
Best CSLU∩Best Berk.	89.1	57.0	69.5	30.5

TABLE II
Overall word alignment results.

anna(1) : A	fifty-six(34) : P	station(18) : K
thompson(2) : B	due(44) : S	story(59) : W
employed(6) : E	four(38) : Q	took(60) : X
boston(5) : D	small(39) : R	collection(63) : X
cook(9) : F	reported(14) : I	for(64) : Y
robbed(32) : O	night(29) : N	her(65) : Y
		eaten(49) : T

Fig. 6. Words from original narrative from the alignment in Figure 5, excluding function words, with corresponding story element IDs.

using those techniques.

VI. Experiment 2: Automated Scoring

A. Method

Alignments between the words of the original narrative and those used in each retelling were produced using the four best performing word alignment algorithms: 1) the CSLU aligner with word identity training; 2) the CSLU aligner with word identity weighting; 3) the Berkeley aligner with word identity training; and 4) the intersection of the latter two. For each retelling, we determined the identities of the story elements used from the word alignments using the following algorithm. Recall that every word in the original narrative is associated with a particular story element ID, as shown in Figure 4. For each word in the original narrative, if the word is aligned to a word in the retelling, the story element ID that it is associated with is considered to be recalled. In Figure 6, we see the list of story elements identified from the word alignments in Figure 5.

Note that with this algorithm to convert from words to story elements, it is possible to generate a story element from a single word-alignment pair, even for function words such as *the* and *of* which are not reliable indicators that the content associated with that story element has been recalled. For this reason, we discarded any word-alignment pair in which a word in the retelling mapped to a function word in the original narrative. The only eligible function words from the original narrative were the final two words, *for her*, which are both function words but together make a single story element. Examples of these restrictions can be observed in Figure 6.

B. Evaluation and Results

The story element identities induced from the automatically derived word alignments for all 124 subjects were evaluated against the manually derived per-element scores. In Table III we report the precision, recall, and F1 measure for the four best alignment strategies. In addition, we evaluate a baseline scoring strategy relying on exact match of the verbatim story elements using `grep`. The CSLU aligner yields the highest F1 measure and a balance between recall and precision, while the Berkeley aligner achieves a lower F1 measure and favors recall at the expense of precision. These results are significant at $p < 0.0001$ using the stratified shuffling test [25].

In Table IV, we report scoring accuracy in terms the mean difference, mean absolute difference, and the median difference between the summary score generated from the alignments and the reference summary score. We see again that the CSLU aligner with word identity weighting yields more accurate scoring than the Berkeley aligner, which very consistently overestimates the number elements recalled.

aligner	prec.	recall	F1
verbatim	92.0	47.4	62.6
CSLU+ID train	81.6	83.6	82.6
CSLU+ID weights	82.2	87.2	84.6
Berkeley+ID train	66.8	96.9	79.1
Best CSLU ∩ Best Berk.	88.3	75.9	81.6

TABLE III
Scoring accuracy results.

Although the word AER for all alignment algorithms was quite high compared to the AER typically reported for large data sets (e.g., 4.9% AER on the Hansards English-French task) [23], the scoring accuracy is remarkably robust, demonstrating the utility of applying these algorithms to this task.

VII. Experiment 3: Classification

A. Method

In order to demonstrate the utility of automatic scoring, we now use these scores as features within a support vector machine (SVM) for classification of subjects into the two diagnostic groups: individuals with and without MCI. Previous work [21] has shown that the Wechsler Logical Memory summary scores provide significant classification power.

From the scores for all 124 subjects extracted as described in Section VI, we generated two summary scores per subject, one each for LM-I and LM-II, ranging from 0 (when no items were recalled) to 25 (when all items were recalled). LibSVM [26], the SVM class available in the WEKA data mining package API [27], was then used to train support vector machine classifiers.

B. Evaluation

To evaluate the SVMs, we chose to use a leave-pair-out validation scheme [28], [29] in which every possible pair consisting of a negative example and a positive example is tested against an SVM classifier trained on the remaining 122 subjects. The resulting pairs of scores can be used to calculate the area under the receiver operating characteristic (ROC) curve [30], a commonly used estimate of classification accuracy. The receiver operating characteristic curve is a plot of the false positive rate of a classifier against its true positive rate. The area under this curve (AUC) can range from 0.5, when the classifier performs no better or worse than chance, to 1.0, when the classifier has perfect accuracy.

Table V shows that the classification accuracy of the SVMs trained on automatically derived scores was strong and comparable to that of an SVM trained on manually derived scores. We note, however, that superior accuracy in automatic scoring

aligner	mean	mean abs.	median
Berkeley	-5.32	5.41	-5
CSLU	-0.71	1.69	-1
Intersection	1.65	2.47	1

TABLE IV
Mean and median scoring differences.

(see Tables III and IV) does not necessarily lead to better classification accuracy.

alignment	AUC	s.d.
Manual	0.822	0.036
Berkeley+ID train	0.795	0.039
CSLU+ID weight	0.784	0.040
Berk+CSLU	0.767	0.042

TABLE V
CLASSIFICATION ACCURACY RESULTS.

VIII. DISCUSSION AND FUTURE DIRECTIONS

In this paper, we presented a novel application of automatic word alignment: the alignment of narrative retellings for the purposes of scoring a narrative memory task widely used in clinical neurological assessment. We also offer new alignment strategies designed specifically for the alignment of small monolingual parallel corpora that yield better alignment and scoring performance than the most commonly used alignment algorithms.

The alignments thus produced can also be a source of additional language and memory features. The retellings analyzed here contain numerous repetitions, reorderings, omissions (when the subject fails to recall one of the story elements), and insertions (when the subject produces off-topic phrases or embellishments). Phenomena such as these, which are likely to provide significant diagnostic information, can be automatically induced from word alignments. In future work, we will investigate methods for extracting these features and using them to improve diagnostic classification. We also plan to adapt our techniques to other narrative tasks.

Although the results presented here are promising, there is considerable room for improvement in the word alignment itself. We plan to improve spoken retelling alignments both by adding new measures that capture orthographic and semantic similarity and by developing different methods of training the alignment models, including discriminative approaches that will allow us to easily incorporate these new features.

ACKNOWLEDGMENT

This research was supported in part by NSF BCS-0826654. Thanks to Meg Mitchell, Margit Bowler, and Jessica Ferguson for their assistance in transcription and data preparation.

REFERENCES

[1] D. Wechsler, *Wechsler Memory Scale - Third Edition Manual*. The Psychological Corporation, 1997.

[2] E. T. Prud'hommeaux and B. Roark, "Extraction of narrative recall patterns for neuropsychological assessment," in *Proceedings of Interspeech*, 2011.

[3] M. Korkman, U. Kirk, and S. Kemp, *NEPSY: A developmental neuropsychological assessment*. The Psychological Corporation, 1998.

[4] C. Lord, M. Rutter, P. DiLavore, and S. Risi, *Autism Diagnostic Observation Schedule (ADOS)*. Western Psychological Services, 2002.

[5] K. Dodwell and E. L. Bavin, "Children with specific language impair ment: an investigation of their narratives and memory," *International Journal of Language and Communication Disorders*, vol. 43, no. 2, pp. 201–218, 2008.

[6] N. Botting, "Narrative as a tool for the assessment of linguistic and pragmatic impairments," *Child Language Teaching and Therapy*, vol. 18, no. 1, 2002.

[7] H. Tager-Flusberg, "Once upon a ribbit: Stories narrated by autistic children," *British journal of developmental psychology*, vol. 13, no. 1, pp. 45–59, 1995.

[8] G. Gomez and D. A. White, "Using verbal fluency to detect very mild dementia of the alzheimer type," *Archives of Clinical Neuropsychology*, vol. 21, pp. 771–775, 2006.

[9] K. Ritchie and J. Touchon, "Mild cognitive impairment: Conceptual basis and current nosological status," *Lancet*, vol. 355, pp. 225–228, 2000.

[10] K. Lyons, S. Kemper, E. LaBarge, F. Ferraro, D. Balota, and M. Storandt, "Oral language and Alzheimer's disease: A reduction in syntactic complexity," *Aging and Cognition*, vol. 1, no. 4, pp. 271–281, 1994.

[11] S. Singh, R. Bucks, and J. Cuerden, "Evaluation of an objective technique for analysing temporal variables in DAT spontaneous speech," *Aphasiology*, vol. 15, no. 6, pp. 571–584, 2001.

[12] K. Sagae, A. Lavie, and B. MacWhinney, "Automatic measurement of syntactic development in child language," in *Proceedings of ACL*, 2005, pp. 197–204.

[13] K. Gabani, M. Sherman, T. Solorio, and Y. Liu, "A corpus-based approach for the prediction of language impairment in monolingual English and Spanish-English bilingual children," in *Proceedings of NAACL-HLT*, 2009, pp. 46–55.

[14] T. Solorio and Y. Liu, "Using language models to identify language impairment in spanish-english bilingual children," in *Proceedings of the ACL Workshop on (BioNLP)*, 2008, pp. 116–117.

[15] B. Roark, M. Mitchell, and K. Hollingshead, "Syntactic complexity measures for detecting mild cognitive impairment," in *Proceedings of the ACL 2007 Workshop on Biomedical Natural Language Processing (BioNLP)*, 2007, pp. 1–8.

[16] E. T. Prud'hommeaux and B. Roark, "Classification of atypical language in autism," in *Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics*, 2011.

[17] C. Callison-Burch, "Syntactic constraints on paraphrases extracted from parallel corpora," in *Proceedings of EMNLP 2008*, 2008.

[18] R. Barzilay and K. R. McKeown, "Extracting paraphrases from a parallel corpus," in *Proceeding of ACL*, 2001.

[19] M. Diab and P. Resnik, "An unsupervised method for word sense tagging using parallel corpora," in *Proceedings of ACL*, 2002.

[20] M. Sahlgren and J. Karlgren, "Automatic bilingual lexicon acquisition using random indexing of parallel corpora," *Natural Language Engineering*, vol. 11, no. 3, 2005.

[21] W. R. Shankle, A. K. Romney, J. Hara, D. Fortier, M. B. Dick, J. M. Chen, T. Chan, and X. Sun, "Methods to improve the detection of mild cognitive impairment," *Proceedings of the National Academy of Sciences*, vol. 102, no. 13, pp. 4919–4924, 2005.

[22] J. Morris, "The clinical dementia rating (CDR): Current version and scoring rules," *Neurology*, vol. 43, pp. 2412–2414, 1993.

[23] P. Liang, B. Taskar, and D. Klein, "Alignment by agreement," in *Proceedings of HLT NAACL*, 2006.

[24] F. J. Och and H. Ney, "A systematic comparison of various statistical alignment models," *Computational Linguistics*, vol. 29, no. 1, pp. 19–51, 2003.

[25] A. Yeh, "More accurate tests for the statistical significance of result differences," in *Proceedings of the 18th International COLING*, 2000, pp. 947–953.

[26] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 27, pp. 1–27, 2011.

[27] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: An update," *SIGKDD Explorations*, vol. 11, no. 1, 2009.

[28] C. Cortes, M. Mohri, and A. Rastogi, "An alternative ranking problem for search engines," in *Proceedings of WEA2007, LNCS 4525*. Springer-Verlag, 2007, pp. 1–21.

[29] T. Pahikkala, A. Airola, J. Boberg, and T. Salakoski, "Exact and efficient leave-pair-out cross-validation for ranking RLS," in *Proceedings of AKRR 2008*, 2008, pp. 1–8.

[30] J. Egan, *Signal Detection Theory and ROC Analysis*. Academic Press, 1975.