Alignment of narrative retellings for automated neuropsychological assessment

Emily Tucker Prud'hommeaux
A.B., Harvard College, 1996
M.A., University of California, Los Angeles, 2000

Presented to the Center for Spoken Language Understanding
within the Oregon Health & Science University
School of Medicine
in partial fulfillment of
the requirements for the degree of
Doctor of Philosophy
in
Computer Science & Engineering

August 2012

Center for Spoken Language Understanding
School of Medicine
Oregon Health & Science University

CERTIFICATE OF APPROVAL

This is to certify that the Ph.D. dissertation of
Emily Tucker Prud'hommeaux
has been approved.

Brian Roark, Thesis Advisor
Associate Professor, OHSU

Jan van Santen, Thesis Advisor
Professor, OHSU

Richard Sproat
Professor, OHSU

Chris Callison-Burch
Associate Research Professor
Johns Hopkins University

Acknowledgments

I am lucky enough to have had two advisors, Brian Roark and Jan van Santen, both of whom provided invaluable instruction and guidance throughout my years at CSLU. I thank Brian and Jan for sharing their expertise, for never losing their patience or sense of humor, and for motivating and inspiring me. Many thanks are also due to the other members of my committee, Richard Sproat and Chris Callison-Burch, for their interest in my research and their helpful feedback on my proposal and dissertation.

I am grateful to all of the faculty at CSLU, who not only taught me everything I know about speech and language processing but also willingly offered their input and encouragement. Zak Shafran, Esther Klabbers Judd, and Alison Presmanes Hill deserve special recognition for their contributions to this thesis.

Those of us at CSLU working on autism owe a great debt to the late Lois Black, without whom none of our work would be possible. I am also indebted to CSLU's clinical team, especially Beth Langhorst and Robbyn Sanger; to our support staff, especially Pat Dickerson and Cullen Conway; and to my labelers, Cheryl Green, Margit Bowler, Rachel Coulston, and Gordon Keepers.

CSLU would not be the mellow and friendly place it is without its graduate students. For making each trip to Beaverton a little bit nicer, I extend my heartfelt thanks to my fellow students, both past and present, especially Maider Lehr, Aaron Dunlop, Meg Mitchell, Géza Kiss, Eric Morley, Masoud Rouhizadeh, Mahsa Yarmohammadi, Rebecca Lunsford, Ethan Selfridge, Andrew Fowler, Russ Beckley, Nate Bodenstab, Steven Bedrick, Taniya Mishra Linger, and Kristy Hollingshead Seitz.

Finally, words cannot adequately express my gratitude to my family for their sage advice and their unshakable confidence in me; and to my husband, Marc, for his patience and his support, both emotional and technical. To Mom, Dad, Tom, and Marc, I offer my appreciation, admiration, and love.

Contents

List of Tables

List of Figures

Abstract

Alignment of narrative retellings for automated neuropsychological assessment

Emily Tucker Prud'hommeaux

Doctor of Philosophy
the Center for Spoken Language Understanding within
the Oregon Health & Science University
School of Medicine

August 2012
Thesis advisors: Brian Roark and Jan van Santen

As the prevalence of neurological disorders such as dementia, autism, and language impairment increases, so will the demand for simple, objective, and unobtrusive screening tools for these disorders. The automated analysis of narratives shows potential as a component of such a screening tool, since the ability to produce accurate and meaningful narratives is impaired in these populations. This dissertation investigates the reliability and diagnostic utility of automatically analyzing responses to a clinically elicited narrative retelling task, in which a subject listens to a brief story and must verbally retell the story to the examiner.

In order to establish the utility of using narrative retellings for diagnostic classification, we first demonstrate that manually assigned narrative recall scores can be used to accurately identify subjects with mild cognitive impairment and language impairment. We then present a method for extracting narrative recall scores automatically and highly

accurately from a word-level alignment between a retelling and the source narrative. We propose improvements to existing machine translation-based systems for word alignment, including a novel method of word alignment relying on random walks on a graph of interconnected nodes, which achieves alignment accuracy superior to that of standard expectation maximization-based techniques for word alignment in a fraction of the time required for EM. In addition, the narrative recall scores and related narrative fidelity features extracted from these high quality word alignments yield classification accuracy comparable to that achieved using manually assigned scores and significantly higher than that achieved using automated scoring techniques proposed in the literature. Finally, we apply these automated scoring and classification methods to spontaneous language samples elicited in other neuropsychological instruments, thereby demonstrating the flexibility and generalizability of these methods.

Chapter 1

Introduction

Virtually all human communication that portrays an event or sequence of events is framed as a narrative, from films, novels, and news reports to clinical notes, legal briefs, gossip, and any response to the question "what happened?". We understand and use narratives almost from the moment we begin to acquire language. Toddlers are able to recall an event from the recent past and accurately relate it to listeners unfamiliar with the event, and they show awareness of the ordering of and relationship between events in stories told to them (Fivush et al., 1987; Peterson, 1990). By they time they reach grade school, children are regularly producing narratives from their own experience and retelling stories they have heard with the expected episodic structure (Applebee, 1978; Stein, 1988). When recalling stories as adults, we tend to order the events according a common narrative framework even when the events are presented to us out of order or interleaved with unrelated events (Stein and Nezworski, 1978). It has been argued that narrative forms are how we store and represent all communicative information (Schank and Abelson, 1995).

In short, narratives give meaning and structure to our communication, and our successful interaction with others often depends upon our ability to coherently and accurately formulate new narratives and retell the ones we have heard or witnessed. This ability in turn depends on a wide range of cognitive functions, including language production and comprehension, long-term and working memory, and theory of mind. As a result, disorders such as language deficits, cognitive impairment, social and communication disorders, brain injuries, and mental illness can lead to difficulties in producing narratives. The analysis of narratives thus has the potential to reveal the presence of these disorders and some of their characteristic features, thereby opening up avenues for screening and diagnosis.

Many neuropsychological assessment instruments and protocols include a task in which the subject must produce a narrative. In some tasks, the subject narrates the events that unfold in a visual stimulus such as a drawing or a wordless picture book. In other tasks, the subject hears a brief narrative and must retell the narrative to the examiner. The narrative paradigm is useful for language analysis because it affords the elicitation of a relatively lengthy stream of uninterrupted spontaneous speech. The language used to convey a narrative can be analyzed in terms of its linguistic features, such as word count, sentence length, number of grammatical errors, sentence complexity, and vocabulary diversity, which can be extracted from a transcript using manual coding procedures (Miller et al., 2011; Scarborough, 1990) or automated techniques based in natural language processing (Sagae et al., 2005; Gabani et al., 2009; Roark et al., 2011). Alternatively, narratives can be evaluated according to their content. In many of the narrative tasks used in clinical settings, for instance, the score of a narrative is the number of predetermined key *story elements* that the subject uses in his narrative. This sort of content scoring is typically done in real time by a trained examiner following the guidelines published for the particular instrument, and few automated scoring procedures have been proposed. Analysis of narratives both in terms of content and form is widely used in neuropsychological evaluation, and differences in performance have been found to characterize certain disorders and to distinguish individuals with a particular disorder from typical controls, which suggests that narratives may be especially well suited for diagnostic screening purposes.

The importance of accurate screening tools for neurological disorders cannot be overstated, especially in light of the increased prevalence of disorders such as dementia and autism currently being observed worldwide (King and Bearman, 2009; Kim et al., 2011; Ferri et al., 2006; Brookmeyer et al., 2007). Dementia and other neurodegenerative disorders are being diagnosed at a higher rate for obvious demographic reasons: namely, improvements in the treatment and prevention of illness have allowed people to live past middle age into their seventies, eighties, and nineties, which are the years that generally mark the onset of dementia. In the industrialized world, for the first time in recorded history, the population over 60 years of age outnumbers the population under 15 years of age, and it is expected to be double that of children by 2050 (United Nations, 2002). As

the elderly population grows and as researchers find new ways to slow or halt the progression of dementia, the demand for objective, simple, and noninvasive screening tools for dementia and related disorders will grow.

The reasons for the increased incidence of neurodevelopmental disorders such as autism are less clear, though they can be partly explained by the broadening of definitions of these disorders, changing societal and cultural demands on children during their development, and improved diagnostic procedures. Although there is considerable disagreement about the best type of intervention for many developmental disorders, particularly autism, there is evidence that earlier treatment leads to more successful outcomes (Kasari, 2002; Blackman, 2002; Myers et al., 2007). Unfortunately, the difficulty of screening and the lack of trained clinicians in many areas of the world lead to delayed and sometimes incorrect diagnoses, underscoring the need for accurate and objective screening tools that can be administered in community and educational settings.

1.1 Problem statement

The demand for simple, objective, and unobtrusive screening tools for neurological disorders will continue to grow as the prevalence of such disorders increases. Because the act of narration taps into so many cognitive functions, narrative analysis could potentially point the way to the development of such screening tools.

Although there is a great deal to be learned from analyzing clinically elicited narratives, scoring approaches are not standardized across different instruments. Furthermore, the scoring procedures for most of the narrative tasks that are widely used in neuropsychological assessment have not been subjected to a rigorous analysis of their validity and reliability on the one hand, or their diagnostic potential on the other, particularly since they are rarely used individually for diagnostic purposes.

Inter-rater correlations are reportedly high under the scoring guidelines for some tests, but the scoring of narrative tests is inherently subjective. Narrative scoring requires training, experience, and expertise, and sufficient levels of reliability can be achieved only

when the examiners administering and scoring the test are experts. Thus, there is a need for objective scoring mechanisms, which can be provided via automated analysis.

In addition, the scoring procedures used in most narrative instruments assign a uniform importance to all content elements, and the score reported in the majority of narrative tests is simply the count of the number of story elements used. Such scoring procedures disregard interesting and potentially meaningful information about the fidelity of a retelling to the source narrative that is independent of the raw count of story elements used.

1.2 Research objectives

In this thesis, I propose solutions to the above problems that are grounded both in findings from neuropsychology research and in algorithms and approaches from natural language processing and machine learning. There are five major objectives of the research presented in this thesis, which I now briefly review.

Diagnostic utility of narrative recall scores. I will show that manually derived narrative recall scores can be used alone and in combination with other neurological data for highly accurate diagnostic classification.

Word alignment for automated scoring. I will demonstrate that narrative recall scores can be very accurately automatically extracted, using limited training data, from unsupervised machine translation-style word alignments of retellings to the source narrative. I will then propose new alignment strategies drawn from other NLP research areas that result in improvements in word alignment accuracy, which in turn improve scoring and classification accuracy.

Alternatives to standard scoring. I will explore alternative manual and automated scoring approaches inspired by work in psychology and in related natural language processing tasks, such as automatic summarization and machine translation.

Language features characterizing narrative fidelity. I will describe other features derived from word alignments that characterize narrative fidelity and validate the significance of these features for diagnostic classification via machine learning.

AD	Alzheimer's disease
ADOS	Autism Diagnostic Observation Schedule
AER	alignment error rate
AES	automated essay scoring
ALI	autism with language impairment
ALN	autism language normal (no language impairment)
ASD	autism spectrum disorder
AUC	area under the receiver operating characteristic curve
BDAE	Boston Diagnostic Aphasia Exam
DAT	dementia of the Alzheimer's type
DLD	developmental language disorder
HMM	hidden Markov model
LI	meeting criteria for a language impairment
LM-I	Logical Memory I
LM-II	Logical Memory II
LSA	latent semantic analysis
MCI	mild cognitive impairment
MMSE	Mini-mental State Exam
NEPSY	A Developmental NEuroPSYchological Assessment
NLP	natural language processing
NNM	NEPSY Narrative Memory
RTE	recognizing textual entailment
SLI	specific language impairment
(S)MT	(statistical) machine translation
SOV	subject-object-verb
SVM	support vector machine
SVO	subject-verb-object
TD	typical development
WLM	Wechsler Logical Memory subtest
WMS	Wechsler Memory Scale
WER	word error rate

Table 1.1: Acronyms and initialisms used in this thesis.

Extensions and applications. I will review our work in incorporating speech recognition into the narrative analysis framework and discuss application of the narrative analysis techniques to other semi-structured spontaneous language elicitation paradigms.

1.3 Organization of the thesis

The next three chapters of this thesis lay the groundwork for the original research that will be presented in the later chapters. In the Chapter 2, I discuss the role of narratives and

narrative analysis in the context of characterizing and diagnosing mild cognitive impairment, autism spectrum disorders, and language impairment. Chapter 3 reviews related work in established natural language processing research areas, including automated essay scoring, recognizing textual entailment, and automated evaluation of machine-generated language. The final background chapter, Chapter 4 discusses the areas of natural language processing that have a direct impact on the research presented in this thesis: word alignment in machine translation and graph-based methods of text analysis.

The remaining chapters contain the results of the original research carried out for this thesis: the collection and characteristics of the two main data sets that I will analyze (Chapter 5); the framework for using manually assigned narrative recall scores to perform diagnostic classification (Chapter 6); the method I propose for extracting narrative recall scores from a retelling using word alignment and improvements in word alignment that result in more accurate scoring and diagnostic classification (Chapter 7); and some manual and automated alternatives to standard scoring procedures, including features independent of element-level scores (Chapter 8). Chapter 9 discusses an extension of the techniques to a narrative task with a visual stimulus, as well as recent collaborative work on incorporating speech recognition into the narrative analysis pipeline in order to develop a truly automatic screening tool. I conclude with a chapter summarizing the contributions of this thesis and describing future work in applying the techniques and strategies developed in the preceding chapters to other clinical data and to more mainstream natural language processing tasks.

Chapter 2

Narratives in psychological assessment

Our ability to communicate with others coherently and effectively depends on our ability to formulate the narratives that explain our motivations and describe the things we have heard or seen. Because narration taps into so many different and important cognitive functions, narrative ability is often compromised in a wide variety of neurological disorders including senile dementia, language impairment, and social and communication disorders such as autism. The analysis of narratives thus has the potential to reveal the presence of these disorders and some of their characteristic features. In this chapter, I present a brief overview of theories on the content and structure of narratives, followed by a discussion of the role of narratives in neuropsychological evaluation.

2.1 Narrative Structure

In his Poetics, Aristotle defined the three basic components of narrative as a beginning, a middle, and an end. From a practical standpoint, most subsequent descriptions of narrative structure are variations on this theme, including the works of Labov and Propp, which are the sources of most of the approaches to narrative analysis employed by researchers in memory and language development. Labov and Waletzky (1967) divide the global structure of a narrative into four obligatory sections: the orientation to the setting and characters; the complication, or series of complicating events; the evaluation, which is the point at which the complication has reached an emotional maximum or *high point*; and the result or resolution. The sections contain different types of clauses possessing specific syntactic, morphological, and lexical properties which enable them to function in

ways that further the goals of those sections. In the clinically focused work on narratives discussed in Section 2.3.2 and Section 2.3.3, this narrative structure and the ideas of evaluation and high point play an important role.

Labov and Waletzky's work is often contrasted with the *narrative schema* approach to narrative analysis proposed by Propp (1968). Propp proposes that all narratives share an underlying structure: each begins with a setting, followed by any number of episodes, each of which consists of an initiating event or problem that the protagonist must address; an emotional or cognitive internal response to that event; an action by the protagonist set in motion by that response; a consequence or result of that action; and an emotional or cognitive reaction to the result. Labov and Waletzky's conception of the narrative is that it grows naturally in a linear or temporally appropriate way from the orientation to the result, while Propp suggests that narratives are constructed in a top-down fashion in which the narrator is aware in advance of the necessary components and their order and need only fill in the open slots with the appropriate content. Although these two conceptualizations of how we represent and generate narratives are different, their surface realizations are, of course, the same. For this reason, clinical studies of narratives, including those in Section 2.3.2 and Section 2.3.3 tend to mix and match ideas derived from both schools of thought.

Another way to analyze narratives is in terms not of a hierarchical structure but of *causal networks*, as outlined by Trabasso and colleagues (Trabasso et al., 1984; Trabasso and Sperry, 1985; Trabasso and van den Broek, 1985). In this framework, a narrative is represented as a network of connections among its various component events. A causal connection is a relationship between two events that is defined in terms of necessity: event A stands in a causal relationship with event B if event B could not have happened without event A first taking place. The more causal relationships an event shares with other events, the greater its perceived importance. A causal chain is the sequence of the causal connections that connects the opening or setting of the story to the resolution, and it is these chains that give structure to a narrative. It is only in the last few years that the idea of causal chains has started to be used in the clinical analysis of narratives, as discussed in Section 2.3.3, below.

2.2 Narrative elicitation paradigms

The act of verbally generating a narrative taps into a wide array of cognitive functions. First, a speaker generating a narrative must have the capacity to produce language; in the developmental psychology literature this is referred to as *expressive language* (American Psychiatric Association, 2000). In order to retell a narrative that has been verbally related, the speaker must also be able to understand language, or have adequate *receptive language* (American Psychiatric Association, 2000). Memory also plays a role in narrative production. General narrative production relies on *working memory*, the brain system that manages temporary and concurrent storage of information (Baddeley and Hitch, 1974; Baddeley, 1992). In addition, *long-term memory* is required for a delayed, as opposed to an immediate, narrative recall task, since the speaker must be able to remember all of the events of a narrative for an extended period of time. A speaker relating a narrative must also be able to understand the thoughts and motivations of the characters in a narrative, as well as the knowledge of the listener. This ability is often referred to as *theory of mind* (Baron-Cohen et al., 1985).

Because of the variety of intact cognitive functions required to generate a narrative, the inability to coherently produce or recall a narrative is associated with many different cognitive and developmental disorders, including dementia, autism, language impairment, intellectual disability, attention deficit disorder, and schizophrenia. Narrative tasks are widely used in neuropsychological assessment, and many commonly used instruments and protocols include a task involving narrative recall or production, including the Test of Memory and Learning (Reynolds and Bigler, 1994); the Luria-Nebraska neuropsychological battery (Golden and Freshwater, 2001); the Woodcock Johnson III Tests of Achievement (Woodcock et al., 2001); the Children's Memory Scale (Cohen, 1997); the Boston Diagnostic Aphasia Examination (Goodglass et al., 2001); the Wechsler Memory Scale (Wechsler, 1997); the NEPSY (Korkman et al., 1998); and the Autism Diagnostic Observation Schedule (ADOS) (Lord et al., 2002).

The narratives used in neuropsychological examination usually fall into one of three categories according to how the narratives are elicited: 1) narratives produced in response

to a narrative recall or episodic memory task; 2) narratives generated in order to describe events depicted visually in drawings, a picture book, or silent film; and 3) personal narratives. Because the narratives are elicited in different ways, the scoring and evaluation procedures vary according to the type of task.

In a narrative retelling task such as the Wechsler Logical Memory task (Wechsler, 1997) or the NEPSY Narrative Memory subtest (Korkman et al., 1998), the subject listens to a brief narrative and must verbally retell the narrative to the examiner immediately and in some cases, after a brief delay. The examiner then scores the retelling according to how many predetermined key points, or *story elements*, the subject uses in his retelling. In most normed narrative retelling tasks, the standard scoring procedures ignore information about the identities, relative importance, and ordering of the story elements recalled, although there is always a common, explicit linguistic reference in the form of the original narrative.

For narrative generation tasks in which a subject generates a narrative from non-linguistic data, such as a picture book or silent movie, elicitation and evaluation are more varied. In a structured scenario, such the Cookie Theft picture description task of the Boston Diagnostic Aphasia Examination (Goodglass et al., 2001) or the Renfrew Bus Story (Glasgow and Cowley, 1994), all subjects narrate the events depicted in a particular drawing, series of drawings, or film. There is often a predetermined list of "main ideas" against which a generated narrative is scored, although such narratives are sometimes analyzed in terms of the linguistic features used during narration. In the ADOS Wordless Picture Book activity (Lord et al., 2002), a less structured scenario, the child narrates a wordless picture book but can be guided or prompted by the examiner. There is no single book that must be used, and the examiner plays a role in furthering the narrative. Since there is no linguistic reference for these types of tasks, and since the reference itself may not be consistent in all administrations of the test, the criteria for evaluation are sometimes more subjective than those used in narrative recall tasks.

Personal narratives are the least structured form of narrative that is typically elicited and analyzed for neuropsychiatric purposes. Sometimes the subject is asked to describe his experience of a momentous historical event or to tell an interviewer about a memorable

occasion in his life, while in other cases, the subject writes an autobiographical sketch. Since there can be no common reference used for evaluation purposes, personal narratives are usually assessed in terms of linguistic features, such as word count, lexical choice, syntactic complexity, and overall features of coherence.

Note that the expected degree of variation in narrative output differs depending on the elicitation. Responses to a particular narrative recall task, when produced by typical speakers, are likely to share a similar lexicon and sequence of events, since the source narrative to be recalled is presented verbally. Tasks in which the speaker must describe a visually presented stimulus have more variation, and personal narratives have the broadest range of vocabulary and content.

2.3 Disorders of interest

In Chatper 5, I will discuss in detail two commonly used neuropsychiatric evaluation instruments that include a narrative component: the Wechsler Memory Scale (Wechsler, 1997) and the NEPSY (Korkman et al., 1998). In Chapters 6, 7, and 8, I describe how these narrative recall tests can be used to diagnose three different neurological disorders: Mild Cognitive Impairment (MCI), Specific Language Impairment (SLI), and Autism Spectrum Disorder (ASD). Detailed descriptions of these three disorders and the diagnostic criteria used to diagnose them are presented in Chapter 5. Here I present a brief overview of previous research in using manual expert analysis of narratives for diagnosis and screening of these three disorders.

2.3.1 Dementia and Mild Cognitive Impairment

When relating a narrative, a speaker must remember the events, characters, ordering, and details of the narrative he wishes to produce, while keeping track of what information he has already imparted and what remains to be told. Since these abilities rely heavily on memory and general executive function, it is not surprising that narrative production might be impaired in individuals with dementia of the Alzheimer's type (DAT) and other

dementias. Narratives produced by individuals with DAT and other dementias are characterized by fewer semantic themes (Vuorinen et al., 2000) and statements about the target content (Hier et al., 1985; Chenery and Murdoch, 1994; Ehrlich et al., 1997; Vuorinen et al., 2000), as well as more revisions (de Lira et al., 2011; Creamer and Schmitter-Edgecombe, 2010), repetitions (de Lira et al., 2011), and word retrieval problems (Hier et al., 1985; Ehrlich et al., 1997; de Lira et al., 2011). Although language fluency, measured variously as number of words, propositions, clauses, or utterances per narrative, was impaired in only some studies of DAT subjects, most of the work that looked at syntax did find significant differences including shorter utterances (Ehrlich et al., 1997), decreased syntactic complexity (Lyons et al., 1994; Hier et al., 1985; de Lira et al., 2011), and more fragments (Lyons et al., 1994; Hier et al., 1985; Ehrlich et al., 1997; Creamer and Schmitter-Edgecombe, 2010). Intrusions, such as incorrect informational statements, off-topic references, and extraneous information, were often observed (Ulatowska et al., 1988; Chapman et al., 1995; Creamer and Schmitter-Edgecombe, 2010), but the inconsistent definition of this phenomenon made coding unreliable.

These obvious problems in narrative expression observed in mild to moderate dementia may begin to appear in subtle ways before dementia can be diagnosed with standard screening instruments such as the Mini-Mental State Examination (MMSE) (Folstein et al., 1975). For this reason analysis of narratives and narrative recall tasks is widely used, sometimes in conjunction with other cognitive measures, for detecting the earliest stages of cognitive decline, known variously as Mild Cognitive Impairment (MCI), Subjective Memory Impairment, preclinical Alzheimer's, very mild dementia, and incipient dementia. (A more detailed discussion of the symptoms and diagnostic procedure is found in Chapter 5.) With the goal of finding such subtle differences specific to narrative ability, Bschor et al. (2001) compared narratives generated for the Cookie Theft picture description task (Goodglass et al., 2001) from subjects with MCI, mild DAT, moderate to severe DAT, and healthy older adults (OA). (More details about the Boston Diagnostic Aphasia Exam Cookie Theft picture description task are provided in Chapter 9.) Responses were scored according to the total number of words used, the number of words used to indicate persons and objects, locations, actions, features, and the total number of words belonging to these

four categories. The authors found that although these scores were able to differentiate the DAT groups from the non-DAT groups, they were not effective in distinguishing MCI from either OA or mild DAT. Chapman et al. (2002) compared memory for details in a narrative and memory for the overall gist in people with mild Alzheimer's disease (AD), MCI, and typical aging, using a complex elicitation and coding scheme. Subjects with MCI again patterned with the mild AD group, performing significantly worse than the typically aging group both on the gisting tasks and on the recall questions testing memory for story details, but showing better performance on gisting than detail memory.

More compelling work on narrative ability in MCI focuses on responses to the Wechsler Logical Memory subtest (henceforth abbreviated as WLM and described in more detail in Chapter 5) of the Wechsler Memory Scale (Wechsler, 1997). As noted earlier and as described in great detail in Chapter 5, under the standard scoring procedure, the WLM is scored only in terms of how many predetermined story elements a subject uses in his immediate and delayed retellings. Multiple studies have demonstrated a significant difference in performance on the WLM between subjects with MCI and typically aging controls under the standard scoring procedure, particularly in combination with tests of verbal fluency and memory (Storandt and Hill, 1989; Petersen et al., 1999; Wang and Zhou, 2002; Nordlund et al., 2005). Further studies have shown that performance on the WLM can accurately predict whether MCI will progress into Alzheimer's disease (Morris et al., 2001; Artero et al., 2003; Tierney et al., 2005). The WLM can also serve as a cognitive indicator of physiological characteristics associated with symptomatic Alzheimer's disease. WLM scores in the impaired range are associated with the presence of changes in Pittsburgh compound B and cerebrospinal fluid amyloid beta protein, two biomarkers of Alzheimer's disease (Galvin et al., 2010). Poor performance on the WLM and other narrative memory tests has also been strongly correlated with increased density of Alzheimer's related lesions detected in post-mortem neuropathological studies, even in the absence of previously reported or detected dementia (Schmitt et al., 2000; Bennett et al., 2006; Price et al., 2009).

The standard scoring guidelines for the WLM, which have been updated at regular intervals in the decades since the test was first introduced, allow for paraphrasing of some

but not all of the story elements. There have been a number of alternative scoring schemes proposed for the WLM, such as awarding more points for correct verbatim responses than paraphrased responses or fewer points for recalled details than for recalled main ideas. In an attempt to differentiate MCI from typical aging and mild dementia, Johnson et al. (2003) scored the responses according to whether the subjects produced the various propositions found in the story veridically (i.e., verbatim with allowance for difference in word order and morphology), in a gist fashion (i.e., paraphrasing but conveying the correct idea), or in a distorted fashion (i.e., generating propositions that were incorrect or did not appear in the source). Johnson found deteriorated veridical recall in the MCI subjects, with veridical recall scores correctly classifying 86% of the healthy and MCI subjects, while scores based on gist degraded classification performance.

These studies demonstrate the power of the Wechsler Logical Memory subtest to distinguish healthy aging from mild cognitive impairment. The alternative scoring procedures like those outlined in Johnson et al. (2003) are promising, but they tend to require expert coders and manual annotation. The automated techniques outlined in this thesis address this issue while also capturing additional features, such as element ordering and coherence, which few of the approaches described above considered, and the phenomenon of intrusion, whose definition in previous work was often too subjective to be reliably coded.

2.3.2 Language disorders

Although narrative recall ability in the elderly seems to be hampered by the problems associated with memory and cognitive decline, narrative recall performance in children is often more strongly correlated with measures of language development (Titley and D'Amato, 2008; Norbury and Bishop, 2003). Poor narrative production in early childhood is highly predictive of continued language impairment in later childhood and adolescence (Bishop and Edmunsson, 1987; Stothard et al., 1998), and language acquisition delays are correlated with poor narrative skills in later childhood (Miniscalco et al., 2007). For these reasons, there is a great deal of interest in the detailed analysis of narratives in children with language and communication disorders.

Early attempts to analyze narratives in the context of typical language development focused on defining narrative structure according to the theories of researchers in other fields and on using lexical choices and syntactic structures to assess understanding and coherence. Following work by Propp (1968) on narrative schema, Stein (1988) outlined six structural components of a narrative (opening, setting, action, obstacle, resolution, ending) and showed that children gradually incorporate these components into their narratives. In their work on the development of narrative competence, inspired by both Propp and Labov and Waletzky (1967), Bamberg and Damrad-Frye (1991) attempted to classify *evaluative devices* that speakers use to provide comments on the content and direction of a narrative for the purposes of engaging the listener such as the use of character speech, and devices characterized by lexical choices, including frames of mind, hedges, negative qualifiers, and causal connectors. The authors found that although adults use these devices frequently in their narratives, children incorporate them gradually into their narrative repertoires.

The structural and evaluative facets of narrative production continue to be explored, but the interest in identifying the core characteristics of language impairment has led to a broader interpretation of narrative ability, resulting in the evaluation of narratives both in terms of their generic linguistic features (e.g., sentence length, syntactic structure) as well as their narrative content and structure (e.g., organization, cohesion, reliability, use of devices). Perhaps not surprisingly, general linguistic features, rather than narrative-specific features, are often more effective in distinguishing children with language impairment. Liles et al. (1995) found that variables tied to linguistic structure, such as grammatical complexity and the use of subordinate clauses, were better predictors of language impairment than variables related to episodic structure. Language impairment among subjects participating in a study of narrative production by Botting (2002) was associated not with poor narrative structure or differences in the use of narrative devices, but rather with increased errors in marking tense, shorter sentences and stories, and the use of fewer subordinate clauses. Similarly, Norbury and Bishop (2003) found that the differences in narrative ability between children with typical development and children with specific language impairment (SLI) were characterized primarily by differences in syntactic

and morphological competence, although children with SLI did exhibit more ambiguous anaphoric references than the controls.

In the above studies, a very large number of annotations and codes for narrative features, borrowed and adapted from Bamberg and Damrad-Frye (1991), Stein (1988), Tager-Flusberg (1995) and others, were assigned to each narrative. Although the studies lack discussion about the reliability of the coding used in the more subjective categories, it is easy to imagine that inter-rater agreement might not have been particularly high. It also seems that there is little evidence in these results for a consistent area of deficit in narrative expression associated with language impairment. Perhaps in response to these concerns, more recent work has focused on developing standardized instruments for narrative analysis that are normed on large and diverse populations (e.g., Expression, Reception and Recall of Narrative Instrument (Bishop, 2004), Test of Narrative Language (Gillam and Pearson, 2004), Index of Narrative Microstructure (Justice et al., 2006), Narrative Assessment Protocol (Justice et al., 2010), and Narrative Scoring Scheme (Heilmann et al., 2010)). In addition, researchers not using these instruments have moved toward narrative evaluation scenarios that rely on coding schemes that are simpler and more objective and that attempt to separate informational content from the linguistic expression of the content.

The technique used by Bishop and Donlan (2005) for scoring the informational content of narratives involved simply counting how many "main ideas" from a predetermined list were used in narratives generated from pictures, although the coding procedures for syntactic analysis were as complex as those used in early work and similarly relied on ad-hoc word lists. The authors found that children with SLI produced fewer main ideas than typically developing subjects, as well as fewer terms expressing cognitive states and propositions. Goldman (2008), in her analysis of personal narratives in children with developmental language disorder (DLD) and autism, also relied on many of the familiar constructs of structure and coherence. Instead, however, of attempting to associate measures of syntactic complexity or lexical choice with these constructs, she coded most of them in a binary or ternary fashion, resulting in relatively high inter-rater reliability; the property of coherence, for instance, is coded as either 0 or 1 depending simply on

whether the events related followed a logical sequence. Despite the large number of variables coded, the author found only that children with DLD used fewer story elements. In an experiment by Dodwell and Bavin (2008), subjects with specific language impairment and two control groups, one consisting of age-matched subjects and one of younger but language-matched subjects, performed a narrative recall task and a task in which they generated a narrative from a picture book. In both cases, their responses were scored according to how many informational story elements from a predetermined list of elements were used. (Note that this is the same scoring technique used for the NEPSY Narrative Memory task and the Wechsler Logical Memory task, both of which will be discussed at length in Chapter 5.) The authors found that children with SLI produced significantly fewer story elements than their aged-matched peers on the story recall task, but generated a similar number of story elements in the story generation task, which suggests that the narrative difficulties experienced by children with SLI might be related to processing and memory for aurally received information.

In all of these studies, two trends emerge. First, children with language impairment use simpler syntactic constructions and make more syntactic and morphological errors, which should not be especially surprising given that language impairment is generally diagnosed using these criteria. Second, children with language impairment tend to produce less information when generating narratives even though they tend to follow the conventional narrative structure and use the expected narrative devices. This latter, less obvious, point will become important in Chapter 6, when I compare the retellings produced by children with developmental language disorder to those produced by children with autism.

2.3.3 Autism Spectrum Disorder

Atypical or idiosyncratic language is a characteristic of autism that has been observed since Kanner (1943) first gave the name to the disorder, and this atypicality is used as a diagnostic criterion in most of the widely used screening instruments for autism spectrum disorder (ASD) (Lord et al., 2002, 1994; Rutter et al., 2003). The current view is that phonological, morphological, and syntactic skills are typically spared in ASD (at least in the subpopulation of children with ASD without a co-morbid language impairment), while

pragmatic expression is impaired (Tager-Flusberg, 2001; Lord and Paul, 1997). Thus, although the atypical language that might be observed in a child with autism is very different from the disordered language of a child with a language impairment, it will likely have an impact on the child's successful social and communicative interactions with others. The study of narrative production in ASD is particularly interesting since relating a narrative requires pragmatic competence as well a developed theory of mind, which is also reportedly lacking in individuals with ASD (Baron-Cohen et al., 1985).

In one of the first studies to examine narrative production in ASD, Loveland et al. (1990) found that children and young adults with autism supplied significantly more bizarre, off-topic, and inappropriate information when generating narratives from a puppet show or video than verbal-age-matched subjects with Down Syndrome. The two groups did not differ in their ability to produce narratives with a recognizable structure or to answer questions testing recall, but the individuals with ASD were more likely to portray the characters as objects rather than agents and to fail to understand the relationships among the events.

Tager-Flusberg (1995), in one of the most widely-cited investigations of narratives in autism, applied many of the same narrative analysis techniques used in the context of language impairment screening (see 2.3.2) and derived from the work of Propp (1968), Labov and Waletzky (1967), and Bamberg and Damrad-Frye (1991). Narratives of a wordless picture book generated by children with low-functioning autism, intellectual disability, and typical development were analyzed in terms of linguistic features (e.g., story length, grammatical complexity), structure/schema (opening, orientation to characters and setting, explicit mention of the theme, resolution, formal ending), and the use of referential devices (e.g., anaphora), affective enhancement devices (e.g., emotionally charged and emphatic words, sounds effects) and social-cognitive enrichment devices (e.g., words related to thought processes, negative, statements of inference and cause). The children with autism produced shorter stories with less grammatical complexity and fewer causal statements. The authors attribute these differences to deficits in theory of mind; that is, a child with autism might have difficulty understanding what information is necessary for a listener to understand a narrative.

In similar work, Capps et al. (2000) considered both linguistic features (length, morphosyntax) and evaluative devices similar to those used in Tager-Flusberg (1995) and Bamberg and Damrad-Frye (1991), including causality, emotion and cognition, negatives, hedges, character speech, sound effects, and intensifiers. The results were similar to those found previously, namely that low-functioning children with autism produced shorter narratives with a lower degree of syntactic complexity and a restricted range of evaluative devices, but their frequency of use of evaluative devices was similar to that of children with typical development. In the ASD group, however, measures of theory of mind were strongly correlated with several features, including number and diversity of evaluations and the use of mental state language, indicating that deficits in theory of mind may lead to difficulties in employing evaluative devices and comments on the mental states of the characters. In a second study, Losh and Capps (2003) investigated narrative competence in high-functioning children with ASD using the same analysis framework described above. There were no between-group differences in linguistic features like those found in previous work, and the correlations between theory of mind and narrative ability reported previously did not hold in this population. The children with ASD, however, did supply more bizarre and irrelevant information in their responses, as found previously by Loveland et al. (1990) and later by Goldman (2008). In addition, the ASD children used fewer evaluative devices, complex syntactic structures, and causal explanations of events.

In recent work, Diehl et al. (2006) moves away from analysis in terms of hierarchical structure and lexically-determined evaluative devices and instead considers the structure defined by the causal relationships among events. Relying on the notions of causal connections and causal chains developed by Trabasso and colleagues (Trabasso et al., 1984; Trabasso and Sperry, 1985; Trabasso and van den Broek, 1985), the authors found that high-functioning children with ASD produced significantly fewer causal connections independent of the length of the narratives they produced, and they seemed not to rely on gist (as defined by the number of causal chains in a retelling divided by the number of causal chains in the source) to aid their story recall. In addition, although children with ASD

produced a similar proportion of story elements, they did produce more narrative intrusions, including both irrelevant information and excessive details taken from the pictures used to elicit the story.

Two of the themes consistently reported in the above research are the difficulties children with ASD have in employing causal language and understanding causal relationships, and the tendency of ASD children to include irrelevant and unnecessary content in their narratives. These two features will be among features explored in Chapter 8.

2.4 Summary

These studies described above revealed interesting observations about narrative ability in the three target populations that I will be discussing in this thesis: adults with MCI, children with language impairments, and children with ASD. With just a few exceptions, however, these approaches for analyzing and scoring narratives beyond the standard element-level scoring involved extensive manual annotation that usually required trained, expert coders. The researchers occasionally even found that the annotations they planned to use were simply too subjective to be reliably coded. In the coming chapters, I will present an approach for using automatically generated word alignments to objectively analyze narrative data gathered in a clinical setting for the purposes of diagnostic classification of mild cognitive impairment, language impairment, and autism spectrum disorder. Several of the diagnostically significant trends in narrative expression described above will be explored, along with some of the more promising techniques for analyzing narratives. First, however, I will present a discussion of related work in NLP, followed by an overview of the algorithms that lay the foundation for my new approach to automatically scoring and analyzing narratives for the diagnosis of neurological disorders.

Chapter 3

Related technical work

Language-based neuropsychological evaluation and analysis of narratives have only recently become the focus of research in computational linguistics or natural language processing. There are, however, similarities between many standard NLP research areas and the work described in this thesis. Tasks such as automated essay scoring, paraphrasing, and recognizing textual entailment have characteristics in common with analysis of narrative fidelity. In addition, the techniques for evaluating of machine generated text in the context of machine translation and automatic summarization may be useful in approximating human scores of narratives. In this chapter, I review work in these areas that might have useful applications for the task of analyzing narrative retellings.

3.1 Automated essay scoring

The standard NLP task most closely resembling analysis of narratives for neuropsychological assessment is automated essay scoring (AES). The similarities between scoring essays written in response to a particular prompt and scoring retellings of a particular narrative are quite obvious: a better retelling, like a better essay, will contain pertinent information and will present that information in a way that can be easily understood. In addition, like narrative retellings, essays can be evaluated in terms of both form (also known as style or usage) and content.

Form in the context of automated essay scoring is usually assessed in terms of linguistic features ranging from very basic numerical values, such as overall word count, to more complex syntactic and discourse-level characteristics, such as number of subordinate

clauses or the presence or absence of a thesis statement. The evaluation of form in essays is not unlike the techniques used to measure linguistic complexity in retellings discussed in Chapter 4. Automated scoring of essay content, which more closely resembles the task discussed in this thesis, typically relies on lexical similarity between the essay and a set of training essays written in response to the same prompt. Systems typically use measures of vocabulary overlap, similarities derived using the vector space model (Turney and Pantel, 2010), or a variant of latent semantic analysis (Landauer et al., 1998) to determine the lexical or semantic conceptual distance between the essay to be graded and a set of manually graded essays.

Most AES systems select a set of of these form and content features and then use some standard statistical technique, such as multiple regression, to determine the best weights for those features according to the human scores assigned to the essays in their training sets (Shermis et al., 2010). In systems that use content features, there is typically a separate training set for each specific essay prompt, while systems that rely only on form and style features can be more domain independent. We note that the details about the techniques and features described in the AES literature often remain undisclosed, as some AES systems, in particular IntelliMetric from the Vantage Learning corporation, are developed by organizations hoping to monetize their particular approach to essay scoring. Thus, although I will not be able to directly compare my approach to evaluating narratives with the evaluation approaches used by the most successful AES systems, I will be able to explore a few related techniques.

The majority of AES systems focus primarily on features related to writing form and quality rather than content. Some systems, such as Project Essay Grade (PEG) (Page, 1966; Page and Petersen, 1995), subscribe entirely to this approach, relying on dozens of easily extracted count-based measures, such as essay length or counts of different word classes, that can serve as approximations for well-formedness, complexity, and sophistication of writing style. The Vantage corporation does not specify the set of features used in their proprietary AES system, IntelliMetric, but their advertising materials indicate that the majority of their hundreds of features are related to form and style.

The Educational Testing Service (ETS) provides far more information about their AES system, e-rater (Burstein et al., 2003; Attali and Burstein, 2006). The e-rater feature set includes features related to errors in grammar, usage, style, and mechanics; lexical complexity features that measure vocabulary level and word length; features related to the structure and organization of the essay; and lexical content features. The lexical content evaluation component of their e-rater uses a vector space model of vocabulary frequency similarity to compare the essay to be graded with the numerous training essays in the six manually assigned score categories (1 through 6) that are used by ETS to grade essays. Each essay to be automatically graded is converted to a vector, where each element in the vector represents a word in the essay and the value of that element is a variant of tf*idf: the term frequency is expressed as a ratio of its frequency in the essay compared to the maximum frequency of any word in that essay; the inverse document frequency is log of the standard measure of inverse document frequency over the entire set of training essays. A similar vector for each score category is constructed from all of the essays in that score category. The two features extracted are (1) the value of the score category whose cosine similarity was the largest, and (2) the cosine similarity of the essay to the set of essays in the highest score category. This particular variation of tf*idf might not be especially relevant for narrative fidelity evaluation, but the vector space model itself could prove to be useful.

At the extreme of content end of the form-content spectrum is the Intelligent Essay Assessor (IEA) (Landauer et al., 2003). IEA uses latent semantic analysis (LSA) to measure the content similarity between an essay and other manually graded essays for the same prompt. The coherence of the essay is also measured using LSA by measuring the similarity between each paragraph and the preceding paragraph. Although the IEA approach focuses primarily, if not exclusively, on content, it achieves as high correlations with manual graders as systems that use features related to form and style. As discussed in Section 4.1, LSA has previously been used to generate automatic scores for neuropsychological narrative recalls tests (Dunn et al., 2002), and I will investigate using LSA with the WLM and NNM data in Chapter 8.

Simple n-gram overlap between an essay and one or more expert-written "gold" essays on the same topic might seem like a useful approximation to human scoring. Although none of the major AES systems seems to use n-gram overlap as a content measure, there has been work on using a modified version of BLEU for essay scoring (Alfonseca and Pérez, 2004; Pérez et al., 2004, 2005). The correlations between scores assigned by humans and both BLEU and their modified version of BLEU are noticeably lower than those reported for the systems described above and the correlations typically reported between human graders. We will see that BLEU scores correlate much more highly with manually-assigned scores of narrative recall in Chapter 8.

On a somewhat intriguing note, recent work by the researchers at ETS (Attali et al., 2010) suggests that while human graders believe that content and topic weigh heavily in their grading decisions, AES systems that disregard the content of an essay correlate as well with human graders as systems that use content-related features. It seems that content, at least in the context of essay grading, does not necessarily play a crucial role. In contrast, we will see in the coming chapters that the analysis of content in a narrative seems to provide more diagnostic classification power than linguistic features relating to structure.

The differing behavior of comparable techniques in automated essay scoring and automated narrative scoring is ultimately not surprising. First, the narratives we are analyzing are spoken, while the essays scored by the systems are written, which makes the two domains difficult to compare. The range of expected acceptable answers in an essay is far broader than that in a retelling, since the retelling should contain only the information explicitly reported in the source narrative. The essays typically being graded by these systems are several hundred words long, and they are graded with the expectation that they will follow a particular structure. Given the length of a typical narrative recall story, the retellings are typically quite short, and they need not follow a particular structure other than the temporal ordering of events, if that. Perhaps most crucially, the WLM scoring procedure is based on recalling specific elements from the source story, some of which must be recalled verbatim. Repetition of the elements suggests better mastery of

the recalled material, and a subject who introduces novel content or chooses to use synonyms or paraphrases for the elements that must be recalled verbatim should be penalized by an automatic scoring system. In contrast, an essay grading systems should reward innovation in both vocabulary and content and should penalize repetition. This last point could explain why content-based measures do not contribute as much to AES accuracy as structural measures.

3.2 Summarization

Although the task of automatically generating a summary from a length of text is not especially applicable to the analysis of narrative retellings, there has been some research in analyzing the relationships between documents and human-generated abstractive summaries of those documents with the goal of building models for automatic abstractive summarization systems. Some work in this area relies on monolingual word alignments between documents and abstracts (Daumé and Marcu, 2005; Jing and McKeown, 2000), just as the approach for scoring narrative retellings that I will present in Chapter 7 relies on monolingual word alignments between retellings and a source narrative. Daumé and Marcu (2005) argue that existing machine translation word alignment packages such as Giza++ (Och and Ney, 2003) are not sufficient for the task of aligning documents and abstracts, particularly because much of the information contained in the document will likely not be included in the abstract. They propose a more complex approach in which sentences are selected from the source document that have some likelihood of containing the same information contained in each sentence of the abstract. A semi-HMM is built over these sentence pairs that aligns both words and phrases and that calculates word correspondence probabilities separately for words, words identities, stems, and synonyms.

Such an approach to monolingual word alignment of abstract and a source document might seem promising for the task of aligning a retelling to a source narrative. There are, however, a number of differences between these two tasks that makes this approach impractical. Although narrative retellings are somewhat like abstracts in the sense that they might lack some of the information contained in the source narrative, there is a good

chance that a retelling produced by an unimpaired subject will contain all or almost all of the source narrative content. Furthermore, it is also likely that a retelling, regardless of the subject, will contain *intrusions*, or entire utterances that are not related to the source narrative at all, such as conversational asides, comments about the story, or confabulations. In the data analyzed in this thesis, the retellings and the source narratives are relatively short, while the documents used in summarization model building are very long. In addition, the difference in length between a retelling and the source narrative is usually quite small, while an abstract, by definition, is significantly shorter than the document it summarizes. All of these factors conspire to render the approach described in Daumé and Marcu (2005), which requires that similar sentences be extracted from the document and abstract, difficult to adapt to the task of aligning retellings and a source narrative.

3.3 Textual entailment and paraphrasing

Textual entailment is another area of NLP research that could provide insight into how to analyze and score narrative retellings. The goal of the task of recognizing textual entailment (RTE) is to determine whether the meaning of some text fragment (the hypothesis) entails or can be inferred from another text fragment (the text). The data used in RTE competitions typically define the texts as sentence- or paragraph-length passages and the hypotheses as short sentences. An example of such a text-hypothesis pair from the 2010 PASCAL challenge is shown below (Bentivogli et al., 2009):

> Text: The Grapes of Wrath, published exactly 70 years ago, can be seen as a prophetic novel, rooted in the tragedies of the Great Depression, but speaking directly to the harsh realities of 2009, writes Steinbeck scholar Robert DeMott. Steinbeck's epic novel, which traces the harrowing exodus of Tom Joad and his family from blighted Oklahoma (where they are evicted from their farm), across the rugged American south-west via Highway 66, and on to what they mistakenly hope will be a more promising future in California, is considered by many readers to be the quintessential Depression-era story, and an ironic reversal of the rags-to-riches tale favoured by many optimistic Americans.
>
> Hypothesis: "The Grapes of Wrath" was written by Steinbeck.

In the context of narrative retelling evaluation, the text might be considered the source narrative and the hypothesis could be a particular retelling. Alternatively, the text could be a retelling, and the hypotheses could be each of the sentences, phrases, or story elements of the source narrative.

Textual entailment need not be limited to passage-to-sentence matching scenarios. Automatic paraphrase recognition can be considered to be bi-directional RTE, in which the text and hypothesis are both words or phrases that convey the same meaning. Many current RTE systems (see citations below) incorporate features derived from multiple levels of analysis, from measures as simple as lexical similarity to more complex features related to anaphora, semantic roles, and temporal relationships. Many of these systems, along with a number approaches to paraphrasing, rely on a lexical alignment between a text and a hypothesis, which seems especially relevant to the work presented in this thesis, as will become clear in the coming chapters.

The techniques used in RTE and paraphrasing to derive lexical alignments are varied, ranging from approaches based on orthographic similarity to co-occurence-based methods to discriminative phrase-matching algorithms to the IBM mixture models normally used in machine translation. In their general-purpose RTE system, Glickman et al. (2006) determined word alignment probabilities by measuring document-level word concurrence statistics collected from external web data. The alignment approach developed by Nielsen et al. (2009), in which "facets" of knowledge in a hypothesis are aligned with corresponding facets in the text, relies on extensive manual annotation of hypothesis-text training pairs. In very recent work, Meurers et al. (2011) performed alignment using a pair-matching technique called the Traditional Marriage Algorithm (Gale and Shapley, 1962), an algorithm for matching individuals in one group (in this case, words and phrases in a text or hypothesis) with individuals in another group according to a set of predefined preferences (in this case, similarity features). The winning system of the 2006 PASCAL RTE challenge, LCC (Hickl et al., 2006), used a maximum entropy based classifier to align text chunks identified in the hypothesis and text by a chunk parser. In subsequent work, Hickl and Bensley (2007) used a variation of a discriminative approach originally proposed by Taskar et al. (2005) to improve word alignment for machine translation.

MacCartney et al. (2008) argue against using any existing machine translation word alignment tools for deriving alignments for textual entailment problems. The authors propose instead using an aligner specifically designed for the task of textual entailment, which they note has many characteristics that make MT-style alignment less than ideal. Their phrase-level alternative, the MANLI aligner, uses a feature-based scoring system to select the phrase-to-phrase alignment with the best combination of substitutions, deletions, insertions, and matches given a set of phrase-level features, including measures of length, constituency, orthographic similarity, semantic similarity, and distributional similarity. The parameter weights of the aligner are trained on hand-aligned sets of text-hypothesis pairs. The authors found that MANLI achieved higher phrase alignment accuracy than either Giza++ or Berkeley aligner, even after supplementing the Giza++ and Berkeley training data with word identities. However, the RTE system relying on MANLI was significantly outperformed by the leading RTE system, LCC, described above (Hickl et al., 2006; Hickl and Bensley, 2007).

The paraphrase recognition and generation literature is more friendly to the word-level alignment approaches normally used in machine translation. The pivoting approach for paraphrasing first outlined by Bannard and Callison-Burch (2005) and expanded in subsequent papers (Callison-Burch et al., 2006; Marton et al., 2009; Wang and Callison-Burch, 2011) relies on phrase tables built from MT-style word alignments. Both Quirk et al. (2004) and Mehdad et al. (2011) also used Giza++ word alignments for generating paraphrases, and Munteanu and Marcu (2006) used Giza++ word alignments as a starting point for their system for identifying sub-sentential fragments. The techniques in Chapters 7 and 8 to score narrative retellings are in part inspired by these sorts of paraphrase generation methods, which assume that the IBM models and related techniques for generating word alignments for machine translation models are likely to be sufficient for other tasks.

Note that most, if not all, of the successful RTE and paraphrasing systems rely heavily on external resources. Many RTE systems look for synonyms and hypernyms using WordNet or paraphrases using an existing dictionary or paraphrase corpus. Many systems require grammars for constituent and dependency parsing. Alignment-based paraphrase

systems are typically built using large parallel or comparable corpora. Template-based systems also require large monolingual corpora from which to extract grammatical and semantic templates.

In contrast, the techniques presented in this thesis for evaluating narratives do not require input or training data from any outside sources. One reason for this is that our early attempts to use outside sources met with little success. When attempting to score retellings by using WordNet to look up synonyms for each of the story elements, we found that very few of the words used in either of the narratives belonged to synsets with more than one member, and their hypernyms and hyponyms were not usually acceptable substitutes according to published scoring guidelines. Similarly, the existing paraphrase corpora did not contain paraphrases for the majority of story elements. Generating paraphrases of story elements using the method described by Callison-Burch et al. (2006) met with little success, since very few of the story elements were found in the phrase tables used to generate paraphrases, and the paraphrases that were generated did not typically correspond to the paraphrases used by the subjects, even when using an in-domain parallel corpus.

The lack of utility of external resources for this task is likely due to distinctive features of the narrative retellings data sets and the manual scoring procedure. The retellings are spoken language rather than written, and they are rendered in a conversational and sometimes interactive style, which is not represented well in most external corpora. Furthermore, the speakers producing the retellings analyzed in this thesis are elderly or very young, and a large number of them are neurocompromised, which would make it difficult to find appropriate in-domain annotated external resources. The WLM story contains numerous proper names and numerical quantities, which cannot be paraphrased and are not general enough to be considered named entities. The NEPSY story deals with themes that are unlikely to be represented in any existing corpus of paraphrases or in any readily available bilingual or monolingual parallel corpus. Interpreting the retellings does not require real-world knowledge, and subjects are not rewarded for recalling information that is implied but not explicitly stated in the source narrative.

The issue, however, is not simply that external resources don't seem to help in this task. Rather, it seems that they are entirely unnecessary for the particular task of scoring

retellings of a single source narrative collected in a clinical setting. The best corpus for finding lexical equivalences is an in-domain corpus, and the only truly in-domain corpus in this case is the set of narrative retellings itself. As I will outline in the following chapters, the techniques I have developed exploit the unique features of the data set. Despite their independence from external resources, these techniques achieve scoring accuracy levels higher than those usually reported in the textual entailment recognition literature. This is not to suggest that our approach could be adapted to the textual entailment task. On the contrary, it demonstrates that that the two tasks require different approaches, despite their surface similarities.

3.4 Automatic evaluation of machine-generated language

Another possible approach to the problem of analyzing narrative retellings might be to treat the comparison of a retelling with the source narrative as a machine-generated language evaluation task. In many approaches to automatic evaluation of language samples, a candidate output, such as a machine-translated sentence or an automatically generated summary, is compared against one or more human-generated reference translations or summaries. In applying such techniques to our narrative retelling data, we can consider the source narrative to be the reference and each retelling to be a candidate that must be evaluated against that reference.

Several packages for evaluating machine translation are readily available and easily adapted to this scenario. The de facto standard for machine translation output evaluation is BLEU (Papineni et al., 2002). BLEU is often described as a measure of n-gram overlap between candidate text generated by an MT system and a human reference translation. More precisely BLEU derives all of the possible n-grams from $n = 1$ to $n = 4$ from a candidate MT output text, and calculates for each n-gram order, the number of n-grams also appearing in the reference. The final score is the geometric mean of the precision for each of the four n-gram orders, scaled by a factor that penalizes candidate sentences whose length is unexpectedly short given the reference translation.

A more complex method for evaluating machine translation output is Meteor (Denkowski and Lavie, 2011). Meteor derives an alignment between a reference and a candidate by finding all of the word-to-word matches, along with matches of stems, synonyms, and paraphrases. Stemming is defined for several languages, and the synonyms and paraphrases are drawn from a predefined language-specific set. The alignment is the largest set of all the possible matches that includes either zero or 1 match for each word, maximizes the number of matched words, minimizes the number of chunks, and minimizes the sum of the distances between the positions of the items matched in the two sentences. From this alignment a precision and a recall score are derived while taking into account different possible weights for function and content words, and from the recall and precision values, an F-measure is calculated. That F-measure is then multiplied by a weighted fragmentation penalty that is the ratio of the number of chunks to the number of matched words. The various weights and parameters are tuned against a development set. The authors report that Meteor correlates more highly than BLEU with human judgements of machine translation output.

A third commonly used MT evaluation metric is TER (Snover et al., 2006) and its variants HTER (Snover et al., 2006) and TER-Plus (Snover et al., 2009). TER, which stands for *Translation Edit Rate*, calculates a particular kind of edit distance between a candidate and its reference that is meant to correspond to the minimum number of edits a human would have to make to convert the candidate translation into the reference. In contrast to an edit distance metric, such as word error rate, TER allows word sequences to be moved as a single unit, reducing the number of individual word shifts and thereby avoiding penalizing what could be acceptable phrasal reordering. The original formulation of TER is the number of edits required divided by the total (or average) number of words in the reference(s). A recent enhancement of TER, TER-Plus, uses the same edit distance metric as TER, but it allows substitutions of stems, synonyms, and paraphrases, which are generated in the same way as they are in Meteor. This latest version of TER correlates more highly than both Meteor and BLEU with human judgements of machine translation output.

The analog to BLEU for automatic summarization is ROUGE (Lin, 2004). ROUGE is also an n-gram overlap measure, but unlike BLEU, which measures precision, ROUGE is a recall measure. ROUGE is calculated by counting all of the n-grams of a given order that appear in both the reference and the candidate and dividing that sum by the total number of n-grams of that order in the reference summary. The n-gram order used by ROUGE is not predetermined and can vary according to the needs of the evaluator. Numerous variations of ROUGE have been proposed, including ROUGE-L which is a measure of the longest subsequence shared by the reference and candidate summaries, and ROUGE-S, which counts non-adjacent bigrams separated by a set number of intervening words.

We again note many differences between our data and the output that is usually analyzed with these techniques. There is no particular reason to believe that techniques developed specifically for evaluating machine generated written language will be relevant when analyzing human generated spoken language retellings. We also note that the majority of these evaluation techniques assume that the reference and candidate sentence are roughly the same length and conveying roughly the same information, something that cannot be safely assumed in the case of narrative retellings. Presenting the words in the correct order will likely result in a higher TER or METEOR score, while it will have no impact on a narrative retelling score. Nevertheless, since these evaluation methods are readily available and easily applied to our data, I will investigate their utility for scoring narrative retellings in Chapter 8.

3.5 Other areas

There are a number of other subfields in NLP that might have something to contribute to the task of evaluating narratives. Automatic evaluation of text coherence for summarization or for essay scoring, for instance, could potentially be used to evaluate the coherence of a narrative. The approaches typically used in the literature, however, would be difficult to adapt to narrative recall data for number of reasons. The LSA and vector space model techniques used to assess coherence in essays, such as those described in Foltz et al. (1998) and Higgins et al. (2004), might not work well with the narratives used in clinical

narrative recall tasks since these techniques expect lengthy essays consisting of multiple paragraphs, each containing multiple sentences, with each sentence functioning in a particular role. The source narratives used in clinical settings and the retellings elicited are generally quite short, and neither would conform to the expected structure of an essay. The grid-based techniques used by Barzilay and Lapata (2008) and Lapata and Barzilay (2005) to evaluate coherence on automatically generated extractive summaries might be more appropriate for the task of evaluating coherence in a brief narrative, but they would also be unnecessarily complicated. Much of the machinery of these approaches depends on extracting dependency relationship and resolving coreference in novel text, neither of which is required to analyze a retelling since the correct information is contained in the source narrative and hence is known in advance. All that is needed is a method for mapping the retelling to the source; from this mapping, the difference between the two versions of the story can be evaluated.

Work on extracting temporal structure in order to determine the order of events in a news story (Mani and Pustejovsky, 2004; Pustejovsky et al., 2005) might seem relevant for analyzing narratives. In the case of clinically elicited narrative retellings, however, the correct order of events is already known because it is contained in the source narrative. The order of events in a retelling can be inferred from the alignment of the retelling to the source narrative. Similarly, research in automatically learning narrative schemata and event chains (Chambers and Jurafsky, 2008, 2010) is not applicable to narrative recall analysis since the narrative structure of a retelling can be inferred from the alignment to the source narrative, whose narrative structure is already known.

3.6 Summary

Although I will appeal to some of the techniques introduced in this section, I am fortunate to be dealing with relatively structured data. These narratives do not need to be evaluated or annotated in isolation relative to an abstract idea of the appropriate structure or components in a well formed narrative. Rather, they must be analyzed in terms of their *narrative fidelity*, that is, how well they recreate the events of the original source narrative.

In both of the data sets I discuss, each subject hears the same story, and each subject's retelling is scored using the same metric. This inherent structure enables me to consider analysis methods that are less complex than those described above and are completely independent of external resources.

Chapter 4

Technical background

In this chapter, I discuss the few instances of previous research in using NLP techniques to automatically analyze narratives for neuropsychological evaluation. I then proceed to an in-depth discussion of the techniques adapted from other areas of NLP to the task at hand in this thesis, namely, the analysis of the content of narrative retellings for the purposes of neuropsychological evaluation and classification. This includes a review of word alignment for machine translation, as well as a discussion of graph-based methods for ranking documents and sentences in information retrieval and summarization, which I adapt to the task of improving word alignment.

4.1 Automated evaluation of narratives

Most previous work in applying automated analysis of unannotated transcripts of narratives for diagnostic purposes has focused not on evaluating properties specific to narratives but rather on using narratives as a data source from which to extract speech and language features. Solorio and Liu (2008) were able to distinguish the narratives of a small set of children with specific language impairment (SLI) from those of typically developing children using perplexity scores derived from part-of-speech language models. In a follow up study on a larger group of children, Gabani et al. (2009) again used part-of-speech language models in an attempt to characterize the agrammaticality that is associated with language impairment. Two part-of-speech language models were trained for this experiment: one on the language of children with specific language impairment and one on the language of typically developing children. The perplexity of each child's utterances was

calculated according to each of the models. In addition, the authors extracted a number of other structural linguistic features including mean length of utterance, total words used in the narrative, and measures of accurate subject-verb agreement. These scores collectively performed well in distinguishing children with language impairment, achieving an F1 measure of just over 70% when used within a support vector machine (SVM) for classification.

Roark et al. (2011) extracted a subset of the features used by Gabani et al. (2009), along with a much larger set of language complexity features derived from syntactic parse trees for utterances from narratives produced by elderly subjects for the diagnosis of MCI. These features included simple measures, such as words per clause, and more complex measures of tree depth, embedding, and branching, such as Frazier and Yngve scores. Selecting a subset of these features for classification with an SVM yielded a classification accuracy of 0.73, as measured by the area under the receiver operating characteristic curve. The data analyzed in the research of Roark et al. (2011) is a subset of the data that will be analyzed in this thesis, which provides an opportunity to compare diagnostic classification based on linguistic features to classification based on features of narrative content and fidelity.

An alternative to analyzing narratives in terms of lower-level linguistic features is to evaluate the content of the narratives themselves in terms of their fidelity to the source. Hakkani-Tur et al. (2010) developed a method of automatically evaluating an audio recording of a picture description task, in which the subject looked at a picture and narrated the events occurring in the picture. After using automatic speech recognition (ASR) to transcribe the recording, the authors measured unigram overlap between the ASR output transcript and a predefined list of key semantic concepts. This unigram overlap measure correlated highly with manually assigned counts of these semantic concepts. The authors did not investigate whether the scores, whether derived manually or automatically, were associated with any particular diagnostic group or disorder, but their results point to the potential for incorporating ASR into narrative analysis, which will be discussed in Chapter 9.

Dunn et al. (2002) are among the only researchers (other than the author of this thesis and her collaborators) to apply automated NLP-based methods specifically to scoring the Wechsler Logical Memory subtest and determining the relationship between these scores and measures of cognitive function. In an investigation comparing standard methods of scoring the WLM with alternative methods, the authors used latent semantic analysis (LSA) to measure the semantic distance from a retelling to the source narrative. The LSA cosine distance correlated very highly with the scores assigned by examiners under the standard scoring guidelines and with independent measures of cognitive functioning. In subsequent work comparing subjects with and without an English-speaking background (Lautenschlager et al., 2006), the authors propose that LSA-based scoring of the WLM as a cognitive measure is less biased against people with different linguistic and cultural backgrounds than other widely used cognitive measures. This work demonstrates not only that accurate automated scoring of narrative recall tasks is possible but also that the objectivity offered by automated measures has specific benefits for tests like the WLM, which are often administered by practitioners working in a community setting and serving a diverse population.

In this thesis, I will investigate the techniques just described for analyzing narratives in terms of their content and fidelity to the source, and I will compare them to the approaches for narrative analysis that I have developed. With that in mind, I now turn to a review of the relevant techniques and methods from the NLP literature that I will employ in my approach to narrative analysis.

4.2 Word alignment

An ideal retelling consists of the same sequence of events contained in the original narrative with variation in wording and word ordering. A good retelling thus resembles a translation of the original narrative into the language of the subject. The idea of treating retelling as a special case of translation is one of the ideas driving the work presented in this thesis.

Most statistical machine translation (SMT) systems begin with a parallel bilingual corpus, in which each sentence on one side of the corpus is a translation of the sentence

in the corresponding location in the other side of the corpus (Lopez, 2008). Figure 4.1 provides a snippet of the Europarl corpus for the language pair French-English, which was built using the transcribed proceedings of European parliamentary sessions (Koehn, 2005). Each English sentence on the right is a manually generated translation of the corresponding French sentence on the left. The parallel corpora used in machine translation are typically very large; the French-English Europarl corpus, for instance, contains over 1.8 million parallel sentences with over 54 million words in French and over 50 million words in English. This is not the case for clinically elicited narrative retellings, as we will see in Chapter 7.

In many widely used SMT software packages, such as Moses (Koehn et al., 2007) and Joshua (Li et al., 2010), the first step in building a model to translate from one language (the source language) to another (the target language) is to determine the alignments between the words on the source language side of the corpus and the words on the target language side of the corpus. The alignment between two words is a way of representing that one of the words in the alignment pair can be considered a translation of the other in the pair. From these word alignment pairs, which indicate some level of translational correspondence between words in two languages, pairs of corresponding strings and structures can be extracted, which can then be used to build a translation model. We refer the reader to Lopez (2008) for a detailed overview of past and current approaches to statistical machine translation.

Deriving the correct word alignment is not simply a matter of looking words up in a dictionary. A single word can have multiple translations in another language, depending not only on the syntactic and semantic context but also on the particular sense of the word that is being used. In addition, word alignment cannot be performed simply by moving monotonically through the words in one language and aligning them to words in the other language via some simple metric such as Levenshtein distance. There is no universal ordering of words or of syntactic structures that all languages follow. In English, for instance, the adjective typically precedes the noun it modifies, while in French, the adjective usually follows the noun it modifies. English follows subject-verb-object word order in a sentence in which the verb can take an argument, while a plurality of the

French	**English**
Reprise de la session	Resumption of the session
Je déclare reprise la session du Parlement européen qui avait été interrompue le vendredi 17 décembre dernier et je vous renouvelle tous mes vœux en espérant que vous avez passé de bonnes vacances.	I declare resumed the session of the European Parliament adjourned on Friday 17 December 1999, and I would like once again to wish you a happy new year in the hope that you enjoyed a pleasant festive period.
Comme vous avez pu le constater, le grand "bogue de l'an 2000" ne s'est pas produit. En revanche, les citoyens d'un certain nombre de nos pays ont été victimes de catastrophes naturelles qui ont vraiment été terribles.	Although, as you will have seen, the dreaded 'millennium bug' failed to materialise, still the people in a number of countries suffered a series of natural disasters that truly were dreadful.
Vous avez souhaité un débat à ce sujet dans les prochains jours, au cours de cette période de session.	You have requested a debate on this subject in the course of the next few days, during this part-session.
En attendant, je souhaiterais, comme un certain nombre de collgues me l'ont demandé, que nous observions une minute de silence pour toutes les victimes, des tempêtes notamment, dans les différents pays de l'Union européenne qui ont été touchés.	In the meantime, I should like to observe a minute' s silence, as a number of Members have requested, on behalf of all the victims concerned, particularly those of the terrible storms, in the various countries of the European Union.
Je vous invite à vous lever pour cette minute de silence.	Please rise, then, for this minute's silence.
Madame la Présidente, c'est une motion de procédure.	Madam President, on a point of order.
Vous avez probablement appris par la presse et par la télévision que plusieurs attentats à la bombe et crimes ont été perpétrés au Sri Lanka.	You will be aware from the press and television that there have been a number of bomb explosions and killings in Sri Lanka.
L'une des personnes qui vient d'être assassinée au Sri Lanka est M. Kumar Ponnambalam, qui avait rendu visite au Parlement européen il y a quelques mois à peine.	One of the people assassinated very recently in Sri Lanka was Mr Kumar Ponnambalam, who had visited the European Parliament just a few months ago.

Table 4.1: Section of the French-English Europarl parallel corpus.

world's language, including Turkish and Basque, place the verb after its object. Argument structure in English is conveyed using word order; in other languages, this information is conveyed primarily with verbal and nominal morphology, which may allow more flexibility in word order. A language might use multiple words to convey the same meaning that is expressed with only a single word in another language, or vice versa, resulting in many-to-one and one-to-many alignments. In addition, there can be words in one language that have no corresponding translation in another. Figure 4.1 provides a few examples

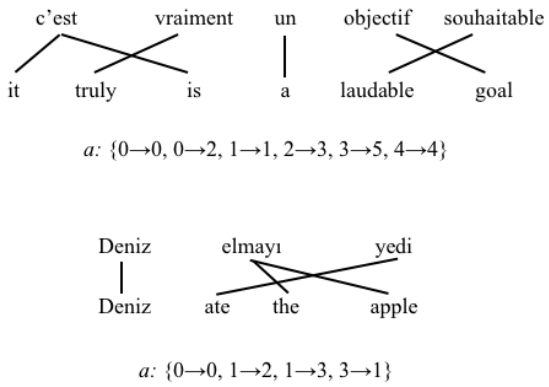

Figure 4.1: Word alignments showing word order differences in French and Turkish.

demonstrating the impracticality of using simple monotonic alignment between words in a pair of parallel sentences.

Statistical machine translation systems that rely on word-level alignments use these alignments to build a translation model. In this thesis, however, I will use word alignments to automatically determine the fidelity of a narrative retelling to the original source narrative, as I describe in Chapter 7. The features characterizing this fidelity that can be extracted from such an alignment include not only the score that is typically recorded for narrative recall (i.e., how many elements were recalled) but also the identities of the elements and their relative importance, the ordering of the elements, and the presence of off-topic or irrelevant speech, embellishments, and repetitions. These features can subsequently be used within a machine learning framework to determine the neurological status of the subject performing the retelling.

In this section, I review a few of the more widely used techniques for unsupervised word alignment used in machine translation research. I will also briefly discuss other alignment approaches proposed in the literature that are not commonly used in large-scale MT systems but may offer some utility for the task of aligning narrative retellings.

4.2.1 IBM Models

The IBM word-based translation models (Brown et al., 1993) form the basis for many of the word alignment approaches that are widely used today for statistical machine translation. These models were developed as a new data-driven approach for word-based machine translation at IBM during the revival of interest in machine translation in the late 1980s and early 1990s. The authors frame the problem of machine translation in terms of the noisy channel model, in which the goal is to recover a hidden input sequence given an output observation sequence. When translating from French to English, the output observation sequence is a sentence in French, f, and the hidden input sequence is a sentence in English, e. The process of translating a sentence is therefore finding the input sentence e with the highest probability, given the output sentence f. This conditional probability can be represented in the following way according to Bayes' rule:

$$\underset{e}{\operatorname{argmax}} \mathrm{P}(e \mid f) = \underset{e}{\operatorname{argmax}} \frac{\mathrm{P}(f \mid e)\ \mathrm{P}(e)}{\mathrm{P}(f)} \tag{4.1}$$

Since the probability of f itself is constant across all possible e, the equation can be rewritten as follows:

$$\underset{e}{\operatorname{argmax}} \mathrm{P}(e \mid f) = \underset{e}{\operatorname{argmax}} \mathrm{P}(f \mid e)\ \mathrm{P}(e) \tag{4.2}$$

The first term on the right-hand side of the equation represents the translation model probability, which is the conditional probability of the string f given the string e, or more precisely for this context, the probability that observed string f is a translation of the string e. The second term represents the language model; it models the probability of the string e according to what is known about its language. The IBM models for word alignment were developed in order to learn the word-to-word translation model from a parallel corpus for a particular language pair.

The best alignment between a source string and a target string might perhaps be most easily learned from a large set of alignments determined by a human speaker of both languages. It would be very laborious and difficult, however, to generate sufficient quantities of this sort of data for any arbitrary language pair. The IBM models, as they are implemented in commonly used word alignment packages such as Giza++ (Och and Ney,

2003), do not learn translation probabilities from such manual word alignments. The only information given is the target language translation for each source language sentence, which is presented in the form of a bilingual parallel corpus. From these sentence pairs, the IBM models use expectation maximization to learn the most probable alignments between words in the two languages. Each IBM model can be used as an initialization "stepping stone" to any of the following models.

IBM Model 1 is the simplest of the IBM models, and it is used to initialize the translation probabilities in both Giza++ (Och and Ney, 2003) and the Berkeley aligner (Liang et al., 2006), which are described in more detail in Sections 4.2.4 and 4.2.5. Model 1 estimates only the word-to-word translation probabilities from a parallel corpus. Let $\mathbf{f}$ be the source language string where f_j represents the word in string f at position j; let $\mathbf{e}$ be the target language string, where e_i represents the word in string e at position i and where e_0 represents the NULL alignment; and let $\mathbf{a}$ be the alignment mapping between the words in those two strings, where a_j is the position in e that aligns to word f_j:

$$\begin{aligned} \mathbf{f} = f_1^J &= f_1 \dots f_j \dots f_J \\ \mathbf{e} = e_0^I &= e_0 \dots e_i \dots e_I \\ \mathbf{a} = a_1^J &= a_1 \dots a_j \dots a_J \end{aligned}$$

We can determine the probability of the string $\mathbf{f}$ and a particular alignment $\mathbf{a}$, given the string $\mathbf{e}$, simply by multiplying the probability of the translation probabilities for the word pairings defined by that alignment:

$$\mathrm{P}(\mathbf{f}, \mathbf{a} \mid \mathbf{e}) = \prod_{j=1}^{J} p(f_j \mid e_{a_j}) \tag{4.3}$$

The probability of a source string f given a target string e is therefore derived by marginalizing over all possible alignments:

$$\begin{aligned} p(\mathbf{f} \mid \mathbf{e}) &= \sum_{\mathbf{a}} p(\mathbf{f}, \mathbf{a} \mid \mathbf{e}) \\ p(\mathbf{f} \mid \mathbf{e}) &= \sum_{\mathbf{a}} \prod_{j=1}^{J} p(f_j \mid e_{a_j}) \end{aligned}$$

IBM Model 1 provides a way to estimate those word-to-word translation probabilities from a parallel corpus by observing patterns of word cooccurrence. Model 1 typically begins by initializing all word-to-word translation probabilities uniformly. For each sentence pair in a parallel corpus, the expectation phase of Model 1 considers every alignment between a word in the source string and a word in the target string as a potential alignment pair. The probability for each of these word-to-word alignments in that sentence pair is normalized by the total probability in that sentence for the target word and then added to the total probability count for that word pair in the entire corpus. In the maximization phase of Model 1, these probability counts for each source-target word pair are normalized by the total probability count for the source word, thereby generating new translation probabilities for each source-target word pair. In theory this process should be repeated until convergence, but in practice most systems that rely on Model 1 allow the user to determine the number of iterations, which is usually larger than 1 but does not exceed 5. With these new translation probabilities, it is then possible to determine for any source-target sentence pair the most probable alignment via some decoding algorithm, such as Viterbi decoding.

IBM models 2 through 5 are incremental improvements upon the preceding models. Note that Model 1 does not restrict in any way the locations of the words it is aligning; a word at a particular position in the source string can align to any word at any position in the target string. Model 2 includes a second parameter to be estimated, namely *distortion*, which models the likelihood that a particular position in the source string will be aligned with a particular position in the target string. The goal of the version of Model 2 that is usually used (Och and Ney, 2003) is to encourage alignment along the diagonal so that a word in a particular position in the source string will be aligned to a word at around the same position in the target string. Model 3 introduces the idea of fertility, which is the notion that a single source word can generate (i.e., align to) multiple output words. Model 4 improves upon the alignment distortion probability developed in Model 2 to model relative rather than absolute distortion, and Model 5 repairs some deficiencies introduced in Model 3. The only IBM model I use in the work presented here is Model 1. I therefore refer the reader to Brown et al. (1993) for further details about Models 2 through 5.

4.2.2 HMMs for word alignment

Another obvious approach to determining word alignments is to use a Hidden Markov Model (HMM), which was first proposed by Vogel et al. (1996). In an HMM, two probabilities must be estimated: the emission (or translation) probability and the transition (or distortion) probability. The emission probability, $b_{f_j e_i}$, in a word alignment HMM is the translation probability: the probability that the word at position j in the source language string f is translated by (or will emit) the word at position i in the target language string, e. The transition probability, $d_{a_j a_{j-1}}$, is the probability of an alignment of word f_j given the alignment of the previous source word, f_{j-1}. Equation 4.2.1, which previously included only the translation probability, will now include this transition probability:

$$p(\mathbf{f} \mid \mathbf{e}) = \sum_{\mathbf{a}} \prod_{j=1}^{J} p(a_j \mid a_{j-1}, I) p(f_j \mid e_{a_j}) \tag{4.4}$$

In the context of translation, the actual numeric value of the position of a word in a string is not meaningful, since except in a few possible cases (perhaps German's requirement for the finite verb to appear in second position in a main clause), the underlying rules of human language do not enforce restrictions on numeric word position. Instead, what is important is the size and direction of the difference between the current position and the previous position; that is, we want to model the likelihood of each *jump width* in the alignment, rendering the above equation as:

$$p(\mathbf{f} \mid \mathbf{e}) = \sum_{\mathbf{a}} \prod_{j=1}^{J} p(a_j - a_{j-1}) p(f_j \mid e_{a_j}) \tag{4.5}$$

In this way, we can model the distortion or reordering of the words in the target that is defined by the alignment. We have seen that local word order variations across languages, such as the variation within a noun phrase of the order of noun and adjective, tend to be small. Global word order variations, however, can require very large jump widths, as is the case in SOV language relative to an SVO language like English. Using an HMM rather than simply IBM Model 1, we can model the probabilities for the size of the jumps that represent these types of word order variations. The emission and transition probabilities can be estimated using the Forward-Backward algorithm. The resulting probabilities can

then be used to determine the most probable alignment between a source language sentence and target language sentence using Viterbi decoding or another decoding approach, such as posterior decoding.

4.2.3 Alignment error rate (AER)

The quality of the output of a word alignment system is generally measured in terms of its alignment error rate (AER), following work in this area by Och and Ney (2000, 2003). The alignments produced are compared to a manually generated set of word alignments. The manual alignments between words can be labeled as *Sure* or simply *Possible*. Possible alignments are used to align idiomatic expressions, free translations, and missing function words, while the sure alignments are those alignments that are unambiguous. The sure alignments are used in the calculation of recall, while the possible alignments are used in the calculation of precision. If S is the set of sure alignments in the manual alignment, P is the set of possible alignment pairs in the manual alignment (including the set of S alignments), and A is the set of alignments proposed by the aligner, we can define recall, precision, and alignment error rate as follows:

$$\begin{aligned} Recall &= \frac{|A \cap S|}{|S|} \\ Precision &= \frac{|A \cap P|}{|A|} \\ AER &= 1 - \frac{|A \cap S| + |A \cap P|}{|A| + |S|} \times 100 \end{aligned} \tag{4.6}$$

Note that if all alignments are sure alignments and there are no alignments labeled possible, alignment error rate is equivalent to F-measure: $AER = 1 - F1$. Although this has become the standard measure of word alignment quality in the machine translation community, the connection between reductions in AER and improvements in translation quality is not entirely clear. For this reason, some researchers in the MT community argue against using AER as a measure of quality (Fraser and Marcu, 2007). In my research on using alignments to extract scoring information from retellings, however, I have found that

reductions in AER, as measured using the manual annotations and metrics just described, do correspond to improvements in scoring accuracy, as will be discussed in Chapter 7.

4.2.4 Giza++

Giza++ (Och and Ney, 2003), the most widely used word alignment package, implements all of the IBM models and an HMM. In its default configuration, Giza++ follows this ordering, with translation probabilities set uniformly for the first model, and each subsequent model's parameters initialized to be the output parameters of the previous model: IBM model 1, an HMM model, IBM model 3, and IBM model 4. When used in building a machine translation model, Giza++ is used to train an alignment model in each direction: source to target and target to source. The output word alignment for a particular sentence is usually some combination of the output of the two directional models for that sentence, such as the union or intersection. Och and Ney (2003) and Koehn et al. (2003) recommend expanding this alignment to allow many-to-one and one-to-many alignments by using a heuristic method of symmetrization, which allows additional alignment points between unaligned words to be added to the alignment.

4.2.5 Berkeley aligner

An alternative to Giza++ is the the Berkeley aligner (Liang et al., 2006). The Berkeley aligner implements IBM Model 1, IBM Model 2, and an HMM. In its default configuration, the Berkeley aligner initializes translation probabilities with Model 1 and uses these as initial probabilities for an HMM. The Berkeley aligner introduces an innovation that enables it to achieve a lower AER than Giza++. Instead of training two separate models, one from source to target and the other from target to source, and then taking the intersection of the resulting alignments, the Berkeley aligner uses a joint training approach that allows it to learn the probabilities from both directions simultaneously. The Berkeley aligner also uses posterior decoding instead of Viterbi decoding, which results in an additional slight improvement in AER over Giza++ and allows the user to select a posterior threshold according to the desired trade-off between recall and precision of alignment. Although the Berkeley aligner generates more accurate alignments than Giza++ in its

default configuration, Giza++ seems, for historical reasons, to be the preferred method for word alignment in the machine translation research community.

4.2.6 Other approaches

Numerous alternative techniques for generating word alignments for machine translation have been proposed. While some of these alternative techniques are independent of the generative approaches described above (Melamed, 2000; Tiedemann, 2003; Moore, 2005; Moore et al., 2006), many of them can be considered extensions to one of the IBM models or to the word alignment HMM. Quirk et al. (2007), in their work on extracting parallel fragments from comparable corpora, extend the standard generative word alignment models by modeling null word alignments within parallel sentence fragments separately from words that are generated monolingually, independently of the opposite side of the parallel corpus. Distinguishing words in one language that align to the NULL word in the other language from words generated monolingually with no corresponding words in the other language affords the extraction of parallel sentence fragments that do not contain spurious or superfluous words. Although I do not distinguish these two types of null alignments in my approach to word alignment, the alignment to NULL will play a very important role in the graph-based method for word alignment described in Section 7.7.

Other extensions to the IBM models and the word alignment HMM rely on externally derived linguistic features, such as dictionary and thesaurus entries, part-of-speech tags, similarity measures, syntactic information, and unigram frequency information. In some approaches, one of the existing alignment packages is modified so that it can take into account this sort of linguistic information (Toutanova et al., 2002; Berg-Kirkpatrick et al., 2010). Other word alignment approaches involve training a discriminative model with these linguistic features along with probabilities generated via one of the IBM models or an HMM (Liu et al., 2005; Fraser and Marcu, 2005; Taskar et al., 2005; Blunsom and Cohn, 2006; Niehues and Vogel, 2008; Riesa and Marcu, 2010). As discriminative approaches, these techniques additionally require a set of manually annotated word alignments for training their models.

In the work presented in this thesis, I appeal to many of the algorithms and approaches described above. In addition, although I do not explore discriminative techniques or using external resources, I do estimate a key probability from a small held-out set of manual alignments which I then use to improve the probabilities estimated by the generative models, both within the Berkeley aligner and within an aligner we developed specifically for this purpose. The primary novel contribution of this thesis to word alignment, however, is the use of graph-based methods, which I discuss in the following section.

4.3 Graph-based methods

Social networks, in which actors are connected to one another according to a set of relations, are often modeled as graphs in which the nodes represent the actors and the edges represent the relations between those actors. Paths, or *random walks* (Lovász, 1993), can be traced through such a graph of interconnected nodes in order to learn more about the relative importance or centrality of the nodes and their relationships to one another and to the graph as a whole. These random walks through the graph are Markov chains in that each step to a new node depends only on the current node. As a Markov process, a random walk on a graph will converge to a stationary probability distribution over the vertices, provided that the graph is irreducible and aperiodic. Random walks can be simulated with matrix operations, as will be discussed shortly.

Graph-based methods involving random walks have been used for various NLP applications, including information retrieval (Brin and Page, 1998; Page et al., 1999), extractive summarization (Erkan and Radev, 2004; Otterbacher et al., 2009), word-sense disambiguation (Mihalcea, 2005; Sinha and Mihalcea, 2007; Agirre and Soroa, 2009), and paraphrase induction (Kok and Brockett, 2010). Graph-based approaches have also been proposed for tasks such as verb clustering (Brew and Schulte im Walde, 2002), determining the correct point of attachment for prepositional phrases in a syntactic structure (Toutanova et al., 2004), and measuring word similarity (Minkov and Cohen, 2008). Below I discuss several of these systems in order to illustrate the flexibility and utility of random walks on graphs. In Chapter 7, I present a novel application of graph-based methods to improve

word alignment between a retelling and a source narrative, in which the nodes in the graph are words and the edges are represented by alignments.

4.3.1 PageRank

Probably the most successful and well-known application of these graph-based methods is PageRank, the algorithm developed by Google to rank web pages that are retrieved in response to a user query (Brin and Page, 1998; Page et al., 1999). The PageRank algorithm interprets the world wide web as a graph in which individual web pages are the nodes, and the hyperlinks between web pages are the edges connecting those nodes. Each node, or web page, has zero or more incoming edges, or hyperlinks from other web pages link to that web page, and zero or more outgoing edges, or hyperlinks that connect that web page to other web pages. Thus, in a graph-based interpretation of the world wide web, more important web pages are those that have many incoming links, since this presumably means that more web page creators found it worthwhile to link to that page. In addition, the higher the ranks of the web pages supplying those incoming links are, the higher the rank of that page will be. In this way, a web page inherits importance from pages that link to it and reassigns the importance it accumulates in this way to the pages connected to it through its outgoing links.

This framework models how a web surfer, starting at a random node in the graph that is the world wide web and following hyperlinks at random from web page to web page, will be more likely to end up on pages with more incoming links from other webpages, especially when those webpages have more incoming links. The simplified version of the PageRank rank score calculation is as follows, where $R(u)$ is the rank of web page u; B_u is the set of pages that link to u; $C(v)$ is the number of outgoing hyperlinks of v:

$$R(u) = \sum_{v \in B_u} \frac{R(v)}{C(v)} \tag{4.7}$$

Random walks on the graph can be simulated with matrix operations in order to determine the stationary distribution. We can render the above equation in matrix notation as follows:

$$p = B^T p \tag{4.8}$$

where B is a matrix in which each element (i, j) represents an outgoing edge from node i to page j, and p represents the stationary distribution, i.e., the probability of ending up at that node on a random walk. The power method can then be used to solve for this stationary distribution, p.

In order for the graph to converge to a stationary distribution, it needs to be aperiodic and irreducible. An aperiodic graph is a graph that has no inescapable cycles; an irriducible graph is a graph in which every node is connected to at least one other node in the graph. For this reason, the full PageRank algorithm additionally includes a small probability that the theoretical random web surfer will "get bored" and move directly to an unlinked web page. This allows a walk through the graph to escape a cycle, and it ensures that any node can be reached from any other node, thereby making the graph aperiodic and irreducible. If the probability of jumping to a random node is uniform over all web pages, this model can be expressed as follows. The rank of a web page u, $R(u)$, is:

$$R(u) = \frac{d}{N} + (1 - d) \sum_{v \in B_u} \frac{R(v)}{C(v)} \tag{4.9}$$

where N is the total number of web pages in the graph and d is the "damping factor", which sets this small probability of jumping to a random web page in the graph. This first term in the equation can be adjusted to present a broad view of the web, where any page on the web has an equal probability of being jumped to, or a focused exploration of the web, in which only certain pages are highly probable destinations for a random jump.

4.3.2 NLP applications of random walks

Random walks on graphs have been applied to a number of NLP tasks, including extractive automatic summarization, paraphrase generation, and word sense disambiguation. In some applications, the goal is to rank nodes according to their centrality, while other applications use random walks to establish the strength of the relationships between nodes. The implementation of the random walk framework in these applications includes both actual random walks and matrix operation simulation.

A popular approach for the task of extractive summarization is to find the sentences in a cluster of documents that are most central to that cluster, or salient, in order to

create a summary with the information most pertinent to that cluster. Ranking the sentences in terms of their centrality or salience is similar to ranking web pages according to their relevance, which allows the graph-based approach to be extended to the task of of automatic summarization. The LexRank (Erkan and Radev, 2004) and Biased LexRank (Otterbacher et al., 2009) algorithms parallel the PageRank algorithm but apply instead to sentences in a document cluster.

Each node in the graph is a sentence in a cluster of documents. The edges between nodes are the similarity relationships between those sentences, as measured by some lexical or semantic similarity metric, such as a bag-of-words cosine similarity. Edges are included in the graph only if their similarity measure exceeds a certain threshold. Again, there is a damping factor, d that determines whether the walk will follow an edge or jump to a random node. The centrality of a sentence in LexRank is calculated in much the same way as the rank of a web page in PageRank:

$$LR(u) = \frac{d}{N} + (1-d) \sum_{v \in adj[u]} \frac{w(v,u)}{\sum_{z \in adj[v]} w(v,z)} LR(v) \tag{4.10}$$

where $adj[u]$ is the set of sentences with high enough similarity scores to be adjacent (i.e., linked) to u; $w(u,v)$ is the similarity score of u and v; and d is the damping factor determining whether the next move will be to a random node or a linked node. Both Erkan and Radev (2004) and Otterbacher et al. (2009) simulate random walks using matrix operations, as described above.

In the Biased LexRank formulation, the first term can be adjusted to bias the random jumps to sentences that are good matches to the topic or query that retrieved the set of documents to be summarized. If $b(u)$ is a score representing this bias toward the query, and C is the totally number sentences in the graph, then biased LexRank is:

$$LR(u) = d\frac{b(u)}{sum_{z \in C} b(z)} + (1-d) \sum_{v \in adj[u]} \frac{w(v,u)}{\sum_{z \in adj[v]} w(v,z)} LR(v) \tag{4.11}$$

Graph-based methods have also recently been applied to the task of word sense disambiguation (WSD) (Mihalcea, 2005; Sinha and Mihalcea, 2007; Agirre and Soroa, 2009; Navigli and Lapata, 2010). The goal in word sense disambiguation is to select, for a particular word in a string, the correct word sense given the context of that string. These

word senses are typically defined as WordNet synsets or structures from a similar lexical ontological resource. In a graph-based approach to WSD, a graph is constructed from a set of nodes representing the possible word senses for each content word in the string. The edges between these nodes can be derived in a number of ways. Mihalcea (2005) and Sinha and Mihalcea (2007) use various measures of pairwise semantic similarity between the definitions for each of the senses in the graph to determine the directions and weights of the edges. Alternatively, the edges can be determined from the implicit and explicit structure of the lexical resource itself (Agirre and Soroa, 2009; Navigli and Lapata, 2010). Some of the more recent work is focused on investigating centrality measures other than the eigenvector centrality of PageRank for exploring the graph space, including measures that consider additional features that can be extracted from the graph, such as the number outgoing links and the average path length from a node to all other nodes (Sinha and Mihalcea, 2007; Navigli and Lapata, 2010).

Relatively recently, random walks have been applied to the task of learning paraphrases from parallel corpora. Kok and Brockett (2010) use a graph-based method to extend the pivot-based approach for deriving paraphrases from a parellel corpus outlined in Bannard and Callison-Burch (2005); Callison-Burch et al. (2006); Marton et al. (2009) and Wang and Callison-Burch (2011). The graph in this case consists of nodes that are phrases extracted from MT phrase tables and edges that are the correspondences between these phrases in the phrase tables, weighted by their translation probabilities. New paraphrases are discovered by sampling random walks through this graph, where the length of the path between a phrase and a potential paraphrase can be used to estimate the strength of that paraphrase. In addition, the authors include "feature nodes" that allow domain knowledge, such as syntactic categories, n-grams, and substring information, to connect potential paraphrases from the same side of the phrase table. In the next chapter, I will present a graph-based approach for improving word alignments which, like the approach used in Kok and Brockett (2010), leverages the output of techniques normally used in machine translation and relies on sampling random walks over graphs. The nodes in these graphs represent words, while the edges represent the alignments between these words

proposed by an EM-based alignment algorithm. A detailed description of this approach appears in Section 7.7.

4.4 Summary

In this final background chapter, I discussed previous work in three areas of research that will play a crucial role in the research I will be presenting in the following chapters. The techniques I have developed for scoring and analyzing narrative retellings depend on deriving high-quality word alignments between the retellings and the source narrative. The available word alignment techniques, however, are designed for large bilingual corpora rather than a small monolingual corpus in which the the source language is a single narrative and the target language is the speech of dozens or hundreds of individuals, many of whom are neurocompromised. In the coming chapters, I will discuss the techniques I have developed for tailoring word alignment to these sorts of corpora, and how scores and other narrative fidelity features can be extracted from these word alignments and used for diagnostic classification.

Chapter 5

Data

In this chapter I describe the two data sets analyzed in this thesis: (1) Wechsler Logical Memory responses from seniors with and without mild cognitive impairment, collected at OHSU's Layton Aging and Alzheimer's Disease Center; and (2) NEPSY Narrative Memory responses from children with autism spectrum disorder, language impairment, and typical development, collected at the Center for Spoken Language Understanding at OHSU. Each of the following sections provides information about the neurological test in question, including a discussion of the administration, scoring procedures, reliability, and utility of the test, followed by information about the subjects, diagnostic criteria, and data collection procedure.

5.1 Wechsler Logical Memory

Although I apply the techniques described in this thesis to several different neuropsychological instruments, my investigation of narrative fidelity for automated neuropsychological assessment began with the Wechsler Logical Memory subtest. The Wechsler Logical Memory subtest is part of the Wechsler Memory Scale (Wechsler, 1945, 1987, 1997; Wechsler et al., 2009), a diagnostic instrument used to assess memory and cognition in adults. In the decades since it was first published, the Wechsler Memory Scale (WMS) has been updated several times in response both to suggestions for improvement from the psychological assessment community and to changing societal and cultural norms.

The current version of the WMS, WMS-IV (Wechsler et al., 2009), includes activities that test visual memory and verbal/auditory memory in an immediate and a delayed

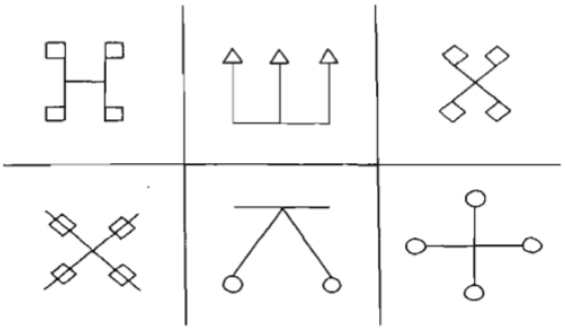

Figure 5.1: Illustration of geometric designs for the Visual Reproduction subtest.

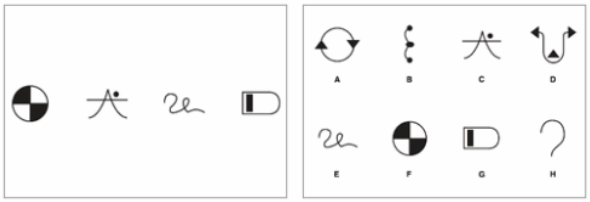

Figure 5.2: Illustration of Symbol Span subtest.

context. In the version of the test currently administered to the elderly population, the visual memory tasks include the Visual Reproduction subtest, in which the subject must select a previously seen geometric design from a set of designs (Figure 5.1), and the Symbol Span subtest, in which the subject is shown a sequence of symbols and then must select the symbols in the correct order from a large set of unordered symbols and foils (Figure 5.2). There are two verbal memory tasks: the Paired Associates subtest, in which the subject must recall pairs of semantically associated and novel words, and the Wechsler Logical Memory subtest, which is the subtest of interest of this dissertation.

The Wechsler Logical Memory (WLM) subtest is a narrative recall task. In the WLM, described in detail below, the subject listens to the examiner read a brief narrative and then retells the narrative to the examiner twice: once immediately and a second time after a delay of 20 to 30 minutes. The current version of the WMS includes two Logical Memory narratives, one of which is the Anna Thompson narrative, which is the narrative used in the research presented here. Despite numerous changes to the content of the WMS

over the almost 70 years since it was first released, the WLM subtest has been included in every edition of the WMS, and the Anna Thompson story in particular has appeared with only slight modifications in every edition of the WMS.

5.1.1 Administration

In the administration of the WLM used in this study, the examiner introduces the task by saying "I am going to read you a little story. Listen carefully, and when I am done, I want you to tell me everything you can remember." The subject listens to the examiner read a brief narrative, shown in Figure 5.3. The examiner then says: "Now begin at the beginning and tell me everything you can remember from the story." The subject must retell the narrative to the examiner. This first retelling is the immediate recall portion of the WLM, referred to here as Logical Memory I or LM-I. If the subject pauses for more than a few seconds, the examiner then says "Anything else?" When the subject has provided all the information he is able to provide, the examiner then says "Later on I will ask you to tell me this story again, so try not to forget it."

After approximately 20 minutes of unrelated activities, the delayed recall portion of the WLM, referred to here as Logical Memory II or LM-II, takes place. The examiner introduces the task by saying "A few minutes ago, I read you a little story. It was about a woman who was robbed. Tell me everything you can remember about that story." The subject then recalls as much as he can remember about the story. Again, when the subject pauses for longer than a few seconds, the examiner asks "Anything else?" allowing the subject to provide any further information he is able to provide.

There are no guidelines about the speed at which the story should be read. In addition, the guidelines do not specify whether any particular intonation or emphasis should be used. Some research suggests that speaking rate can have an impact on a listener's ability to process and recall the information, and for this reason, some practitioners recommend playing a recording of the narrative rather than reading the story aloud (Shum et al., 1997). In the data used in this study, however, the story was read aloud by the examiner rather than presented as an audio recording.

5.1.2 Scoring

The narrative is divided into 25 *story elements*. In Figure 5.3, the boundaries between story elements are denoted by slashes. While the subject is retelling the story, the examiner may either (1) transcribe, in real time, what the subject says and score the retelling afterwards from that transcript, or (2) note on the published WLM scoresheet, in real time, which story elements the subject uses. In the administration of the WLM used for this study, the examiner followed the latter procedure. Although the identities of the elements are noted on the scoresheet, the final score that is reported under standard administration of the task is a *summary score*, which is simply the raw number of story elements recalled per retelling. Story elements do not need to be recalled in the correct temporal order or the order in which they appeared in the story. In addition, some paraphrasing of elements is allowed. The published scoring guidelines describe the permissible substitutions for each story element, shown in Table 5.1. Some lexical substitutions and paraphrases are allowed: the first story element, *Anna*, for instance, can be replaced in the retelling with any variant of the name, such as *Annie* or *Ann*, while the 16th story element, *fifty-six dollars*, can be replaced with any number of dollars between fifty and fifty-nine. Other elements, however, must be recalled verbatim, such as the second element *Thompson* and the eighth element *cafeteria*.

An example LM-I retelling from our data, which will be discussed at length in Section 5.1.5, is shown in Figure 5.4. The correctly recalled story elements are in bold. According to the published scoring guidelines, this retelling receives a score of 12, since it contains the following 12 elements: *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*. We see in this example that the subject's replacement of *Thompson* with "Taylor" is disregarded, while the phrase "was sympathetic" counts as a match for the story element *touched by the woman's story*. The ordering of the story elements in the retelling, which is somewhat different from that of the source narrative, has no impact on the final summary score. Additionally, the subject is not penalized for including information in his retelling

> 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 St. / 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.

Figure 5.3: Text of WLM-III/IV 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.

Figure 5.4: Sample WLM retelling by a subject before MCI diagnosis (score=12).

> She was **robbed**. And she had a couple **children** to feed. She **had no food for them**. And people **made a collection for her** and to pay for her, for the food for the children.

Figure 5.5: Sample WLM retelling by same subject after MCI diagnosis (score=5).

that does not correspond to elements in the story, such as "Is that right?" and "so that she can feed the children".

5.1.3 Reliability and Utility

The Anna Thompson story appeared in the original formulation of the Wechsler Memory Scale from 1945 (Wechsler, 1945), but it was first used to test the memory and cognition of veterans returning from combat in World War I (Yerkes, 1921). The text of the original story used by Yerkes is shown in Figure 5.6, along with the 1945, 1987, and 1997/2009 Wechsler versions of the story. We see that the gist of the story has remained unchanged while the language and details have been updated to reflect contemporary usage and expectations. (Despite these modernizations, many subjects in our study have pointed out that fifty-six dollars would be insufficient to pay the rent and feed the children. In addition, a few subjects wondered why, if she worked in a cafeteria, she was unable to bring home food for her children.)

The boundaries between the story elements have changed since the WMS was first published, as have the guidelines for scoring the elements. The 1945 WLM scoring instructions ("record verbatim and score according to the number of ideas as marked off

Story Element	**Scoring Criteria**
Anna	Anna or variant of the name
Thompson	Thompson is required
of South	South in any context
Boston	Boston in any context
employed	indication that she held a job
as a cook	cook or some form of the word is required
in a school	school is required
cafeteria	cafeteria is required
reported	indication that a formal statement was made to someone in authority (in any context)
at the police	police in any context
station	station (in any context) or a word or phrase denoting a police station
that she had been held up	indication that she had been held up (i.e., gunpoint or knife)
on State Street	State Street (in any context)
the night before	indication that the hold-up occurred the previous night
and robbed of	indication that a robbery took place
fifty-six dollars.	indication that an amount of money greater than $49 but less than $60 was taken from her
She had four	four is required together with an indication that the children were hers
small children,	children or a synonym is required
the rent was due,	a phrase indicating that the rent was due
and they had not eaten	indication that her children or the family were without food
for two days.	two days is required or a phrase meaning about two days
The police,	a word or phrase indicating one or more members of the police department
touched by the woman's story.	indication that her story evoked sympathy
took up a collection	a phrase indicating that money was collected
for her.	indication that the money collected was for her or her children

Table 5.1: Published scoring guidelines for WLM-III Anna Thompson story.

in selection" (Wechsler, 1945)) did not specify the degree of paraphrasing that should be tolerated, which led to poor inter-rater reliability and agreement (Prigatano, 1978; McCarty et al., 1980; Crosson et al., 1984; Mitchell, 1987). Subsequent versions of the test included more explicit guidelines for scoring the story elements, ultimately resulting in the guidelines presented in Table 5.1. In the intervening years, however, numerous scoring alternatives were proposed with the sometimes conflicting goals of improving inter-rater

> Yerkes (1921): Anna Thompson of South Boston, employed as a scrub woman in an office building, reported at the City Hall Station that she had been held up on State Street the night before and robbed of about five dollars. She had four little children and the rent was due. The officers made up a purse for her.
>
> Wechsler Memory Scale (1945): Anna Thompson of South Boston, employed as a scrubwoman in an office building, reported at the City Hall station that she had been held up the night before on State Street and robbed of fifteen dollars. She had four little children, the rent was due, and they had not eaten for two days. The officers, touched by the woman's story, made up a purse for her.
>
> Wechsler Memory Scale - Revised (1987): Anna Thompson of South Boston, employed as a cook in a school cafeteria, reported at the City Hall station that she had been held up the night before on State Street and robbed of fifty-six dollars. She had four small children, the rent was due, and they had not eaten for two days. The officers, touched by the woman's story took up a collection for her.
>
> Wechlser Memory Scale III (1997) and IV (2009): Anna Thompson of South Boston, employed as cook in a school cafeteria, reported at the police station that she had been held up the night before on State Street and robbed of fifty-six dollars. She had four small children, the rent was due, and they had not eaten for two days. The police, touched by the woman's story, took up a collection for her.

Figure 5.6: Evolution of the Anna Thompson story.

agreement and test-retest consistency on the one hand, and distinguishing more precisely between diagnostic groups on the other.

Before the guidelines were made more explicit for the revised version of the WMS (WMS-R) (Wechsler, 1987), a number of researchers argued that elements should be scored according to whether they were recalled verbatim; that is, recalled elements using the exact wording of the source narrative should contribute more to the summary score than elements recalled in a non-verbatim or "gist" fashion. The strictest scoring criteria, which resulted in the highest inter-rater reliability, required exact verbatim recall (Abikoff et al., 1987). Other approaches awarded different point values depending on the extent of the verbatim recall of an element. Power et al. (1979) assigned a full point to verbatim responses but a half-point to non-verbatim gist responses, while Haaland et al. (1983) awarded a half-point to incomplete verbatim responses. In an approach proposed by Russell (1975) and expanded by Johnson et al. (2003), recall was scored in a loose verbatim fashion, which Johnson terms *veridical*, in which variations in morphology and substitutions, deletions,

and insertions of function words are tolerated. The scoring approach that was eventually chosen for the WMS-R and expanded for the WMS-III combined the scoring guidelines of Russell (1975) with those of Schwartz and Ivnik (1980), who outlined specific permissible gist-level substitutions for each story element. Another important contribution to improving the administration and scoring of the WMS was the introduction of a delayed recall portions to some of the subtests. Russell (1975) demonstrated the importance of delayed recall of the WLM and the Visual Reproduction subtest for detecting specific types of memory deficits. Delayed recall of the WLM has been included in the WMS since the revised version was released in 1987.

Although the published guidelines have an air of arbitrariness, in that verbatim or veridical recall is sometimes required and sometimes not, they do allow the test to be scored with very high inter-rater reliability, which was not possible under the scoring instructions outlined in the original 1945 version of the subtest (Mitchell, 1987). Inter-rater agreement on the summary scores, as measured by the Pearson product-moment correlation coefficient, has been reported to range from as low as 0.87 (McCarty et al., 1980) to as high as 0.99 (Wechsler, 1997), with most researchers reporting r in the mid to high nineties (Haaland et al., 1983; Abikoff et al., 1987; Woloszyn et al., 1993; Sullivan, 1996; Johnson et al., 2003). Because the identities of the individual story elements are not reported in standard scoring, inter-rater reliability as measured by Cohen's kappa is not often reported in the literature, but Johnson et al. (2003), in their investigation of veridical scoring, reports an average kappa of 0.87, with a range of 0.60 to 1.00.

5.1.4 Diagnostic Utility

Summary scores on the WLM have long been reported to distinguish typical aging from Alzheimer's disease and other dementias (Logue and Wyrick, 1979; Brinkman et al., 1983; Storandt et al., 1984; Storandt and Hill, 1989). In addition, alternative scoring techniques, such as those described in Chapter 2 have demonstrated sensitivity of the test to mild and very mild dementia (Dunn et al., 2002; Johnson et al., 2003). Scores on the WLM subtest have been found to correlate both with age and years of education (Crosson et al., 1984; Haaland et al., 1983), although the decline in performance associated with age stabilizes

after the age of 75 (Haaland et al., 1983). For this reason, it is important to make sure that diagnostic groups are matched for both of these features. In addition, it has been recently noted that the performance of women on the WLM follows a very different pattern from that of men (Chapman et al., 2011), which suggests that comparisons of diagnostic utility across gender are necessary when using the WLM.

5.1.5 Data collection

The data examined in this study was collected from participants in an NIA-funded longitudinal study on brain aging at the Layton Aging and Alzheimer's Disease Center at the Oregon Health and Science University. The WLM subtest of WMS-III was administered to each of the experimental subjects as part of an hour-long interview and set of structured activities designed to elicit responses that can be used to assess cognitive function. The audio of the hour-long session for each subject was recorded using a stationary microphone attached to either a laptop or a digital recorder. From these recordings, the segments corresponding to the two full WLM retellings, typically ranging between 30 and 60 seconds, were extracted for each subject. The recordings were sometimes made in an informal setting, such as the subject's home or a senior center, but efforts were made to reduce the number of distractions and the amount of extraneous noise for all participants.

The WLM immediate and delayed retellings for 261 participants in the Layton Center study were transcribed at the word level in three batches by three different skilled labelers according to the rules outlined in the 2004 EARS Official Annotation Guidelines for conversational speech (Strassel, 2004). The words and phrases corresponding to each story element in each retelling were manually identified by a skilled labeler according to published guidelines shown in Table 5.1. I manually verified these annotations, and from these annotation, element-level and summary-level scores were calculated. Agreement between the summary scores assigned by me from transcripts and those assigned in real time during test administration, as measured by the Pearson product-moment correlation coefficient, was 0.97. This value is well within the range of agreement reported in the literature.

Because the algorithmic contributions of this thesis are related to improvements in monolingual word alignment, which in turn are shown to improve automated scoring and diagnostic classification accuracy, I also manually produced word-level alignments between each retelling and the source narrative presented in Figure 5.3. Selecting the "correct" alignment for a particular word can be difficult and subjective. Following Och and Ney (2003), I manually aligned words using the sure/probable system often used in machine translation, in which alignments between words can be labeled S (sure) or simply P (probable). These word alignments were used as a gold standard against which to evaluate the output of the various automatic word alignment techniques explored in this thesis. Extensive details on word-level alignment algorithms and evaluation are given in Chapter 7.

5.1.6 Mild Cognitive Impairment

The neurological disorder I will be investigating with the Wechsler Logical Memory test is Mild Cognitive Impairment (MCI), the stage of cognitive decline between the sort of decline expected in typical aging and the decline associated with dementia or Alzheimer's disease (Petersen et al., 1999; Ritchie and Touchon, 2000; Petersen, 2011). MCI is characterized by subtle deficits in functions of memory and cognition that are significant but do not prevent carrying out the activities of daily life. This intermediary phrase of decline has been identified and named numerous times: mild cognitive decline, mild neurocognitive decline, very mild dementia, isolated memory impairment, questionable dementia, and incipient dementia. Although there continues to be disagreement about the diagnostic validity of the designation (Ritchie and Touchon, 2000; Ritchie et al., 2001), a number of very recent studies have found evidence that seniors with some subtypes of MCI are significantly more likely to develop dementia than the population as a whole (Busse et al., 2006; Plassman et al., 2008; Manly et al., 2008). (For a more extensive discussion of this disorder and these phenomena, please see Section 2.3.1). For this reason, there is increasing interest in finding ways to quickly, easily, objectively, and unobtrusively determine whether a patient has MCI. Early detection can benefit both patients and researchers investigating treatments for halting or slowing the progression of dementia. Identifying MCI, however,

can be problematic, as most dementia screening instruments, such as the Mini-Mental State Exam (MMSE) (Folstein et al., 1975), lack sufficient sensitivity to the very subtle cognitive deficits that characterize the disorder (Morris et al., 2001; Ravaglia et al., 2005; Hoops et al., 2009). Diagnosis of MCI therefore typically requires both a lengthy neuropsychological evaluation of the patient and an interview with a family member or close associate that should be repeated at regular intervals in order to have a baseline for future comparison.

In this thesis, we define MCI using the Clinical Dementia Rating (CDR) scale (Morris, 1993; Morris et al., 1997), following earlier work on MCI (Petersen et al., 1999; Morris et al., 2001), as well as the work of Shankle et al. (2005) and Roark et al. (2011), who have attempted differential diagnosis using neuropsychological instrument subtest responses. The CDR is a numeric staging scale that indicates the presence of dementia and the level of its severity. The scale ranges from 0, indicating the absence of dementia, to 3, signifying severe dementia. A CDR of 0.5 corresponds to MCI (Petersen et al., 1999; Morris et al., 2001; Shankle et al., 2005; Roark et al., 2011). The CDR is determined via a semi-structured interview with the individual and a family member or caregiver that allows the examiner to assess the individual in six key areas of cognitive function: memory, orientation, judgement and problem solving, community affairs, home and hobbies, and personal care. Although it is a subjective measure, the CDR has high inter-annotator reliability when conducted by trained experts, with kappa values ranging from mid seventies to mid eighties (Burke et al., 1988; McCulla et al., 1989; Tractenberg et al., 2001). It is important to note that the calculation of CDR is completely independent of the neuropsychological test investigated in this thesis, the Wechsler Logical Memory subtest of the Wechsler Memory Scale.

5.1.7 Experimental subjects

Out of the set of 261 participants in the Layton Center study whose retellings were transcribed, there were 72 subjects with MCI (CDR=0.5) and 163 typically aging seniors (CDR=0). Table 5.2 shows the mean age and mean years of education for the two diagnostic groups. There were no significant between-group differences in either measure.

Diagnosis	n	Age	Education
MCI	72	88.7	14.9 yr
Non-MCI	163	87.3	15.1 yr

Table 5.2: Layton Center subject demographic data.

The remaining 26 subjects whose retellings were transcribed had either an unconfirmed diagnosis at the time this thesis was written or did not meet the eligibility criteria for this study.

5.2 NEPSY Narrative Memory

The second narrative recall instrument I investigate is the Narrative Memory subtest of the NEPSY (Korkman et al., 1998). The NEPSY is a large and comprehensive battery of tasks that test neurocognitive functioning in children over six separate domains: attention and executive functioning, language, memory and learning, sensorimotor ability, social perception, and visuospatial processing. Like the Wechsler Logical Memory subtest, the NEPSY Narrative Memory (NNM) subtest is a narrative recall task. The examiner reads a brief story to the child, who must then verbally retell the story to the examiner. There are a number of differences, however, between the NNM and the WLM in terms of administration and scoring, which will be discussed in the following two sections. Although the NEPSY is commonly used in schools and community settings in order to identify specific deficits in children with developmental disabilities, it is a relatively new instrument and has therefore not been as widely or thoroughly studied as the Wechsler Memory Scale.

5.2.1 Administration

In the 1998 formulation of the NNM, which is the version of the test administered in our study, the examiner begins by saying "I am going to read you a story. Listen carefully so you can tell me the story when I am finished." The examiner reads aloud the NNM narrative, which is shown in Figure 5.7, then says "Now you tell me the story." If the child hesitates or seems to have difficulty beginning his retelling, the examiner can say "How did the story start?" If the child does not finish the story, the examiner can say "Tell

> Jim was a boy whose best friend was Pepper. Pepper was a big black dog. Jim liked to walk in the woods and climb the trees. Near Jim's house was a very tall oak tree with branches so high that he couldn't reach them. Jim always wanted to climb that tree, so one day he took a ladder from home and carried it to the oak tree. He climbed up, sat on a branch, and looked out over his neighborhood. When he started to get down, his foot slipped, his shoe fell off, and the ladder fell to the ground. Jim held onto a branch so he didn't fall, but he couldn't get down. Pepper sat below the tree and barked. Suddenly Pepper took Jim's shoe in his mouth and ran away. Jim felt sad. Didn't his friend want to stay with him when he was in trouble? Pepper took the shoe to Anna, Jim's sister. He barked and barked. Finally, Anna understood that Jim was in trouble. She followed Pepper to the tree where Jim was stuck. Anna put the ladder up and rescued Jim. Wasn't Pepper a smart dog?

Figure 5.7: Text of NEPSY narrative.

me more" or "What happened next?" The examiner may prompt for information three times. When the child can supply no further information, the examiner moves from the Free Recall portion of the test to the Cued Recall section, in which the examiner asks the child questions about the parts of the story that he was unable to recall. The responses from the Cued Recall section are not analyzed in the work presented in this thesis. There is no delayed recall task. Examples of retellings from children in our study are provided in Figure 5.11.

5.2.2 Scoring

The guidelines for scoring the NNM are significantly more subjective than those currently used to score the WLM. The instructions in the manual are simply: "The child's formulation need not be verbatim from the story, but it must capture the essential information of the details listed in the Record Form." The Record Form is a scoresheet included with the NEPSY materials that lists the key story elements that must be recalled and provides, for some elements, a few example paraphrases that may be used. The examiner typically scores the NNM in real time using the Record Form, though our scoring procedure is more rigorous, as described in Section 5.2.4. The information contained in the Record Form is reproduced in Figure 5.3. The scores reported are a summary score for the Free Recall section, a summary score for the Cued Recall section, and a summary score derived from

both sections, in which the child receives two points for every element recalled in the Free Recall section and one point for every element recalled in the Cued Recall section.

Problems with the scoring procedure

The scoring procedure of the NNM diverges significantly from that of the WLM. Recall that the scoring procedure of the WLM considers every word of the source narrative to be part of a story element. In the NNM, only certain parts of the story are considered worthy of being recalled, as can be seen on examining the list of story elements shown in Figure 5.3. Very important parts of the story are excluded from the story element list, including *Jim always wanted to climb that tree*, *sat on a branch*, *Jim held onto a branch*, *Pepper took Jim's shoe in his mouth*, *he was in trouble*, *Finally Anna understood that Jim was in trouble*, and *Wasn't Pepper a smart dog?*. Thus, a child who recalls every detail of the story, even those that are not included in the Record Form, is not distinguishable from a child who recalls only the listed elements. Furthermore, a child who, for whatever reason, focuses on these parts of the story, is penalized for not remembering the seemingly arbitrary list of elements from the Record Form. The most notable omission from the list of elements is the word *dog*. As I will discuss at length in Chapter 8, the omission of these parts of the narrative from the list of required story elements becomes important when using the test to differentiate subgroups of autism spectrum disorder and language impairment.

We also observe that there are no guidelines in the Record Form or in the NEPSY manual for acceptable substitutions or paraphrases other than the few examples given. In our use of the NNM, this has led to some scoring disagreement among raters. Our clinicians did not agree, for instance, on whether the element *barked and barked* should be counted as correctly recalled if the child simply said "barked", "barked a lot", or "kept barking". There was also uncertainty about whether to score *climbed the tree* as correctly recalled if the child said that Jim had "climbed the ladder" and then indicated that Jim was sitting on a branch, thereby implying that he had used the ladder to climb the tree. In fact, the wording of the part of the story corresponding to this element, *climbed up*, is entirely ambiguous.

Another weakness in the Record Form is that story elements listed do not always use the same wording and syntax that was used in the source narrative. The element *branches too high for Jim to reach* has an entirely different syntactic structure from the expression used in source narrative: *branches so high that he couldn't reach them.* The elements are also not presented on the Record Form in the order in which they are presented in the source narrative: *climbed the tree* appears before *got a ladder* in the story element list, while *Anna* appears in the story element list before *took her Jim's shoe.* This reordering also seemed to make scoring the retellings more difficult.

Perhaps the most troubling feature of the Record Form is that some of the story elements capture multiple distinct events in the story. Element number 10 is particularly egregious, since it includes an entire series of events within a single element, each of which is dependent on the previous event. In the scoring procedure outlined here, it is sufficient for a child to say that Jim slipped without ever mentioning that the ladder fell or that Jim couldn't get down. Collapsing these multiple events into a single story element means that a child is not rewarded for recalling more information, even when this information is what indicates that he actually comprehended the gist of the story. Again, the insensitivity of the standard scoring procedure to these differences in narrative recall will become important when the retellings are used to differentiate diagnostic subgroups.

We note that the authors of the NEPSY do not provide any internal or external justification for the selection of story elements included in the Record Form. Narrative recall and narrative generation tasks are widely used in neuropsychological assessment, but there is no indication in the NEPSY materials or in any publications by the authors that earlier work in this area was considered during the development of this subtest.

5.2.3 Utility and Reliability

Although the score on the NNM subtest is included in the index score for the memory and learning domain of the NEPSY, it is reportedly more strongly correlated with language measures and verbal IQ (Korkman et al., 1998; Schmitt and Wodrich, 2004; Titley and D'Amato, 2008). In fact, the data collected at CSLU (see Section 5.2.4) show the same pattern: the NEPSY Narrative Memory score correlates most highly with measures of

I remember that Jim was a boy and his best friend was a big black dog Pepper. And Jim, he liked to go in the woods. And he loved to climb trees. And there's a tree in his neighborhood. It was a really tall oak tree. And the branches were too high and he couldn't reach them. Until one day he took a ladder, and he put the ladder up there and he climbed up. And he got on the branch, and he looked over his whole neighborhood. And he started to climb back down, and his foot flipped and he lost his shoe. And he was holding on to a branch. And then Pepper the dog. I think her name was Pepper. Pepper the dog sat below the tree barking. And then she grabbed his shoe and he or she ran away. And Jim was sad. He was lonely. He didn't know why his best friend didn't want to be with him when he was in trouble. And meanwhile Pepper the dog was at the house barking at his sister Anna. I have a girl named Anna in my class. And it took a long time but finally Anna understood, and she followed Pepper, and Pepper led Anna out to the oak tree. And Anna took the ladder, and she put it up, and she saved Jim. And that was the end of the story.

Figure 5.8: Sample retelling from a child with typical development (score=15).

Well there was a guy named Jim and he had his best friend is Pepper the dog. And I can't remember a little bit but think there was a tree so big that he like couldn't reach the branches or something. And then like so he always wanted to climb that tree so got a ladder from his house and stuck on there and put on there. Bow, I mean bam! Shake shake grew. And he started climbing up and his foot slipped and his shoe fell off and then the ladder fell. He was sad because he couldn't get down. And Pepper ran with the shoe to a girl in the village, I think. And then she came and rescued him. The dog took his shoe and went to her and finally she understood what he meant.

Figure 5.9: Sample retelling from a child with ASD without LI (score=8).

He had a friend named Pepper. Pepper was a black dog. Pepper. I forgot. Pepper got his shoe. I don't know. Jim was a little boy. Pepper was his friend. Pepper was a black dog, and Pepper rescued his shoe when he brought it to Hannah. That 's all I know. And then they, then Hannah rescued him.

Figure 5.10: Sample retelling from a child with SLI (score=5).

The way he go down and hurt himself. His shoe fell off. And the ladder go down to the ground. The boy took a picture of the girl. And he stopped taking a picture. And he was about to walk to the best thing. He went off to the zoo, and the girl went on with us, too because she went to the zoo. She sold lots of animals, and the boy sells lots of animals, too.

Figure 5.11: Sample retelling from a child with ASD and LI (score=1).

Story Element
1. Jim
2. Pepper
3. big
4. black
5. liked to walk in the woods *or* climb trees
6. tree/oak with branches too high for Jim to reach
7. climbed the tree/oak
8. got a ladder *or* carried a ladder to the tree/oak
9. looked out over the neighborhood *or* looked around
10. slipped *or* shoe fell *or* ladder fell *or* got stuck *or* couldn't get down
11. Pepper ran for help *or* went to get help *or* ran away
12. Jim was sad *or* thought Pepper didn't want to stay
13. Anna
14. Jim's sister
15. took her Jim's shoe
16. barked and barked
17. Anna put the ladder back up *or* rescued Jim *or* helped Jim

Table 5.3: Story element list for the NNM narrative.

semantic and verbal fluency (e.g., name as many foods as possible in 60 seconds) and language production (e.g., use a particular word or phrase in a sentence). The authors of the NEPSY indicate that although they have chosen to include the NNM in the memory domain of their instrument, poor performance on the NNM is likely related to language deficits rather than deficits in memory (Korkman et al., 1998).

Inter-rater reliability for the NNM has not been reported in any research or in the administration materials for the NEPSY or the NEPSY-II, to our knowledge. The authors of the NEPSY report inter-rater reliability only for the tests whose scoring they believe is subjective, a group that does not include the NNM. Split-half reliability estimates, in which scores on the first half of the subtest are correlated with scores on the second half of the subtest, range between 0.68 and 0.86, depending on the age group. Test-retest correlations ranged from 0.29 to 0.84, with the 7- and 8-year-old age group having the weakest correlation (Korkman et al., 1998).

5.2.4 Data collection

The data examined in this study was collected from participants in an NIH-funded study on impaired expressive and receptive prosody in autism. The NEPSY Narrative Memory subtest was one of a large number neuropsychological and neurodevelopmental assessment tests administered to the participants in the study in order to gain a complete picture of each participant's neurological and cognitive functioning. All of the tests administered as part of this study were recorded using a stationary microphone attached to a laptop computer. In addition, video recordings were made of most of the tests. The examiner scored the tests in real time and later listened to the recordings and verified, or when necessary, corrected their real-time scores. We found that this step was crucial in obtaining accurate element-level scores.

From each NNM recording, a word-level transcription of the retelling was made according to the transcription rules outlined in the 2004 EARS Official Annotation Guidelines for conversational speech (Strassel, 2004). The data was transcribed by a single labeler with extensive training in child language transcription, and the transcriptions were reviewed and corrected by the author of this thesis. The words and phrases corresponding to each of the 17 story elements in each retelling were identified by one of two trained labelers. I manually verified these annotations and from those annotations derived element-level and summary-level scores. Agreement between the summary scores that I assigned from the transcripts and those assigned in real time by the examiners during test administration, as measured by the Pearson product-moment correlation coefficient, was 0.97.

From each transcript, a trained labeler produced word-level alignments between the retelling and the source narrative presented in Figure 5.7. I then verified and corrected these word-level alignments, again following the sure/possible technique used in MT (Och and Ney, 2003). In addition to serving as a gold standard against which to evaluate the output of the word alignment techniques explored in this thesis, which will be discussed as length in Chapter 7, these manual alignments will play a role in demonstrating the utility of word alignment and alternative scoring techniques for distinguishing diagnostic groups, as described in Chapter 8.

5.2.5 Autism Spectrum Disorder and Specific Language Impairment

The neurological disorders I investigate with the NEPSY Narrative Memory subtest are Autism Spectrum Disorder (ASD) and Specific Language Impairment (SLI). ASD and SLI suffer from frequent diagnostic confusion and substitution, and there is still disagreement about whether SLI can exist as co-morbidities or whether the language impairment observed in some individuals with ASD is distinct from SLI (Kjelgaard and Tager-Flusberg, 2001; Bishop et al., 2008; Tomblin, 2011). One of the minor aims of this thesis is to identify characteristics of narrative retellings that might distinguish ASD from SLI and potentially provide evidence for the existence of a subgroup within ASD that is language-impaired but distinct from SLI.

In defining ASD, we appeal to the criteria of the current Diagnostic and Statistical Manual of Mental Disorders (American Psychiatric Association, 2000). Disorders on the autism spectrum are characterized by (1) a qualitative impairment in social interaction, (2) qualitative impairments in communication, and (3) restrictive and repetitive interests and behaviors. Delayed or impaired language is not necessary for a diagnosis of ASD, although it is one of a number of symptoms that can indicate ASD, according to the DSM-IV criteria. The language deficits typically associated with ASD are usually framed in terms of their communicative function; that is, they are pragmatic or semantic in nature and unrelated to phonology, morphology, and syntax (Rapin and Dunn, 2003).

In the research presented here, we define SLI as the delayed or impaired acquisition of language without accompanying physical abnormalities or comparable delays or deficits in hearing, cognition, and intellectual development (Tomblin et al., 1997; McCauley, 2001; Bishop, 2006). SLI language problems may include deficits in any area of language production and comprehension, including syntax, morphology, phonology, as well as pragmatics, semantics, and vocabulary. Under some of the definitions provided in the literature, ASD is one of the exclusionary criteria for SLI; that is, an individual may not be diagnosed with SLI if he also has ASD (Tomblin, 2011).

Given previous work that found strong correlations between NNM scores and unrelated language measures, it is expected that performance on the NNM subtest may be able

Diagnosis	n	Age
TD	45	6.3
ALN	18	6.4
ALI	27	6.9
SLI	29	6.7

Table 5.4: Subject data for CSLU autism study participants.

to distinguish children with typical development (TD) from children with SLI. I will also explore using alternative scoring techniques and narrative fidelity features that are independent of the published scoring guidelines that may reveal distinct patterns of narrative recall in ASD subjects with a comorbid language impairment.

5.2.6 Experimental Subjects

Subjects for the original NIH-funded study on prosody in autism were recruited from OHSU's CDRC Autism Center, the Hearing and Speech Institute in Beaverton, OR, and the Kaiser Permanente Center for Health Research. Experienced clinicians, with expertise in developmental psychology, speech-language pathology, and occupational therapy, administered a large battery of neuropsychological tests to each child in order to generate a complete neurocognitive profile and definitive diagnosis for that child.

The subjects participating in this study ranged from 4 years to 8 years of age. There were no significant differences in age across groups, except between the TD and ALI group. All subjects were required to have a full-scale IQ of greater than 70, indicating no intellectual disability, and a mean length of utterance (MLU) of at least 3, indicating an adequate level of verbal fluency. IQ was measured using the WPPSI-III (Wechsler, 2002) for children under 7, and the WISC-IV (Wechsler, 2003) for children 7 and older. Exclusion criteria included any brain lesion or neurological condition such as cerebral palsy; orofacial abnormalities; bilinguality; extreme unintelligibility; gross sensory or motor impairments; or identified mental retardation. In addition, the exclusion criteria for the typical development (TD) group included a history of psychiatric disturbance, such as bipolar disorder or ADHD, and a family history ASD or language impairment.

We categorize out subjects into four groups according to their diagnoses: typical development (TD); autism spectrum disorder without a language impairment (ALN); autism spectrum disorder meeting the criteria for a language disorder (ALI); and specific language impairment (SLI). We can group the ALN and ALI groups into a single group, which we call autism spectrum disorder (ASD). Similarly, we can group the ALI and SLI groups into a single group called language impairment (LI).

A diagnosis of ASD was determined using two instruments: the Autism Diagnostic Observation Schedule (ADOS) (Lord et al., 2002), a semi-structured series of activities designed to allow an examiner to observe behaviors associated with autism; and the Autism Diagnostic Interview-Revised (ADI-R) (Lord et al., 1994), a parental interview. The results of these tests, along with observations and scores from the other neuropsychological tests, were then reviewed in a meeting of all of the clinicians, who then generated a consensus diagnosis according to the instrumental findings (i.e., a score on both tests at or above the cut-off for ASD) and the diagnostic criteria for ASD listed in the DSM-IV-TR (American Psychiatric Association, 2000). A majority of the ASD group also met the criteria for SLI described above. We therefore sometimes divide the ASD group into two subgroups: the individuals with ASD who do not meet the criteria for a language disorder (ALN) and those who do meet the criteria for SLI (ALI). The total number of participants meeting the criteria for ASD, SLI, and TD and not meeting any of the exclusionary criteria was 119. Table 5.4 provides information about the subjects' diagnoses and the mean age for each diagnostic group.

A subject received a diagnosis of SLI if he or she did not receive a diagnosis of ASD and met one of two commonly used criteria: 1) The Tomblin Epi-SLI criteria (Tomblin, 1996), in which diagnosis of language impairment is indicated when scores in two out of five domains (vocabulary, grammar, narrative, receptive language, and expressive language) are greater than 1.25 standard deviations below the mean; or 2) the CELF-Preschool-2/CELF-4 (Paslawski, 2005) criteria, in which diagnosis of language impairment is indicated when one out of three index scores and one out of three spontaneous language scores are more than one standard deviation below the mean.

5.3 Summary

In this chapter, I presented in detail the two narrative memory tasks that I will be analyzing, along with information about our data collection process, experimental subjects, and diagnostic procedures. The next several chapters will introduce the techniques I use to extract scores and information about narrative fidelity from this narrative recall data, and discuss how these scores and features can be used to perform diagnostic classification.

Chapter 6

Classification with manual scores

In the previous chapters, I noted that manually assigned scores on narrative recall tests can be useful in distinguishing diagnostic groups. Lower scores on the Wechsler Logical Memory subtest (WLM) are associated with Mild Cognitive Impairment (MCI) (Storandt and Hill, 1989; Petersen et al., 1999; Wang and Zhou, 2002; Nordlund et al., 2005), and performance on the NEPSY Narrative Memory (NNM) is correlated with many measures of language ability (Korkman et al., 1998). In this chapter, I discuss our research in establishing the diagnostic utility of the WLM and NNM scores for MCI and language impairment, respectively, both in terms of the summary scores reported under standard administration and in terms of individual element-level scores. I will also explore the relationships between these summary scores and other standardized index scores used to measure abilities that may be related to narrative recall performance.

6.1 Wechsler Logical Memory

6.1.1 Brief review of WLM

In the Logical Memory subtest of the Wechsler Memory Scale, a subject listens to a brief story and then retells the story to the examiner twice: once immediately upon hearing the story (Logical Memory I, LM-I), and a second time after a 30-minute delay (Logical Memory II, LM-II). Figure 6.1 shows the text of the Logical Memory narrative used in this study, with slashes indicating the boundaries between the brief phrases that constitute the story elements. During examination, the examiner notes on the scoring sheet included in the test materials which story elements the subject uses in each of his retellings. Under

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.

Figure 6.1: Text of WLM narrative, segmented into 25 story elements

the published scoring guidelines, the subject's score is calculated from the score sheet by counting the number of elements used in his retelling. This scoring procedure does not consider the identity of the story elements recalled. Rather, the summary score (i.e., the raw number of elements recalled) is the only score reported, even though the score sheet itself indicates which of the story elements were recalled. We refer the reader to Chapter 5 for extensive details on the standard administration and scoring procedure.

6.1.2 Data collection

Subjects in this study came from existing community cohort studies of brain aging at the NIA-funded Layton Aging & Alzheimer's Disease Center at Oregon Health & Science University. As described in Chapter 5, following Shankle et al. (2005) and Roark et al. (2011), we defined our MCI and non-MCI groups based on the Clinical Dementia Rating score (Morris, 1993). The CDR has been shown to have high expert inter-annotator reliability (Morris et al., 1997) and, importantly, is assigned independently of the neuropsychological tests that we are investigating in this paper. We refer readers to the above cited papers for a full definition of the CDR and to Section 5.1.6 for a full discussion of our motivations for using this particular measure.

We collected the original paper scoring sheets from just over 400 study participants, half of whom had received a CDR of 0.5, which corresponds to MCI, and the other half roughly age-matched individuals who have never had a CDR greater than 0. We chose the earliest available visit where the individuals had received the CDR of interest: for MCI subjects the earliest visit where they received a CDR of 0.5, and for non-MCI subjects, their earliest visit. We note that this is a distinct set of subjects from the subjects described in Chapters 5 and 7, for whom word-level transcripts were produced. There

Diagnosis	n	Age	Education
MCI	192	79.7	14.5 yr
Non-MCI	201	81.2	14.5 yr

Table 6.1: Demographic information for WLM subjects.

were no significant between-groups differences in age or years of education. Details are presented in Table 6.1.

We then manually entered the per-item results of the Wechsler Logical Memory test, both immediate and delayed, from these paper scoring sheets and reconciled the newly compiled results with what was in the database of scores assigned during examination. Several subjects could not be included in this study due to mismatches between the data collected and the scores that should have been found for that session leaving 201 subjects with CDR 0 and 192 subjects with CDR 0.5. For all of these subjects, we have fully audited and validated per-item results for both immediate and delayed retellings of the Wechsler Logical Memory test.

6.1.3 Tests for statistical significance

Previous work in applying WLM scores to the task of distinguishing subjects with MCI from those with typical aging has focused on statistical tests rather than machine learning (Petersen et al., 1999; Nordlund et al., 2005). Table 6.2 shows that summary scores for both immediate and delayed recall were highly significantly different between the two groups, with the non-MCI group scoring higher by a wide margin. The total number of elements recalled over both retellings was also highly significantly different between the two groups. For each retelling, there are 25 story elements; thus in the two retellings, there are 50 story elements in total. Using the chi-square test for goodness-of-fit, we found that 18 of the elements in LM-I and 24 of the elements in LM-II were recalled significantly more often by non-MCI subjects than MCI subjects, with p values ranging from 2.783507e-11 to 0.0428.

The participants in our study also complete a word list recall task (Rosen et al., 1984) that is part of the Consortium to Establish a Registry for Alzheimer's Disease (CERAD)

Score	MCI	Non-MCI	*t*	*p*
LM-I	9.8	12.8	-6.2919	8.4130e-10
LM-II	6.2	10.5	-7.7589	7.5384e-14
LM-I + LM-II	15.9	23.3	-7.3516	1.1564e-12

Table 6.2: Mean WLM scores and significance testing for between-group differences.

	LM-I	LM-II	CERAD acq.	CERAD rec.
LM-I	1.0000	0.7781	0.5721	0.5468
LM-II	0.7781	1.0000	0.5769	0.6366
CERAD acq.	0.5721	0.5769	1.0000	0.8718
CERAD rec.	0.5468	0.6366	0.8718	1.0000

Table 6.3: Correlations between Logical Memory scores and CERAD scores.

protocol (Morris et al., 1989; Morris, 1993). It has previously been shown that these CERAD word-list recall scores are also good predictors of MCI (Shankle et al., 2005). Table 6.3 also includes the correlations between the CERAD word-list recall scores and the WLM scores. Although the correlations are reasonable, they are not especially high, which suggests that word list recall and narrative recall might not rely on the same set of linguistic, cognitive, and memory-related skills, which might prove useful for improving diagnostic classification.

6.1.4 Machine learning classification

As previously noted, the score sheets contain information not normally reported when scoring the Wechsler Memory Scale, namely, the identities of the recalled story elements. Thus, for each subject, we were able to assemble a feature vector composed of 52 features: one for each story element in LM-I, one for each element in LM-II, and summary scores for LM-I and LM-II. Each story element feature was assigned a binary value of 1 if the story element was recalled and 0 otherwise. Summary scores ranged from 0 (none of the 25 elements recalled) and 25 (all 25 elements recalled).

We used LibSVM (Chang and Lin, 2011), as implemented within the Waikato Environment for Knowledge Analysis (Weka) API (Hall et al., 2009), to train support vector

machine (SVM) classifiers, using radial basis function kernal and default parameter settings. Summary scores were scaled in both the training and testing data to range between 0 and 1, according to the minimum and maximum of the scores in the training data.

The performance of the SVM classifiers was evaluated using leave-one-out validation. In this validation method, each subject is tested against an SVM trained on all of the other subjects. The SVM per-subject scores can be used to evaluate the classifier quality according to one of the most commonly used classification evaluation methods: the Receiver Operating Characteristic (ROC) (Egan, 1975). The ROC plots the false positive rate of a classifier against the true positive rate. The area under the resulting curve (AUC) is the measure typically reported for accuracy. A classifier performing at chance would have an AUC of 0.5, which would be the area under the line from (0,0) to (1,1)). A perfect classifier would have an AUC of 1.0. Note that all AUC values and AUC standard deviation values reported in this thesis have been multiplied by 100 to improve readability and comparability with other accuracy scores; thus, a random classifier will yield an AUC of 50, while a perfect classifier will result in an AUC of 100.

We use the Wilcoxon-Mann-Whitney statistic (Hanley and McNeil, 1982) to calculate the AUC. Taking each negative example's SVM score as a threshold, we count the number of positive examples with a higher SVM score. The AUC is the sum of these counts divided by the product of the number of positive examples and the number of negative examples. This can be expressed at follows, where $s(e)$ is the score, s, of some example, e; P is the set of positive examples, N is the set of negative examples, and $[s(p) > s(n)]$ is 1 if true and 0 if false:

$$AUC(s, P, n) = \frac{1}{|P||N|} \sum_{p \in P} \sum_{n \in N} [s(p) > s(n)] \tag{6.1}$$

The standard deviation for the AUC is calculated as follows, where AUC is abbreviated as A to improve readability:

$$\sigma_a^2 = \frac{A(1-A) + (|P|-1)(\frac{A}{2-A} - A^2) + (|N|-1)(\frac{2A^2}{1+A} - A^2)}{|P||N|} \tag{6.2}$$

Feature set	**AUC**	**s.d.**
LM summary scores	71.1	2.60
LM story elements	82.7	2.11
LM summary scores + story elements	82.7	2.10
CERAD	83.6	2.05
CERAD + LM summary scores	83.7	2.05
CERAD + LM story elements	85.1	1.97
CERAD + 7 chi-square-selected informative LM elements	88.5	1.92

Table 6.4: Classification performance.

In the final trial, we performed attribute selection to reduce the feature space of the set of story elements by ranking those features according to their chi-square statistic. Feature selection was performed separately on each training set to avoid introducing bias from the testing example. We trained and tested the SVM with the two CERAD scores and the top n story element features, from $n = 1$ to $n = 52$. We report here the accuracy for the top seven story elements ($n = 7$), which yielded the highest AUC measure. We note that over all of the folds, only 8 of the 52 features ever appeared among the 7 most informative.

To provide a baseline, we tested the SVM using a feature set consisting of the two LM summary scores alone as features. Subsequent trials used all story element features, and all story elements features together with the summary scores. Classification performance for these three features sets, is reported in rows 1-3 of Table 6.4. We observe a dramatic increase in classification accuracy over the baseline by using the identities of the individual story elements as features. Including the summary scores together with the story elements did not improve performance.

Recall that it has previously been shown that the CERAD word-list recall scores are also good predictors of MCI (Shankle et al., 2005). Since these scores are available for our pool of subjects, we now compare, in rows 4-7 of Table 6.4, the classification power of those scores with that of the Logical Memory summary scores and story elements scores.

The CERAD scores alone yield higher classification accuracy than the LM summary scores and slightly higher accuracy than the LM story element scores. However, including the LM story element scores with the CERAD scores in the SVM improves classification

performance significantly over both of these feature sets individually. Furthermore, including only a subset of LM story elements, selected according to their predictive significance as measured by the chi-square statistic, improves accuracy dramatically, to 0.885.

6.1.5 Discussion

The significant improvement in classification using the CERAD scores together with an informative subset of seven of the story element scores suggests that certain story elements may be more difficult to recall for subjects with MCI. Although we were careful to perform feature selection on the training data only, the set of most informative story element features always included 7 of the same 8 set of elements. Figure 6.2 shows the percentage of subjects in the two diagnostic groups and the two retelling contexts who recalled each of the story elements. The elements chosen with our feature selection method are denoted with asterisks, with one asterisk indicating that the LM-I element and two asterisks indicating the LM-II element.

We observe that primacy and recency effects for both diagnostic groups are not as marked in narrative recall scenarios as they are typically reported to be in word-list recall scenarios (Shankle et al., 2005). The two most commonly recalled elements for both diagnostic groups, *small children* and *was robbed of*, fall very near the middle of the story. These frequently recalled elements are crucial plot points in the narrative, while the more rarely recalled items, such as *on State Street* and *the night before*, are minor details.

These two frequently recalled elements number among the eight most informative elements. We also see, however, that another of the most informative elements is *Thompson*, which is both early in the story and an incidental detail. Previous work has shown that event details with more structural and causal importance are more likely to be recalled in the unimpaired adult population (Johnson, 1970; Trabasso et al., 1984). These trends will be investigated in detail in Chapter 8.

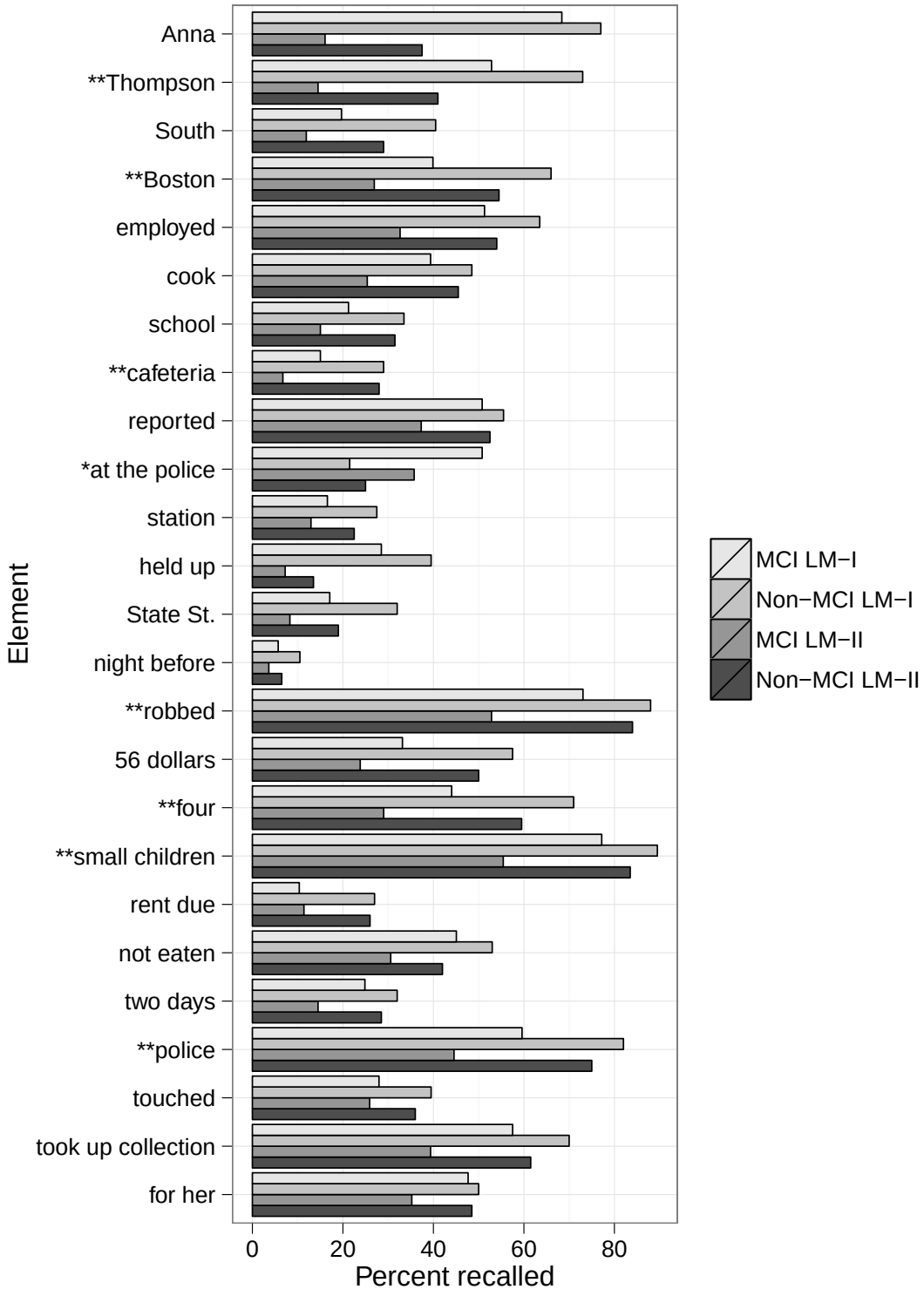

Figure 6.2: Percent of MCI and control subjects recalling each story element, with asterisks indicating the most informative features.

[Jim] was a boy whose best friend was [Pepper]. Pepper was a [big] [black] dog. Jim [liked to walk in the woods and climb the trees]. Near Jim's house was a very tall [oak tree with branches so high that he couldn't reach them]. Jim always wanted to climb that tree, so one day he [took a ladder] from home and carried it to the oak tree. He [climbed up], sat on a branch, and [looked out over his neighborhood]. When he started to get down, [his foot slipped, his shoe fell off, and the ladder fell to the ground.] Jim held onto a branch so he didn't fall, but he couldn't get down. Pepper sat below the tree and barked. Suddenly Pepper took Jim's shoe in his mouth and [ran away]. [Jim felt sad. Didn't his friend want to stay] with him when he was in trouble? [Pepper took the shoe] to [Anna], [Jim's sister]. He [barked and barked]. Finally, Anna understood that Jim was in trouble. She followed Pepper to the tree where Jim was stuck. [Anna put the ladder up and rescued Jim.] Wasn't Pepper a smart dog?

Figure 6.3: Text of NEPSY narrative with story elements in square brackets.

6.2 NEPSY Narrative Memory

6.2.1 Brief review of NNM

All of the subjects participating in our prosody in autism study complete the NEPSY Narrative Memory (NNM) subtest (Korkman et al., 1998). The NEPSY is a large and comprehensive battery of tasks that test neurocognitive functioning. The Narrative Memory subtest is administered and scored in roughly the same fashion as the Wechsler Logical Memory task: the examiner reads a story, and the subject is then asked to retell the story. A list of 17 required story elements is provided in a scoring worksheet, and the examiner notes which of these elements the subject uses in his retelling. As in the WLM, the score reported is the total number of elements recalled, where each recalled element is awarded 2 points. Figure 6.3 shows the NNM narrative. The portions of the narrative roughly corresponding to the required story elements are underlined. The subjects analyzed here are those described in Section 5.2.6: forty-five typically developing (TD) children, 18 children with autism spectrum disorder not meeting criteria for a language impairment (ALN), 27 with ASD and meeting the criteria for a language disorder (ALI), and 29 with specific language impairment (SLI).

Score	TD mean	(S)LI mean	t	p
TD vs. SLI	6.6	3.6	3.2927	0.0008
TD vs. LI	6.6	2.9	5.0796	0.000002

Table 6.5: Mean NNM scores and significance testing for between-group differences.

6.2.2 Tests for statistical significance

Although the score on the NNM is included in the index score for the memory and learning domain of the NEPSY, it is reportedly more strongly correlated with language measures (Korkman et al., 1998; Titley et al., 2008). For this reason, we might expect to see significant differences between the TD group and one or both of the language impaired groups, as reported under the published scoring guidelines. There are two groups of language impaired subjects that can be compared to the TD group: all subjects who meet the criteria for a language impairment including those with ASD, we refer to as LI, and the subjects with only SLI. Table 6.5 shows the mean NNM free recall score, t-score, and p-value for the two relevant comparisons across groups. The TD group performs significantly better than both the SLI-only group and the group comprising all children meeting the criteria for LI, including those with ASD.

The participants in the ERPA study completed numerous hours of testing, and there are dozens of standardized measures that might enhance classification in the same way CERAD scores improved classification of MCI. One such measure, which was not used to identify the presence of a language impairment, is the McCarthy verbal fluency score. In a verbal fluency task, the child has to name as many items from a particular category, such as animals or foods, as he can in a minute. Counts are tabulated after 20 seconds, 40 seconds, and 60 seconds. Here we will use the 40-second score in combination with the NNM scores to improve diagnostic classification.

6.2.3 Machine learning classification

There is one summary score for the free recall section of the NNM, and there are 17 story elements. Therefore, as in Section 6.1.4, we can use two feature vectors for classification:

one containing the summary score with a value ranging from 0 to 17, and one containing the 17 element-level binary scores. Since the number of subjects is quite small, we choose the leave-pair-out cross-validation method (Cortes et al., 2007; Pahikkala et al., 2008). In this approach, every pairwise combination of a positive and negative example is tested against an SVM built on the remaining examples. Again, we use the Wilcoxon-Mann-Whitney statistic (Hanley and McNeil, 1982) to calculate the AUC. Table 6.6 shows the results of the classification for the two comparisons of interest: (1) between TD and SLI only, and (2) between TD and all LI subjects including those with ASD. Both comparisons achieve fairly high levels of classification accuracy, but it appears that, in contrast to the results shown earlier for the WLM, using the element-level recall features collectively does not improve classification accuracy.

Using the McCarthy verbal fluency score as the single feature in the classifier yields higher results than the NNM scores for the small SLI-only group but lower results for the larger LI group including the ASD subjects. Including the McCarthy verbal fluency score as a feature with the NNM scores does not improve classification. Performing feature selection has an effect on accuracy for both comparison groups. As in Section 6.1.4, I considered every possible number of features from 1 to the total number of features, using the chi-square statistic for each feature as the selection criterion. The values of the chi-square statistics were calculated separately for each training fold in the cross-validation. In Table 6.6, I report the classification accuracy for the highest-performing feature selection configuration, which results in classification accuracy improvements for both comparisons.

6.2.4 Discussion

Although classification accuracy for SLI and LI using the NNM data was not quite as high as that achieved for MCI using the WLM data, the results here were nevertheless very promising. The standardized tests required to diagnose LI numbers in the dozens, and a diagnosis cannot be confirmed without analyzing a manually annotated speech transcript. In light of this, it is somewhat remarkable that a simple test like the NNM is able to distinguish children with LI and SLI from those with typical development.

Comparison	Feature set	AUC	s.d.
TD vs. SLI	summary score	74.9	6.1
TD vs. SLI	story elements	74.9	6.2
TD vs. SLI	McCarthy verbal fluency	77.1	5.9
TD vs. SLI	McCarthy verbal fluency and all WLM features	75.2	6.
TD vs. SLI	6 chi-square selected features	78.4	5.7
TD vs. LI	summary score	79.6	4.4
TD vs. LI	story elements	79.1	4.4
TD vs. LI	McCarthy verbal fluency	75.7	4.8
TD vs. LI	McCarthy verbal fluency and all WLM features	79.3	4.4
TD vs. LI	6 chi-square selected features	81.1	4.2

Table 6.6: Classification performance.

Note that in both cross-group comparisons of the NEPSY retellings, the most valuable story element by far was consistently story element 10 (*slipped, shoe fell, ladder fell, got stuck, or couldn't get down*). Note that element 10 actually contains multiple distinct events that occurred in the source narrative, something that was noted with some concern in Chapter 5. It is possible that this element was consistently recalled by TD children simply because there were so many possible ways of correctly recalling this element. Note also that element 10 encompasses all of the crucial Labovian "high-point" events around which the rest of the narrative is built. Understanding that this is the high point and is thus required in a retelling, could be complicated by the problems with language comprehension associated with LI. The memory and processing speed problems sometimes reported in LI might also contribute to the difficulty of recalling an event occurring toward the middle of a narrative.

Regardless of the underlying cause of this particular phenomenon, it bodes very well for diagnostic classification based on automatic scoring. If there are just a few story elements that can collectively distinguish the two diagnostic groups, then our overall element-level scoring accuracy is somewhat inconsequential as long as we can correctly identify those particular story elements. Perhaps the power of these features to predict LI will help us overcome the difficulties we will encounter in the next chapter when we extract scores from word alignments of retellings to the source narrative.

6.3 Summary

In this chapter I demonstrated the diagnostic utility of manual narrative recall scores assigned according to published guidelines for both the Wechsler Logical Memory subtest and the NEPSY Narrative Memory subtest. In particular, I showed that the identities of the recalled elements, which are not normally reported for these instruments, can be used to achieve diagnostic classification accuracy far above that achieved using only the standard summary scores. Selection of specific story elements as features for classification resulted in even more profound accuracy gains. Although element-level scores are not reported for these instruments, gathering this information places no burden on the examiner, as the story element identities are recorded during the standard administration of both the WLM and the NNM. The work presented in this chapter shows the great potential for using these narrative recall tests for diagnostic screening. In the next chapter, I will present techniques for deriving these scores automatically from written transcripts of retellings and compare the classification accuracy of models built using automatically extracted element-level scores with that of the models built from manual scores.

Chapter 7

Automatic scoring from alignments

In the previous chapter, I demonstrated that narrative retelling scores can be used to distinguish diagnostic groups and to classify individuals according to their diagnosis. Those results, however, were generated using manually assigned scores. In this chapter I discuss the techniques I have developed for automatically extracting these scores from transcripts of narrative retellings. I then use these scores for diagnostic differentiation and classification, and show that the automatically derived scores achieve classification accuracy comparable to and sometimes higher than the manually assigned scores.

This approach to automatic scoring relies on word alignments of the type used in machine translation for building word translation models. The algorithm for extracting the scores from the alignments is simple, but the process of getting high quality word alignments from the corpora of narrative retellings is challenging. The bulk of the research presented in this chapter outlines the various approaches I have explored for improving word alignment in the context of aligning narrative retellings and the impact of these different alignment strategies on scoring and diagnostic classification using both the Logical Memory subtest of the Wechsler Memory scale (WLM) and the Narrative Memory subtest of the NEPSY (NNM).

Recall that the research in this thesis is driven by the idea that there are similarities between narrative retelling and translation. In translation, a sentence in one language is converted into another language; the translation will have different words presented in a different order, but the meaning of the original sentence will be preserved. In narrative retelling, the source narrative is "translated" into the idiolect of the subject retelling the

story. Again, the retelling will have different words possibly presented in a different order, but most of the meaning will be preserved.

Although researchers in other NLP tasks that rely on alignments sometimes eschew the sort of word level alignments that are used in machine translation, we have no a priori reason to believe that this sort of alignment will be inadequate for the purposes of scoring narrative retellings. In addition, unlike many of the alignment algorithms proposed for textual entailment, the methods for unsupervised word alignment used in MT require no external resources, which will make it simpler to adapt our automated scoring techniques to new narrative generation scenarios. In the following sections, I will show that the word alignment algorithms used in machine translation, when modified in particular ways, provide sufficient information for highly accurate scoring of narrative retellings and subsequent diagnostic classification of the individuals generating those retellings.

In the coming sections, I will first present the techniques I have developed for word alignment, scoring, and classification in the context of the Wechsler Logical Memory subtest. Then I will discuss the applications of these techniques to the NEPSY Narrative Memory subtest. As I will show, the approaches I have developed are general enough to be easily adapted to a new instrument, given sufficient example retellings.

7.1 Preliminaries

7.1.1 Review of WLM data

Extensive information about the Wechsler Logical Memory (WLM) subtest is presented in Chapter 5. We provide a few details here to remind the reader of the task, the experimental subjects, and the data collected.

In the WLM, the subject listens to the examiner read a brief narrative and must retell immediately and after a delay of 20 to 30 minutes. The narrative and an example retelling are shown in Figure 7.1. The retellings are scored in real-time according to how many key story elements were used. The retellings produced by our subjects, both immediate and delayed, were transcribed at the utterance level and manually aligned, as described

below in Section 7.1.2. Further details about the administration of the task can be found in Section 5.1.1.

The experimental subjects for the work in this thesis include 72 subjects with MCI and 163 typically aging seniors roughly matched for age and years of education. The retellings for an additional 26 subjects who did not meet the eligibility requirements for the study due to age or diagnostic status were also transcribed and used as training data for learning word alignment models. Demographic and diagnostic information about the subjects are found in Section 5.1.7. The transcriptions themselves were tokenized and downcased, and all punctuation and pause-fillers were removed in preparation for alignment.

7.1.2 Example alignment

Let us consider an example from the WLM data that we will be analyzing. Figure 7.1 shows the source narrative used in our administration of the WLM subtest and a sample retelling from our data set. Figure 7.2 shows a visual grid representation of a *manually generated* word alignment between the source narrative and the retelling, along with close-up views of the four main regions of alignment. The source narrative is on the vertical axis and the retelling is on the horizontal axis in Figure 7.2.

When creating these manual alignments, I assigned the "possible" denotation under one of these two conditions: (1) when the alignment was ambiguous, as outlined in Och and Ney (2003); and (2) when a particular word in the retelling was a logical alignment to a word in the source narrative, but it would not have been counted as a permissible substitution under the published scoring guidelines. For this reason, we see that "Taylor" and "sixty-seven" are considered to be possible alignments because although they are logical alignments, they are not permissible substitutions according to the published scoring guidelines. Note that the word "dollars" is considered a possible alignment, as well, since the element *fifty-six dollars* is not correctly recalled in this retelling under the standard scoring guidelines.

In Figures 7.2 sure alignments are marked in black, while possible alignments are marked in grey. Figure 7.4 shows the word-index-to-word-index alignment, in which the

> 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 St. 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.
>
> 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.

Figure 7.1: The WLM narrative and an example retelling.

first index of each sentence is 0 and in which null alignments are not shown. Sure alignments are marked with S, and possible alignments are marked with P.

Manually generated alignments like this one are the gold standard against which any automatically generated alignments can be compared to determine the accuracy of the alignment. The goal of the work in this chapter is thus to produce such alignments accurately using automated techniques. From a word-to-word alignment, the identities of the story elements used in a retellings can be extracted, and from that set of story elements, the score that is assigned under the standard scoring procedure can be calculated, as I will presently show.

7.1.3 Story element extraction and scoring

As was described in great detail in Chapter 5, the published scoring guidelines for the WLM specify the source words that compose each story element. Figure 7.5 displays the source narrative with the element IDs ($A - Y$) and word IDs ($1 - 65$) explicitly labeled. Element Q, for instance, consists of the words 39 and 40, *small children*.

Using this information, we can determine which story elements were used in a retelling from the alignments as follows: for each word in the source narrative, if that word is aligned to a word in the retelling, the story element that it is associated with is considered to be recalled. For instance, if there is an alignment between the retelling word "sympathetic" and the source word *touched*, the story element *touched by the woman's story* would be counted as correctly recalled. Note that in the WLM, every word in the source narrative is part of one of the story elements. Thus, when we convert alignments to scores in

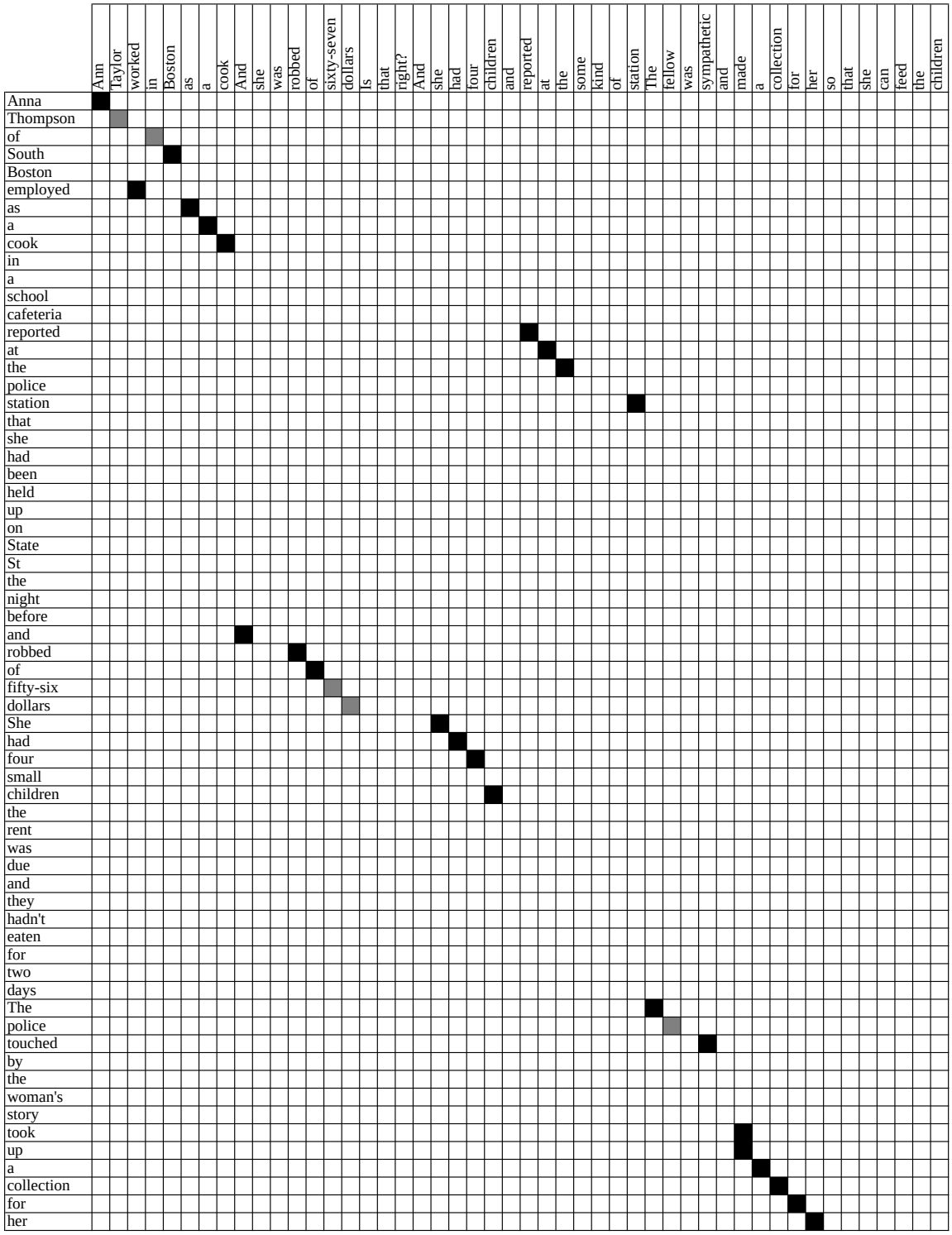

Figure 7.2: Visual representation of word alignment of narratives in Figure 7.1.

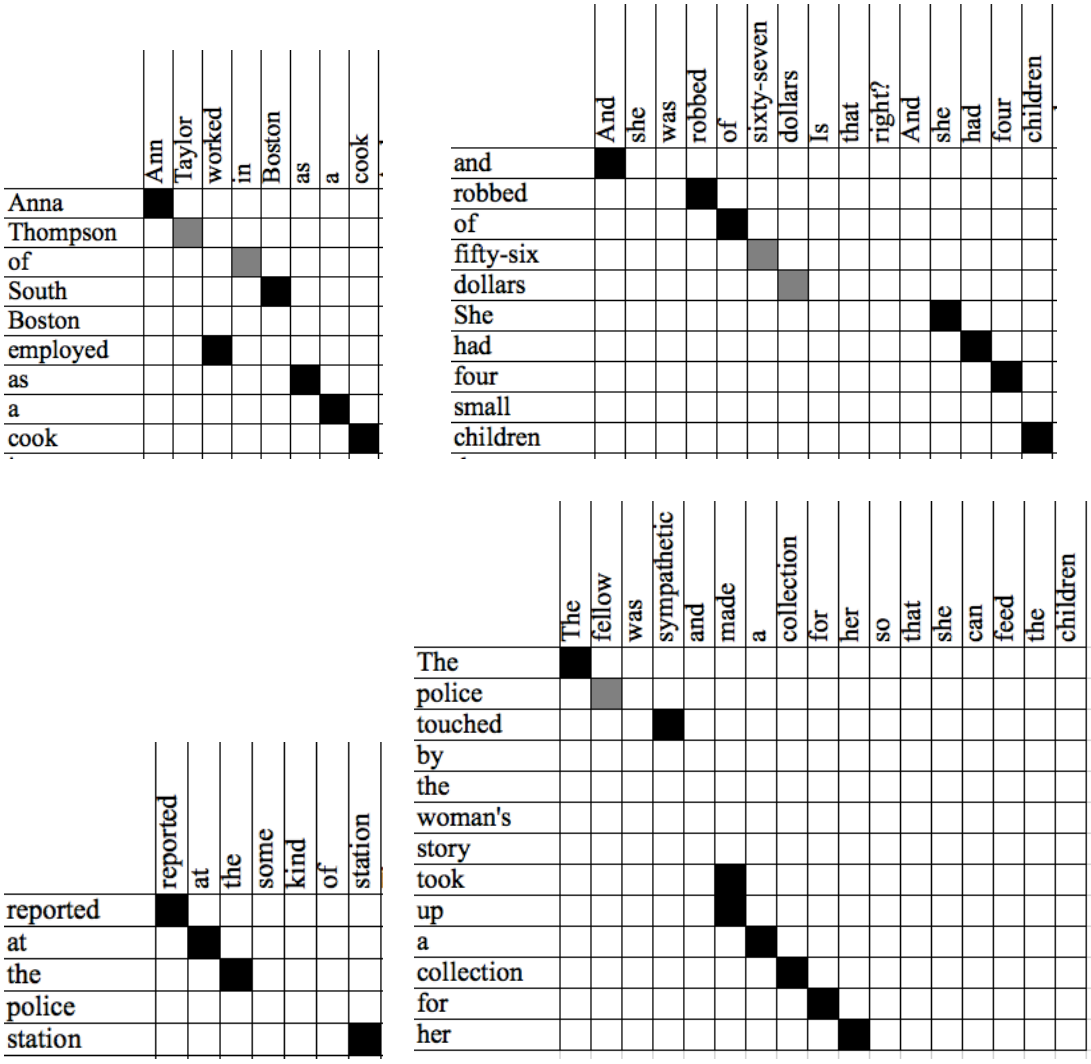

Figure 7.3: Close-up views of visual representation of word alignment of narratives.

the way just described, any alignment can generate a story element. This is true even for an alignment between function words such as *the* and *of*, which would be unlikely individually to indicate that a story element had been recalled. To avoid such scoring errors, we disregard any word-alignment pair containing a function word from the source narrative. The two exceptions to this rule are the final two words, *for her*, which are not content words but together make a single story element.

Source	Retelling	S/P
anna(0)	ann(0)	S
thompson(1)	taylor(1)	P
employed(5)	worked(2)	S
of(2)	in(3)	P
boston(4)	boston(4)	S
as(6)	as(5)	S
a(7)	a(6)	S
cook(8)	cook(7)	S
robbed(31)	robbed(11)	S
of(32)	of(12)	S
fifty-six(33)	sixty-seven(13)	P
dollars(34)	dollars(14)	P
she(35)	she(19)	S
had(36)	had(20)	S
four(37)	four(21)	S
children(39)	children(22)	S
reported(13)	reported(24)	S
at(14)	at(25)	S
the(15)	the(26)	S
station(17)	station(30)	S
the(52)	the(31)	S
police(53)	fellow(32)	P
touched(54)	sympathetic(34)	S
took(59)	made(36)	S
up(60)	made(36)	S
a(61)	a(37)	S
collection(62)	collection(38)	S
for(63)	for(39)	S
her(64)	her(40)	S

Figure 7.4: Sure (S) and Possible (P) index-to-index word alignment of the narratives in Figure 7.1.

Recall that in the manually derived word alignments, certain alignment pairs were marked as *possible* if the word in the retelling was logically equivalent to the word in the source but was not a permissible substitute according to the published scoring guidelines. When extracting scores from a manual alignment, only *sure* alignments are considered. This enables us to extract scores from a manual word alignment with 100% accuracy. The *possible* manual alignments are used only for calculating AER of an automatic word alignment model.

[A anna$_0$] [B thompson$_1$] [C of$_2$ south$_3$] [D boston$_4$] [E employed$_5$] [F as$_6$ a$_7$ cook$_8$] [G in$_9$ a$_{10}$ school$_{11}$] [H cafeteria$_{12}$] [I reported$_{13}$] [J at$_{14}$ the$_{15}$ police$_{16}$] [K station$_{17}$] [L that$_{18}$ she$_{19}$ had$_{20}$ been$_{21}$ held$_{22}$ up$_{23}$] [M on$_{24}$ state$_{25}$ street$_{26}$] [N the$_{27}$ night$_{28}$ before$_{29}$] [O and$_{30}$ robbed$_{31}$ of$_{32}$] [P fifty-six$_{33}$ dollars$_{34}$] [Q she$_{35}$ had$_{36}$ four$_{37}$] [R small$_{38}$ children$_{39}$] [S the$_{40}$ rent$_{41}$ was$_{42}$ due$_{43}$] [T and$_{44}$ they$_{45}$ had$_{46}$ n't$_{47}$ eaten$_{48}$] [U for$_{49}$ two$_{50}$ days$_{51}$] [V the$_{52}$ police$_{53}$] [W touched$_{54}$ by$_{55}$ the$_{56}$ woman's$_{57}$ story$_{58}$] [X took$_{59}$ up$_{60}$ a$_{61}$ collection$_{62}$] [Y for$_{63}$ her$_{64}$]

Figure 7.5: Text of WLM narrative with story element bracketing and word IDs.

Element ID	**Source** : **Retelling**
A	anna(0) : ann(0)
E	employed(5) : worked(2)
D	boston(4) : boston(4)
F	cook(8) : cook(7)
O	robbed(31) : robbed(11)
Q	four(37) : four(21)
R	children(39) : children(22)
I	reported(13) : reported(24)
K	station(17) : station(30)
W	touched(54) : sympathetic(34)
X	took(59) : made(36)
X	collection(62) : collection(38)
Y	for(63) : for(39)
Y	her(64) : her(40)

Figure 7.6: Alignment from Figure 7.4, excluding function words, with associated story element IDs.

From the list of story elements extracted in this way, the summary score reported under standard scoring guidelines can be determined simply by counting the number of story elements extracted. Table 7.6 shows the story elements extracted from the manual word alignment in Table 7.4.

7.1.4 Baseline classification

We will be performing diagnostic classification using these automatically extracted story elements as features within an SVM classifier, in the same way that we used examiner assigned manual scores for classification in the previous chapter. (See Section 6.1.4, above, and Section 7.4.2, below, for further details.) We will compare the diagnostic accuracy,

Model	AUC (s.d.)
Manual summary scores	73.3 (3.8)
Manual element scores	81.3 (3.3)
MMSE	72.3 (3.8)
LSA	74.8 (3.7)
Unigram overlap precision	73.3 (3.8)
Exact match closed-class summary score	74.3 (3.7)
Exact match closed-class unigrams	76.4 (3.6)

Table 7.1: Baseline classification accuracy results.

measured in terms of AUC, of a classifier trained on automatically extracted features to the accuracy of classifiers built on the expert-assigned manual WLM summary and element-level scores, shown in Table 7.1, below. In addition, we provide the accuracy of a classifier trained only on the expert-assigned scores for the Mini-Mental State Exam (MMSE) (Folstein et al., 1975), an instrument that is independent of the WLM and is widely used to identify mild to moderate dementia. Finally, we present the classification accuracy of three features sets derived automatically from the WLM retellings: (1) cosine similarity between a retelling and the source narrative measured using LSA, proposed by Dunn et al. (2002), with two scores for each subject, one for each of the two retellings; (2) unigram overlap precision of a retelling relative to the source, proposed by Hakkani-Tur et al. (2010), with two scores for each subject, one for each of the two retellings; and (3) a set of scores corresponding to the exact match via `grep` of each of the closed-class unigrams in the source narrative and a summary score thereof.

In Table 7.1, we see that all of the WLM-based features yield higher accuracy than the MMSE, which bodes well for using the WLM for classification of MCI. In addition, although all of the automatically derived feature sets yield higher classification than the MMSE, the manually derived WLM element-level scores are by far the most accurate feature set for diagnostic classification. Thus, the goal of the work presented in this chapter is to accurately automatically extract the story elements from the WLM retellings in order to try to achieve classification accuracy comparable to that of the manually assigned WLM story elements and higher than that of the other automatic scoring methods.

7.1.5 Training data

In Chapter 4, I described the two word alignment packages that are widely used in phrase-based machine translation: Giza++ and the Berkeley aligner. These word alignment packages take as input a sentence-aligned parallel corpus or bi-text, in which a sentence on one side of the corpus is a translation of the sentence in that same position on the other side of the corpus. Since we are interested in learning how to align words in the source narrative to words in the retellings, our primary parallel corpus must consist of source narrative text on one side and retelling text on the other. Because the retellings contain omissions, reorderings, and embellishments, we are obliged to consider the full text of the source narrative and of each retelling to be a "sentence" in the parallel corpus.

We compiled three parallel corpora to be used for the word alignment experiments:

- **Corpus 1:** A roughly 500-line source-to-retelling corpus consisting of the source narrative on one side and each retelling on the other. The set of retellings consists of the immediate and delayed retellings for each of the experimental subjects as well as for the each of 26 participants in the larger study who were not eligible for this particular investigation.

- **Corpus 2:** A roughly 900-line word identity corpus, consisting of every word that appears in every retelling and the source narrative. Each line on the source side of the corpus contains a single word and corresponds to a line on the target side containing that same word.

- **Corpus 3:** A roughly 250,000-line pairwise retelling-to-retelling corpus, consisting of every possible pairwise combination of retellings.

I will discuss how each of these corpora is used in the following sections. Note that all three of these corpora are constructed using only the clinically elicited WLM retellings and the source narrative itself. No external resources were used.

7.2 Baseline aligner performance

I begin by running Giza++ and the Berkeley aligner on Corpus 1. Giza++ follows this sequence of steps in its default configuration: 5 iterations of Model 1, 5 iterations of an HMM, 3 iterations of Model 3, and 3 iterations of Model 4. (The IBM models are discussed in more detail in Chapter 4.) When Giza++ is used to build a phrase-based translation model, two word alignment models are usually built: one from the source language to the target language and one from target language to the source (Och and Ney, 2003; Koehn et al., 2003). The output alignments of the two models are then combined to create a single symmetrized alignment using some heuristic method. The most commonly used heuristic is GROW-DIAG-FINAL-AND, which is usually used in the MT system Moses (Koehn et al., 2007), but simple methods such as taking the intersection or union of the two directional alignments is another alternative. There are also a number of additional preprocessing steps that Giza++ performs, including one that clusters words into classes, which are then used in the HMM and Model 4. We refer the reader to Och and Ney (2003) for further details.

The default configuration of the Berkeley aligner includes 5 iterations of Model 1 followed by 5 iterations of their HMM (Liang et al., 2006). The Berkeley aligner uses joint training, in which translation and distortion probabilities are estimated to jointly optimize the alignments in both directions simultaneously. In addition, the Berkeley aligner uses posterior decoding rather than Viterbi decoding. By default, the posterior threshold is set to 0.5. The default alignment combination technique is a soft union, in which the two posteriors for every alignment pair in the union of the two directional alignments are averaged, and every pair whose average across the two directions is above the posterior threshold is included in the final alignment.

All of the word alignment techniques used in this thesis, including both Giza++ and the Berkeley aligner, are entirely unsupervised. For this reason, in most experiments involving word alignment, the word alignment model is trained on the data for which alignments must be produced. It is not possible to use Giza++ without modification to test a word alignment model on a new set of test data. Similarly, the Berkeley aligner documentation

Aligner	P	R	AER
Giza++ with GROW-DIAG-FINAL-AND	4.9	4.5	95.3
Giza++ with intersection	55.7	3.5	93.4
Berkeley	35.1	1.63	96.8

Table 7.2: Baseline performance of aligners trained on Corpus 1.

recommends including all test data as training data and will do so automatically under some configurations. In this paper, we do not distinguish between testing data and training data in word alignment.

We evaluate the performance of the aligner in terms of precision, recall, and alignment error rate (AER), following Och and Ney (2000) and Och and Ney (2003). The alignments produced are compared to the manually generated set of word alignments, described in Section 7.1.2. AER is calculated as described in Section 4.2.3. Recall that the manual alignments between words can be labeled as *Sure* or simply *Possible*. The sure alignments are used in the calculation of recall, and the possible alignments are used in the calculation of precision. Note that if there are no alignments labeled possible, alignment error rate is equivalent to F-measure: $AER = 1 - F1$.

All three aligners performed dismally on the very small and very coarse Corpus 1, with 500 source-to-retelling pairs, as shown in Table 7.2. These poor results are not at all surprising. There are simply not enough examples in the training data for either algorithm to realistically determine the correct word alignments, especially since the "sentences" are all very long and contain largely the same set of lexical items.

Note the dramatic differences in precision and recall between the Giza++ alignments generated with the GROW-DIAG-FINAL-AND symmetrization heuristic and the other Giza++ alignments and the Berkeley alignments. The GROW-DIAG-FINAL-AND heuristic is designed to fill in gaps between aligned words, under the assumption that most words should probably be aligned. This is a reasonable assumption for a standard bilingual parallel corpus that would be used to build a machine translation model, since each target sentence should be a complete translation of its corresponding source sentence. For the task of aligning retellings, however, this assumption is not justified, since the retellings

include many omissions and insertions. For this reason recall is higher than in the Berkeley model, but precision is much lower.

The difference between the intersected Giza++ alignments and the Berkeley alignments is likely due to the two aligners' differing approaches to decoding. The Berkeley aligner uses posterior alignment; thus it returns only those alignments that exceed a certain posterior threshold, which in this case is very few because the model was not able to accumulate enough evidence for most of the word-to-word alignments. This results in virtually no recall but some amount of precision. In all three cases, the alignments will be of little use for any application, including the scoring of narrative retellings.

7.3 Improvement 1: Word identity training

In word alignment of two distinct languages for machine translation, there are likely to be few surface similarities between the two languages that can be easily exploited to improve the alignments. In the alignment of a source narrative to a retelling in the same language, however, we can take advantage of this fact: it is highly probably that a word in a source sentence will align to that same word if it appears in the target sentence. All that is needed is a way of informing the alignment model of this important fact. Following Daumé and Marcu (2005) and MacCartney et al. (2008), in our first approach to improving the baseline word aligner, we add a new corpus to the training data: Corpus 2, described in Section 7.1.5. Corpus 2 contains 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. As we will see in Section 7.6, we can increase the influence of the word identity data by including it multiple times in the training data.

The advantage of this method is that it can be incorporated trivially into existing word alignment frameworks. This very simple improvement, which requires no external resources, improves word alignment tremendously, as is shown in Table 7.3. Without the inclusion of the word identities in the corpus, there are virtually no usable alignments.

Aligner	Corpora	P	R	AER
Giza++ with GROW-DIAG-FINAL-AND	1	5.4	4.5	95.1
Giza++ with GROW-DIAG-FINAL-AND	1, 2	50.6	70.8	41.2
Giza++ with intersection	1	58.2	3.5	93.3
Giza++ with intersection	1, 2	69.2	48.5	42.8
Berkeley with soft union	1	35.1	1.6	96.8
Berkeley with soft union	1, 2	54.7	85.1	33.9

Table 7.3: Alignment quality for aligners trained on Corpora 1 and 2.

Aligner	AER	P	R	F1
Giza++ with GROW-DIAG-FINAL-AND	41.22	65.0	92.1	76.3
Giza++ with intersection	42.82	85.1	71.6	77.8
Berkeley with soft union	33.94	65.6	95.6	77.8

Table 7.4: AER and accuracy of scores from aligners trained on Corpora 1 and 2.

Including the word identity corpus enables the models to learn much more accurate translation probabilities.

7.3.1 Automatic Scoring

Given the alignments generated by these, we can now begin to extract scores using the procedure described in 7.1.3. For each of the above models with reasonable AER, Table 7.4 shows the precision, recall, and F-measure of the scores extracted from the alignments, relative to the manually assigned scores. We see that despite relatively high alignment error rates, the overall scoring accuracy is not that terrible, although there is a very poor balance in all three sets of scores between recall and precision. Since the diagnostic classification procedure outlined in the previous chapter relies not just on summary scores but on element-level scores, this balance between recall and precision may prove to be important. We therefore continue to seek improvements in alignment quality.

7.4 Improvement 2: Word identity weighting

Perhaps a more principled way to encourage the alignment of identical words would be to build this information into the algorithm itself. There may be, as I will discuss in Section

7.6, some optimal number of times that Corpus 2 should be included in the training corpus in order to find a balance between encouraging the identity alignment and rigidly enforcing it. That number is likely to depend on the size and lexical diversity of the source-to-retelling corpus and on the alignment algorithm itself. Such a parameter would have to be determined empirically for every dataset, which would make the alignment technique and subsequent scoring and classification techniques less general and less easily adapted to other narrative recall instruments. In addition, adding more lines to the training data will increase the number of computations required to build a model.

In order to be able to experiment with incorporating word identity into the alignment algorithm, I developed a word aligner, which I refer to here as the CSLU aligner. As in the Berkeley aligner, the translation parameters are initialized using IBM Model 1. During estimation, however, we include a prior probability, ψ, that an alignment is between identical words. The value of this prior probability was determined 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

$$\omega(a_j, j) = \begin{cases} \psi & \text{if } f_j \text{ and } e_{a_j} \text{are identical} \\ 1 - \psi & \text{otherwise} \end{cases} \tag{7.1}$$

In our held-out data, the value of ψ 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:

$$Pr(f_1^J | e_1^I) = \prod_{j=1}^{J} \sum_{i=1}^{I} p(f_j | e_i) \cdot \omega(i, j) \tag{7.2}$$

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 (such as when a speaker remembers only that

the police gave Anna some money), we model all jumps from 0 to the largest possible jump individually. Our HMM differs from the Berkeley HMM in two other ways. First, we train our models independently in both directions (retelling to original, and original to retelling), while the Berkeley aligner jointly trains its translation models. Secondly, we use Viterbi decoding to extract the alignments in both directions and then take the intersection of the resulting alignments, while the Berkeley aligner uses posterior decoding and, by default, takes the soft union of the alignments from the two directions. The CSLU aligner performs 5 iterations of the identity weighted IBM Model 1 but only 2 iterations of the HMM, since we found that performance converged or began to degrade after the first few iterations.

In addition to implementing the CSLU aligner, I also modified the Berkeley aligner source code to include this prior probability of identity alignment during Model 1 estimation. This allows us to compare the impact of word identity training with word identity weighting, as well as the joint training and posterior decoding used in the Berkeley aligner with the independent training and Viterbi decoding used in the CSLU aligner.

Table 7.5 first shows the previous results from Giza++ and the Berkeley aligner trained on Corpus 1, the source-to-retelling corpus, with Corpus 2, the word identity corpus. This information is followed by the word alignment accuracy results of both the CSLU aligner and the Berkeley aligner trained only on Corpus 1 with word identity weighting during estimation. We first see that including the prior probability of the identity alignment within the estimation phase of Model 1, both in the CSLU aligner and in the Berkeley aligner, results in an AER much lower than that of Giza++ trained on the identity corpus and comparable to that of the Berkeley aligner trained on the identity corpus.

We also, see however, that the CSLU aligner outperforms the Berkeley aligner when using identity weighting, an unexpected result given the benefits of joint training and posterior decoding over independent training and Viterbi decoding reported by Liang et al. (2006). Recall that the performance of the CSLU HMM began to degrade after a few iterations, leading us to include just 2 iterations of the HMM. For this reason, we also present results for both the identity-trained Berkeley aligner and the identity-weighted Berkeley aligner run with only 2 HMM iterations. This modification improves AER for

Aligner	Corpora	P	R	AER
Giza++ GROW-DIAG-FINAL-AND	1, 2	50.6	70.8	41.2
Giza++ intersection	1, 2	69.2	48.5	42.8
Berkeley ID training, soft union	1, 2	54.7	85.1	33.9
CSLU ID weighting, intersection	1	71.4	60.5	34.4
Berkeley ID weighting, soft union	1	50.8	85.2	36.9
Berkeley ID weighting, soft union, HMMx2	1	55.2	83.9	33.8
Berkeley ID training, soft union, HMM x2	1, 2	60.9	84.2	29.7

Table 7.5: Performance of aligners.

Aligner	Corpora	AER	P	R	F1
Giza++ GROW-DIAG-FINAL-AND	1, 2	41.2	65	92.1	76.3
Giza++ intersection	1, 2	42.8	85.1	71.6	77.8
Berkeley IDy training, soft union, HMMx5	1, 2	33.9	65.6	95.6	77.8
CSLU ID weighting, intersection	1	34.4	84	82.7	83.4
Berkeley ID weighting, soft union, HMMx5	1	36.9	63.2	96.6	76.4
Berkeley ID weighting, soft union, HMMx2	1	33.8	68.3	95.3	79.6
Berkeley ID training, soft union, HMMx2	1, 2	29.7	74.5	94.3	83.2

Table 7.6: Accuracy of scores comparing identity training and weighting.

both versions of the Berkeley aligner by several points to a level below the AER reported for the CSLU aligner, which is more in line with what might be expected given the training and decoding optimizations used in the Berkeley aligner.

7.4.1 Automatic scoring

We again extract scores using the procedure described in 7.1.3 for each of the alignments produced by the above models. Table 7.6 shows the precision, recall, and F-measure of these scores. Improvements in alignment quality seem to be somewhat tied to improvements in scoring accuracy, but the alignments from which the most accurate scores can be extracted are those from the CSLU aligner, which is not the aligner with the lowest AER. The CSLU aligner also produces scores with a good balance between precision and recall, which may prove to be important for classification. Although we are able to achieve promising automatic scoring accuracy from the alignments with the lowest AER, it remains to be seen whether this will translate to high classification accuracy.

7.4.2 Classification

Now that we have achieved reasonable scoring accuracy, we can use the automatically extracted scores as features within a support vector machine (SVM) to perform diagnostic classification for distinguishing subjects with MCI from those without. The process is similar to the one described in the previous chapter. For each of the 235 experimental subjects, we generate 2 summary scores: one for the immediate retelling and one for the delayed retelling. The summary score ranges from 0, indicating that no elements were recalled, to 25, indicating that all elements were recalled. In addition to the summary score, we also provide the SVM with a vector of 50 per-element scores: for each of the 25 elements in each of the two retellings per subject, there is a vector element with the value of 0 if the element was not recalled, or 1 if the element was recalled. The results presented in the previous chapter using manual scores indicate certain elements are more powerful in their ability to predict the presence of MCI. Therefore, we expect that giving the SVM these per-elements scores may improve classification performance. To train and test our classifiers, we use the WEKA API (Hall et al., 2009) and LibSVM (Chang and Lin, 2011), with an RBF kernel and default parameter settings.

We evaluate the performance of the SVMs using a leave-pair-out validation scheme (Cortes et al., 2007; Pahikkala et al., 2008). In the leave-pair-out technique, every pairing between a negative example and a positive example is tested using a classifier trained on all of the remaining examples. The resulting pairs of scores can be used to calculate the area under the receiver operating characteristic (ROC) curve (Egan, 1975), which is a plot of the false positive rate of a classifier against its true positive rate. The area under this curve (AUC) has a value of 0.5 when the classifier performs at chance and a value 1.0 when perfect classification accuracy is achieved. Recall that all AUC values and AUC standard deviation values reported in this thesis have been multiplied by 100 to improve readability and comparability with other accuracy scores. Further details about this cross validation technique are provided in Chapter 6.

Table 7.7 shows the classification accuracy of each SVM in terms of AUC for the set of summary score features and the set of element-level features. The AER and automatic

Aligner	AER	Score F1	Summary AUC	Element AUC
Giza++ GROW-DIAG-FINAL-AND	41.2	76.3	68.3 (3.9)	75.3 (3.7)
Giza++ intersection	42.8	77.8	72.1 (3.8)	78.6 (3.5)
Berkeley ID training HMMx5	33.9	77.8	72.1 (3.8)	78.2 (3.5)
Berkeley ID weighting HMMx5	36.9	76.4	70.8 (3.9)	73.8 (3.7)
Berkeley ID training HMMx2	33.8	79.6	70.7 (3.9)	75.8 (3.6)
Berkeley ID weighting HMMx2	29.7	83.2	71.4 (3.8)	72.9 (3.8)
CSLU ID weighting	34.4	83.4	74.4 (3.7)	75.3 (3.7)
Manual scores	N/A	N/A	73.3 (3.8)	81.3 (3.3)

Table 7.7: Classification accuracy of scores automatically extracted from alignments.

scoring F-measure are presented alongside the classification accuracy results, where applicable. When using manual scores, the element-level features distinguished the MCI subjects from the typically aging subjects with much higher accuracy than the summary score features. This holds true across the board when using automatically extracted scores. Remarkably, the automatically extracted summary score features under many of the models reach AUC levels approaching those of manual summary scores. The classification performance of the manual element-level scores, however, is still noticeably higher than that of any of the automatically derived element-level scores.

In general, as scoring accuracy increases, the classifier accuracy of models trained on summary score features also increases, but there is no clear relationship between scoring accuracy and the accuracy of a classifier trained on element-level scores. We do note that classifiers built on scores derived from alignments with very high recall tended to achieve higher classification accuracy. Nevertheless, since the classification accuracy of the manual element-level scores exceeds that of all of the automatically derived scores, we will continue to search for ways to improve alignment quality which should lead to gains in scoring accuracy and in turn classification accuracy. We will also explore other techniques for scoring narrative retellings in Chapter 8.

7.5 Improvement 3: Aligner configuration optimization

From this point forward, I will abandon Giza++ and the CSLU aligner to focus on the Berkeley aligner. There are several reasons for this decision. First, the Berkeley aligner

produces more accurate word alignments than both Giza++ and the CSLU aligner, which is not unexpected given the reductions in AER reported in Liang et al. (2006) when using joint training and posterior decoding. Secondly, although neither the Berkeley aligner code nor the Giza++ code is designed for easy modification by the user, the Berkeley aligner code is far more manageable. Finally and most importantly, there is no obvious mechanism in Giza++ for saving out a model built on a particular training corpus and then using that model to align a new, previously unseen corpus. The ability to build a model on a small corpus and then use that model to align a much larger corpus will become important when we discuss using graph-based methods in Section 7.7. First, however, I will discuss ways of optimizing the configuration parameters of the Berkeley aligner for improving the error rate of alignment of narrative retellings to a source narrative.

7.5.1 EM iteration configuration

In Table 7.5, we observe that the AER of the alignments generated by the Berkeley aligner varies with the number of HMM iterations performed. This suggests that the default behavior of the Berkeley aligner, which includes 5 iterations of both Model 1 and the HMM, might not be optimal for the task of aligning narrative retellings. This same tendency to degrade rather than improve or converge with additional iterations has been observed by others using the Berkeley aligner for tasks outside machine translation (Lardilleux et al., 2010).

Figure 7.7 shows how the AER of alignments produced the Berkeley aligner, trained on Corpora 1 and 2, changes as the number of iterations of the component training models increases. Figure 7.8 shows the same information but for the Berkeley aligner when trained only on Corpus 1 with lexical identity weighting during Model 1 estimation. In general, AER increases as number of iterations of both models increases, but we see that the best score for both systems results from 2 iterations of Model 1 followed by 1 iteration of the HMM; the next best combination is 1 iteration of Model 1 followed by 2 iterations of the HMM. These two configurations also produce the best balance between precision and recall. Figures 7.9 and 7.10 show for the Berkeley aligner with identity training and with identity weighting, respectively, the difference between precision and recall for each

of the combinations of number of iterations of Model 1 and the HMM. A value close to zero indicates that recall and precision are balanced. Under both identity weighting and identity training, the worst performing sequence of models, both in terms AER and precision-recall balance, is the most computationally expensive: 5 iterations of Model 1 and 5 iterations of the HMM.

Overall the alignment models built using lexical identity training data perform somewhat better than the models build using lexical identity weighting during estimation. Thus, although it might be more principled to encourage the identity alignment algorithmically, the technique relying on increasing the training data yields the best results on the source-to-retelling corpus.

7.5.2 Symmetrization and combination heuristic selection

Tables 7.3 and 7.5 also highlight the differences in the balance of recall and precision that are obtained according to the symmetrization heuristic that was used to combine the directional alignments. Techniques that rely on intersection, which eliminate alignments for which there is not strong evidence from both directional alignments, result in higher precision, while the more inclusive soft union and "grow" approaches yield higher recall.

The machine translation system, Moses (Koehn et al., 2007), offers several methods for symmetrization of Giza++ directional alignments, including union, intersection, and variants of the "grow" heuristic, which begins with the intersection of the two sets of alignments and begins to fill in gaps in the alignment by introducing alignments from the union that meet certain conditions. Och and Ney (2000) and Och and Ney (2003) argue that alignments should be optimized for recall when they are to be used for building machine translation models, which favors the union and growing approaches. For this reason, users of Moses typically select the GROW-DIAG-FINAL-AND heuristic. Och and Ney (2003) do suggest, however, that lexicographic applications might benefit from high precision alignments.

The Berkeley aligner also offers numerous alignment decoding combination options. Because the Berkeley aligner uses posterior decoding, these options differ from those used by Moses for Giza++ alignments. The default option is the soft union, described above

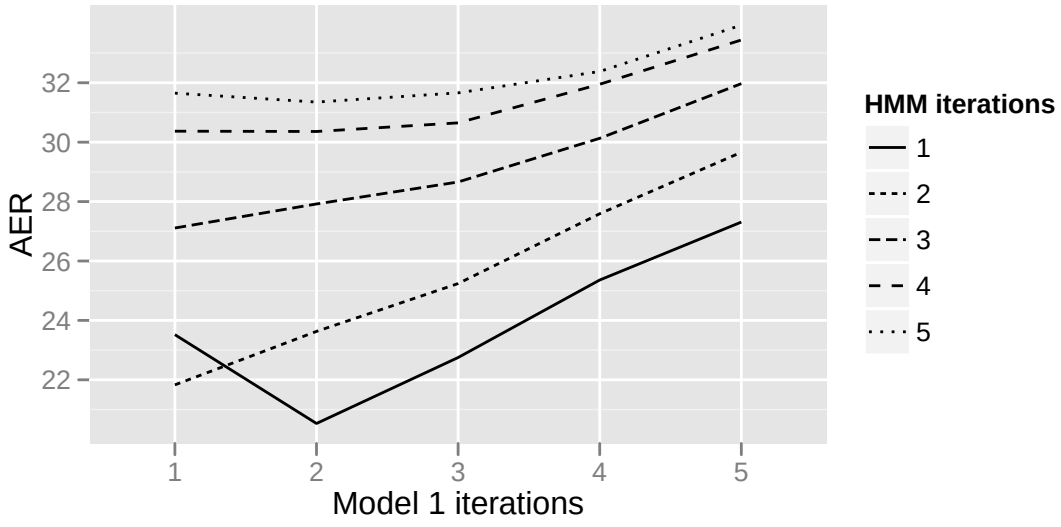

Figure 7.7: Berkeley aligner with identity training: changes in AER as number of Model 1 and HMM iterations increase.

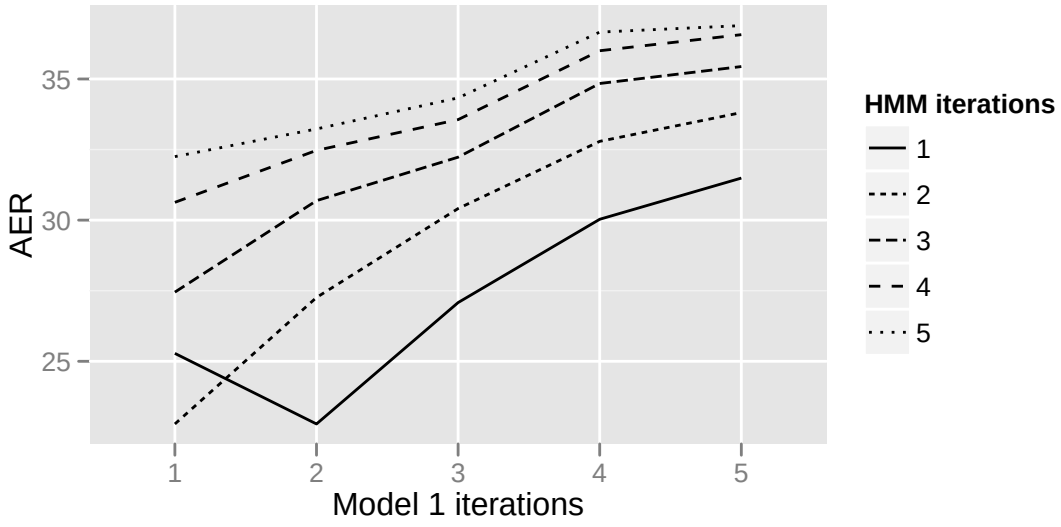

Figure 7.8: Berkeley aligner with identity weighting: changes in AER as number of Model 1 and HMM iterations increase.

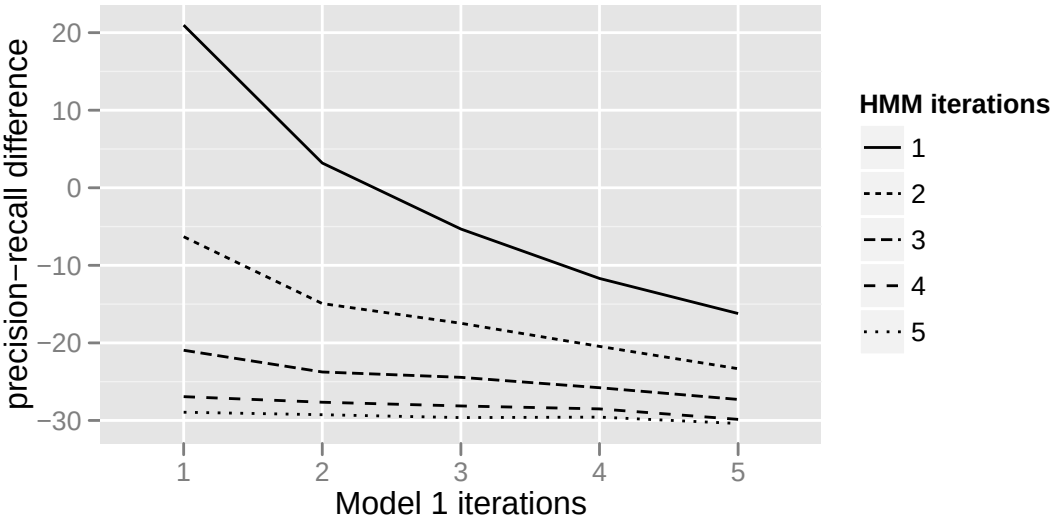

Figure 7.9: Berkeley aligner with identity training: difference between precision and recall as Model 1 and HMM iterations increase.

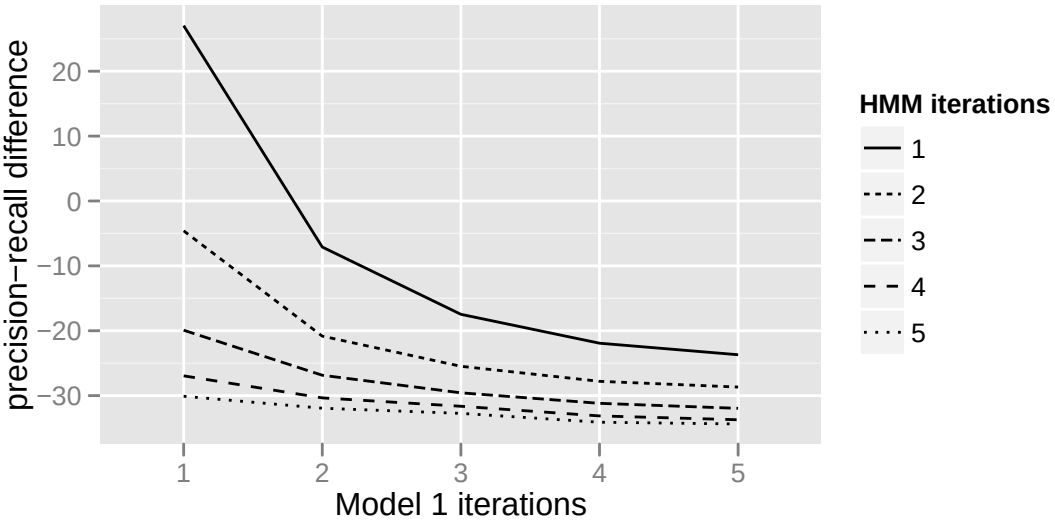

Figure 7.10: Berkeley aligner with identity weighting: difference between precision and recall as Model 1 and HMM iterations increase.

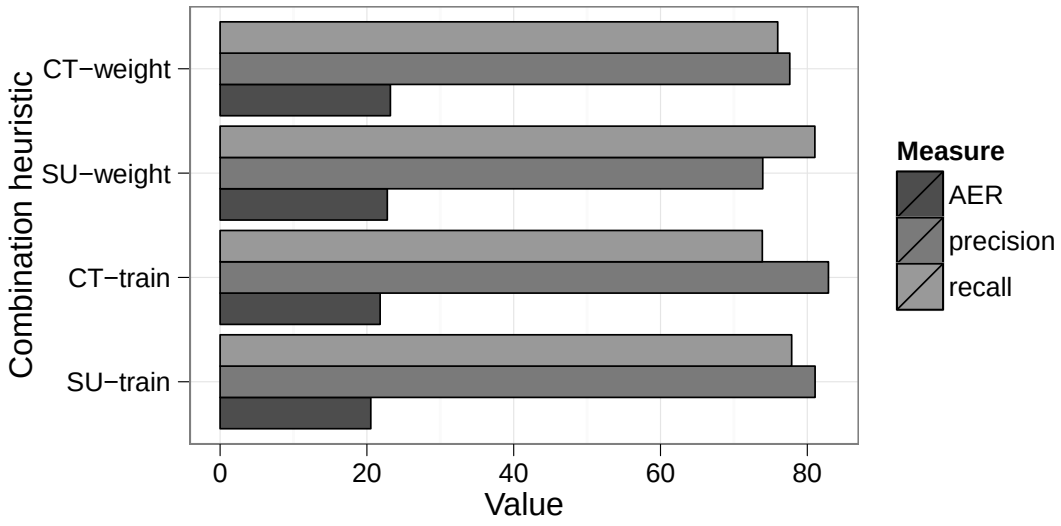

Figure 7.11: Comparison of soft union and competitive thresholding combination heuristics on the Berkeley alignments built using identity training and identity weighting.

in Section 7.2. In addition, the user can select hard union (all alignments from both directions above the posterior threshold), hard intersection (only the alignments that are above the threshold in both directions), and soft intersection (only the alignments whose product over the two directions is over the threshold). In the documentation for the software and in related work on syntactically informed word alignment (DeNero and Klein, 2007), the authors also recommend using competitive thresholding to balance the tradeoff between precision and recall when using one of the union combination heuristics. From the word-to-word matrix of weighted posteriors produced by one of the combination methods, competitive thresholding selects those alignments whose weighted posterior exceeds the posterior threshold and is either the maximum for both words or is adjacent to the maximum for both words.

Figure 7.11 shows, for the best training sequence (2 iterations of Model 1, followed by 1 iteration of the HMM), the precision, recall, and AER values for the Berkeley model with identity training using both the soft union combination (sutrain) and competitive thresholding of soft union (cttrain), and for the Berkeley model with identity weighting

using soft union alignment combination (suweight) and competitive thresholding of the soft union combination (ctweight). We see that both alignment combination methods for the models relying on identity training achieve a lower AER and higher precision than the models relying on identity weighting. Competitive thresholding typically results in higher precision, at the expense of lower recall; the soft union combination alone generally favors recall, although in this optimal training sequence, precision is slightly higher than recall for the identity trained model. Although I will continue to explore all four methods to some degree in the remainder of this chapter, I note that the soft union alignment combination method typically yields lower AER than competitive thresholding, and identity training yields lower AER than identity weighting.

7.5.3 Posterior threshold selection

The default posterior threshold in the Berkeley aligner is 0.5, and the authors report the best performance in their data using this posterior threshold. This value seems to be the optimal value in our data as well. Figure 7.12 shows the effect on AER of varying the posterior threshold from 0.1 to 0.9 on the best performing EM iteration sequence under both soft union and competitive thresholding of the soft union.

7.6 Improvement 4: Increasing size of training corpus

The authors of both Giza++ and the Berkeley aligner report that AER steadily decreases as the size of the training corpus increases. Thus, one obvious way to improve the alignments is to include more training data. Unlike most parallel corpora used in machine translation, the corpus we are analyzing is monolingual. We could therefore include additional monolingual data, such as different English translations of the same foreign text or news stories about the same event. This would involve, however, extensive manual intervention. We would ideally want to find in-domain parallel monolingual corpora; that is, we would need conversational, narrative monolingual parallel corpora produced by potentially neurocompromised seniors and children, something that would most probably not be readily available. Even if we were able to easily construct such a corpus it might only

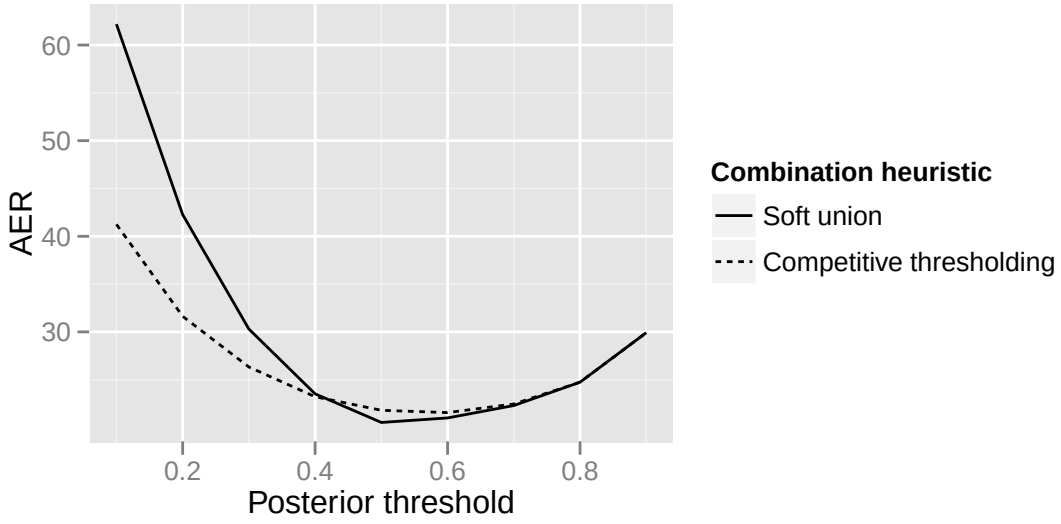

Figure 7.12: Change in AER of best performing Berkeley model as posterior threshold is increased.

prove helpful if the data happened to share vocabulary with the WLM story. Rather than seeking external resources, I instead propose to leverage the existing data in two ways.

Note that Corpus 1 includes the source narrative on the source side of the corpus and the retellings on the target side. This is an extremely small corpus from which to learn word alignment information. One way to increase the amount of training data is to include retelling-to-retelling sentence alignments; that is, we can add Corpus 3, which includes every possible pairing of retellings, to the training corpus. The size of the corpus will increase from roughly 500 lines to 250,000 (500^2) lines.

Another less obvious way to increase the amount of training data is to increase the size of Corpus 2, the word identity corpus: each unique word can be included once or multiple times, simply by increasing the number of times Corpus 2 is included in the training data. This could potentially improve the aligner's ability to correctly estimate the probability of the identity alignment. Table 7.8 shows the precision, recall, and AER values for the Berkeley aligner when trained on Corpus 1 with no identity optimizations and with identity weighting. This is followed by these same metrics for the Berkeley

Aligner	Corpora used	P	R	AER
Berkeley no ID information	Corpus 1	35.1	1.6	96.8
Berkeley with ID weight	Corpus 1	64.2	85.0	27.3
Berkeley with ID train	Corpus 1, Corpus 2 x1	69.9	84.8	23.6
Berkeley with ID train	Corpus 1, Corpus 2 x10	68.0	84.8	24.8
Berkeley with ID train	Corpus 1, Corpus 2 x100	65.7	83.4	26.8

Table 7.8: Performance of aligners trained on Corpora 1 and 2.

Aligner	Corpora used	P	R	AER
Berkeley with ID train	Corpus 1, Corpus 2 x1, Corpus 3	78.3	8.50	85.4
Berkeley with ID train	Corpus 1, Corpus 2 x10, Corpus 3	90.4	54.7	31.6
Berkeley with ID train	Corpus 1, Corpus 2 x100, Corpus 3	79.5	79.7	20.4
Berkeley with ID train	Corpus 1, Corpus 2 x1000, Corpus 3	72.2	85.5	21.9

Table 7.9: Performance of aligners trained on Corpora 1, 2, and 3..

aligner whose training data includes Corpus 1 with an increasing number of instances of Corpus 2. All models were built using two iterations of Model 1 and two iterations of the HMM with the soft union combination heuristic. We see that including exactly 1 instance of Corpus 2 in the training data achieves the lower AER and the highest precision.

Although the performance of the Berkeley aligner degrades slightly as more instances of Corpus 2 are included, it is possible, and perhaps likely, that the optimal number of times to include Corpus 2 is dependent on the size of the baseline corpus. Since Corpus 3, the full pairwise retelling-to-retelling corpus, is significantly larger than Corpus 1, it is possible that including the word identity Corpus 2 only once will not be sufficient to encourage the identity alignment. I ran the same experiments as were run above, but this time Corpus 3 was included in the training data. In contrast to what was observed when training on the small Corpus 1, we see in Table 7.9 that a single instance of Corpus 2 is grossly inadequate, resulting in an extremely high AER. AER steadily improves as the number of times Corpus 2 is included increases before degrading slightly when Corpus 2 is included 1000 times.

Determining, via brute force tuning, the optimal number of times to include Corpus 2 when training on the full retelling-to-retelling corpus is not a realistic goal, as each

HMM training iteration over Corpus 3 takes between 6 and 18 hours. Similarly, it is not realistic to train a model for every possible combination of Model 1 and HMM iterations, particularly since it is likely that the sparsity of the data will result in overfitting. For the moment, I will take as my new baseline the following two models that represent a compromise between using default parameters and tuning to this particular data set: (1) the Berkeley model built on Corpus 1 and ten instances of Corpus 2, with 2 iterations of both Model 1 and the HMM; and (2) the Berkeley model built on Corpus 1, Corpus 3, and 100 instances of Corpus 2, with 2 iterations of both Model 1 and the HMM.

7.7 Improvement 5: Random walks on a graph

We now compare these word alignment approaches to an entirely new approach that uses Berkeley-derived word alignments between retellings as the input for graph-based exploration of the alignment space in order to improve alignment accuracy. As we will see, this technique improves alignment quality, as measured by AER, as much as it is improved by increasing the size of the training corpus for the Berkeley aligner's EM-based approach. This graph-based method has the distinct advantage of requiring only a few minutes of computation. Scaling up the size of the training corpus for EM, although it does result in reductions in AER, requires significantly more resources, both in terms of processing and time.

Graph-based methods, in which paths or *random walks* are traced through an interconnected graph of nodes in order to learn more about the nodes themselves, have been used for various NLP and related tasks, including web-page ranking (PageRank (Page et al., 1999)), extractive summarization (LexRank (Erkan and Radev, 2004; Otterbacher et al., 2009)), and word sense disambiguation (Mihalcea, 2005; Sinha and Mihalcea, 2007; Agirre and Soroa, 2009). In the PageRank algorithm, the nodes of the graph are web pages and the edges connecting the nodes are the hyperlinks leading from those pages to other pages. The nodes in the LexRank algorithm are sentences in a document and the edges are the similarity scores between those sentences. In the word sense disambiguation graphs, the nodes are often WordNet word senses, and the edges are ontological or

similarity relationships between those senses. The number of times that a particular node is visited in a random walk reveals information about the importance of that node and its relationship to the other nodes and the graph itself. In many applications of random walks, the goal is to determine which node is the most central or has the highest prestige. For word alignment, however, the goal is to uncover new relationships and strengthen existing relationships between words in a retelling and words in the source narrative.

In the case of this graph-based method for word alignment, each node represents a word in one of the retellings or in the source narrative. The edges can be modeled in a number of ways, but we will begin by interpreting them simply as the set of unweighted directed alignments proposed by the Berkeley aligner to (1) words in the other retellings, and (2) words in the source narrative. To distinguish these two types of edges, we refer to the first as *links* and the second as *alignments*. We generate these edges by using an existing alignment model to align every retelling to every other retelling (creating *links*) and to the source narrative (creating *alignments*).

Starting at a word in one of the retellings, represented by a node in the graph, the algorithm can either walk from that node to another retelling word in the graph to which it is linked or to a word in the source narrative to which it is aligned. At each step in the walk, there is a set probability λ that determines the likelihood of transitioning to another retelling word versus a word in the source narrative. This probability functions similarly to the damping factor used in PageRank and LexRank, although its purpose is quite different. In the first implementation I will discuss, the destination word is selected at random from the set of possible links for the current word, but these links can be weighted, as I will discuss presently. When the walk arrives at a source narrative word, that word is the new alignment for the starting word proposed by that particular random walk, and a new random walk begins. For each retelling word, multiple random walks are performed, resulting in a distribution of proposed source word alignments for that word over all of the source words. The most frequently observed destination source word is the new alignment for the retelling word.

Consider the following set of retellings. In each retelling, the word that should align to the source word *touched* is rendered in bold:

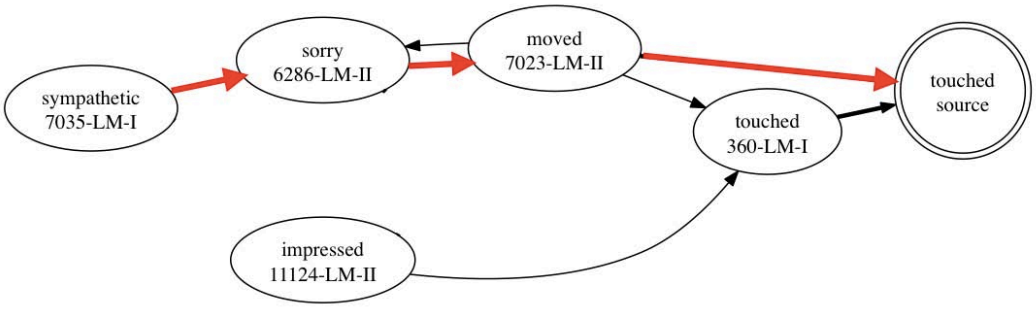

Figure 7.13: Subgraph of the full pairwise and source-to-retelling alignment.

anna thompson of south boston was employed as a cook at a school she reported to what was that station that she'd been held up and robbed the night before fifty-six dollars the rent was due she had four children the rent was due and they had not had anything to eat for two or three days and the police were so **moved** by the story that they took up a collection for her

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

anna thompson is a is a cook at a elementary school or a school and she was robbed on state street of fifty-six dollars she reported it to the police she had four children they hadn't eaten for two days and the police were **touched** by their story so they took up a collection

she lived in state street on state street some place in massachusetts boston she had four children and she was robbed of fifty-six dollars and the police took up a collection they were so **impressed** with her story oh they hadn't eaten for two days so the police were so impressed with her story they took up a collection

anna thompson worked as a cook in boston she stopped at a station and reported to the police that she had been robbed of fifty-six dollars she couldn't pay the rent or she hadn't paid the rent she had four children they had not eaten for two days and the police felt **sorry** for her and took up a collection

In Figure 7.13 is a small idealized subgraph of the pairwise alignments of these five retellings that illustrates the links between the relevant words in the retellings and their alignments to the word *touched* in the source narrative. We see that a number of these

words were not aligned to the correct source word, *touched*. They are all, however, linked to other retelling words that are in turn eventually linked to the source word. Starting at any of the nodes in the graph, it is possible to walk from node to node and eventually reach the correct source word. Although *sympathetic* was not aligned to *touched* by the Berkeley aligner, its correct alignment can be recovered from the graph by following the path through other retelling words. After hundreds or thousands of random walks on the graph, evidence will accumulate that these words should be aligned with the correct source word.

The approach as described might seem most beneficial to a system in need of improvements to recall rather than precision. Our baseline systems, however, are already favoring recall over precision. For this reason we introduced the NULL word to the set of words in the source narrative. Any retelling word that is not aligned to a source word by the input system will implicitly be aligned to the hidden source word NULL. This allows us to also model the likelihood of being unaligned. A word that was unaligned by the original system can remain unaligned, and a word that should have been left unaligned but was mistakenly aligned to a source word by the original system can recover its correct (lack of) alignment. In Figure 7.14, we see that although *food* was incorrectly aligned to *touched*, its correct alignment to NULL can be recovered by traversing retelling nodes in the graph.

We note that most implementations of both IBM Model 1 and HMM-based alignment also model the probability of aligning to a hidden word, NULL. In word alignment for machine translation, alignment to NULL usually indicates that a word in one language has no equivalent in the other language because the two language express the same idea or construction in a slightly different way. Romance languages, for instance, often use prepositions before infinitival complements (e.g., Italian *cerco di andare*, French *j'essaye d'aller*) when English does not (e.g., *I try to go*). In the alignment of narrative retellings, however, alignment to NULL can also indicate that the word in question is part of a concept, element, or piece of information that was not expressed in the source narrative.

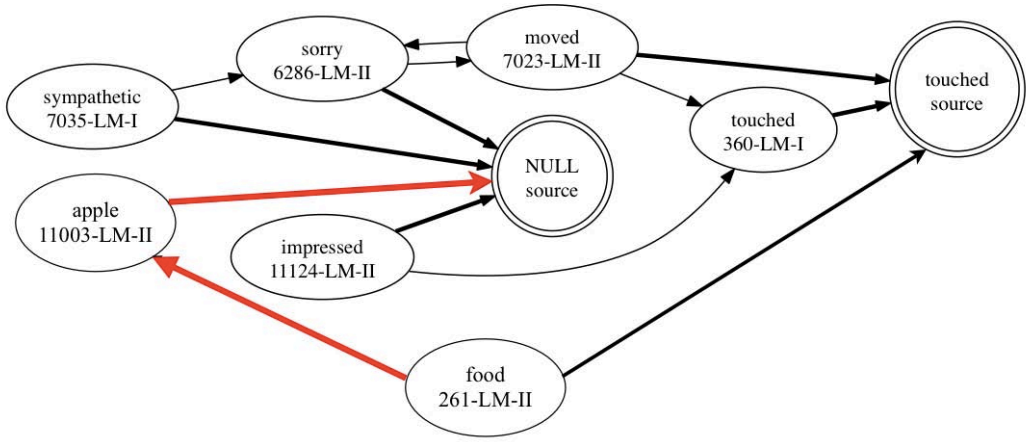

Figure 7.14: Subgraph of the full pairwise and source-to-retelling alignment including the NULL word.

7.7.1 Baseline random walk alignment model

We begin with the Berkeley alignment model that was trained on Corpus 1, the source-to-retelling corpus, and one instance of Corpus 2, the word identity corpus. This model will be referred to as the small Berkeley model. We then use that model to align not only the source narrative to the retellings but also every retelling to every other retelling. From this set of retelling-to-retelling alignments and the retelling-to-source alignments, we build the graph of nodes, in which each node represents a word and each edge represents an alignment to another word that was proposed by the Berkeley model. For each word in each retelling, we perform 1000 random walks on the graph. The single most frequent destination source word over the distribution defined by those 1000 random walks is the new alignment for that word. This is the small graph-based model.

We then build a second graph on the alignments generated by the large Berkeley alignment model that was trained on Corpus 1, 100 instances of Corpus 2, and Corpus 3, the full pairwise retelling-to-retelling corpus. We refer to the former as the large graph-based model and the latter as the large Berkeley model. We follow the same procedure for performing random walks and determining the new alignments.

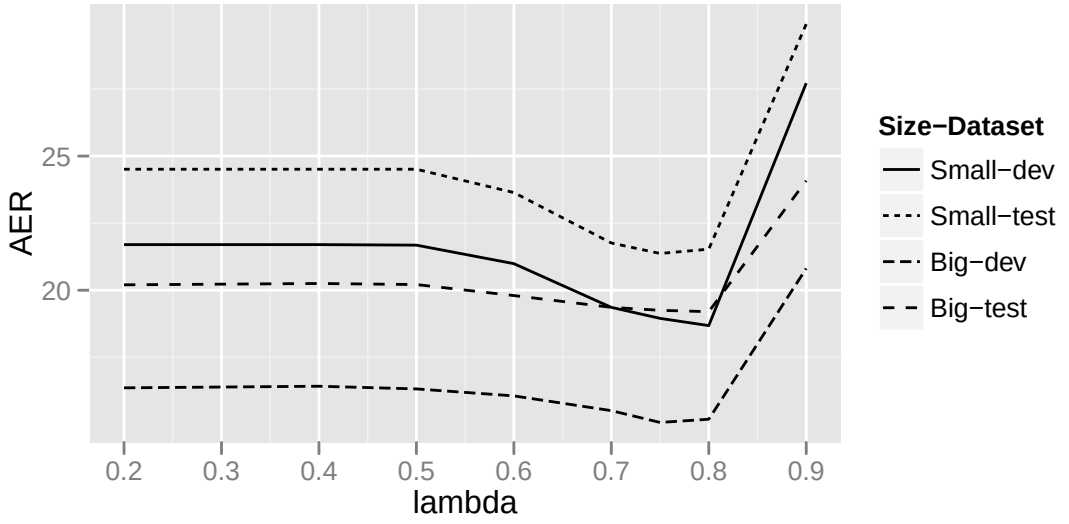

Figure 7.15: Changes in AER as λ increases.

We briefly note here that the Berkeley aligner occasionally fails to return an alignment for a sentence pair, either because one of the sentences is too long or because the time required to perform the necessary calculations exceeds some maximum allotted time. In these cases, in order to build a complete graph that includes all retellings, we back off to the alignments and posteriors generated by the second iteration of IBM Model 1.

Recall that we still need to determine the optimal value for λ, the probability of transitioning to another retelling word rather than a source word. We now separate the retellings into two sets: the retellings produced by the 235 experimental subjects and the retellings produced by the 26 subjects who were ineligible for the study. We will use this latter set as a development set for tuning this probability, and we will report AER results on the former, since these are the retellings that will be scored for use in diagnostic classification.

Figure 7.15 shows how AER changes as the value of λ increases from 0.2 to 0.9 on both the development set and the test set. Using this information, we select $\lambda = 0.80$ as the optimal value for the graph built using the small Berkeley model and $\lambda = 0.75$ as the optimal value for the graph built using the large Berkeley model. We report the

Model	P	R	AER
Berkeley-Small	72.3	78.5	24.8
Berkeley-Large	79.0	79.4	20.9
Graph-based-Small	83.0	74.2	21.5
Graph-based-Large	84.6	77.1	19.3

Table 7.10: Comparison of Berkeley and graph-based models with unweighted edges.

Model	P	R	AER
Berkeley-Small	72.3	78.5	24.8
Berkeley-Large	79.0	79.4	20.9
Graph-based-Small	77.9	81.2	20.6
Graph-based-Large	85.4	76.9	18.9

Table 7.11: Comparison of Berkeley and graph-based models with weighted edges.

results on the test set using that value, alongside the precision, recall, and AER values for the small Berkeley model and the large Berkeley model on the test set. We see that each of the graph-based models outperforms the Berkeley model of the same size. The performance of the small graph-based model is especially remarkable since it achieves an AER comparable to the large Berkeley model while requiring significantly less processing time. Running the millions of random walks required to generate new alignments requires only a few minutes, while building a full EM model on 250,000 very long sentences takes many hours.

7.7.2 Posterior weighted edges

In the graphs described above, the edges between nodes are directional but unweighted. The Berkeley aligner can be configured to output the posteriors for each possible word-to-word alignment of the words in a sentence pair. Using these posteriors we can build a graph with weighted edges, such that at each step in the walk, the choice of the next destination node can be determined according to the strength of the outgoing links, as measured by the posterior probability of that link. Recall that the posterior threshold is set to 0.5 in the Berkeley aligner's default configuration and that this value yields the lowest AER for the Berkeley aligner with the optimal configuration, as discussed in Section

7.5.3. In a graph built using Berkeley alignments and posteriors, we can adjust the value of this threshold again. In this case, the threshold will determine whether or not to include a link in the graph; if a link's posterior probability exceeds the threshold, then it will be included in the graph.

Figures 7.16 and 7.17 show how AER changes as both the link vs. align factor, λ, and the link inclusion threshold are varied. The AER values in Figure 7.16 are for alignments generated for the development set using a graph with posterior-weighted edges built from the small Berkeley model, while the AER values in Figure 7.17 are for the alignments generated for the development set by the graph built from the large Berkeley mode. We see that in both the small and large graph-based model, the tuning set performs best when the value of λ is 0.8 and the link inclusion posterior probability threshold is 0.5. The precision, recall, and AER for the alignments of the experimental subjects using these parameter settings are presented in Table 7.11. Again, each graph-based model outperforms the Berkeley model of the corresponding size by a large margin. In addition, weighting the links with the posterior probabilities improves the performance of the small graph-based model to levels lower than that of the large Berkeley model, which is remarkable given the difference in speed and computing requirements between the two approaches.

Figures 7.18 and 7.19 show the results of aligning the retelling presented in Figure 7.1 using the small Berkeley model and the large graph-based model, respectively. Comparing these two alignments, we see that the latter model yields more precise alignments with very little loss of recall, as is borne out by the overall statistics shown in Table 7.11.

Scoring and classification

The element-level scores induced from the four word alignments for all 235 experimental subjects were evaluated against the manual per-element scores. We report the precision, recall, and f-measure for all four alignment models in Table 7.12. In addition, we report Cohen's kappa, which is usually used as a measure of inter-rater agreement for manual raters, as a measure of reliability between our automated scores and the manually assigned scores. We see that as AER improves, scoring accuracy also improves, with the large graph-based model outperforming all other models in terms of precision, f-measure, and

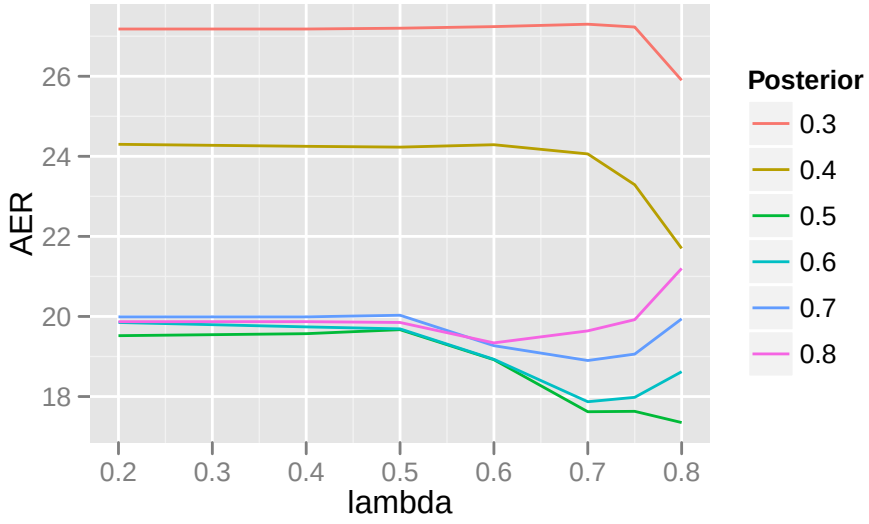

Figure 7.16: Changes in AER on the development set for the small posterior-weighted graph as λ and the posterior threshold vary.

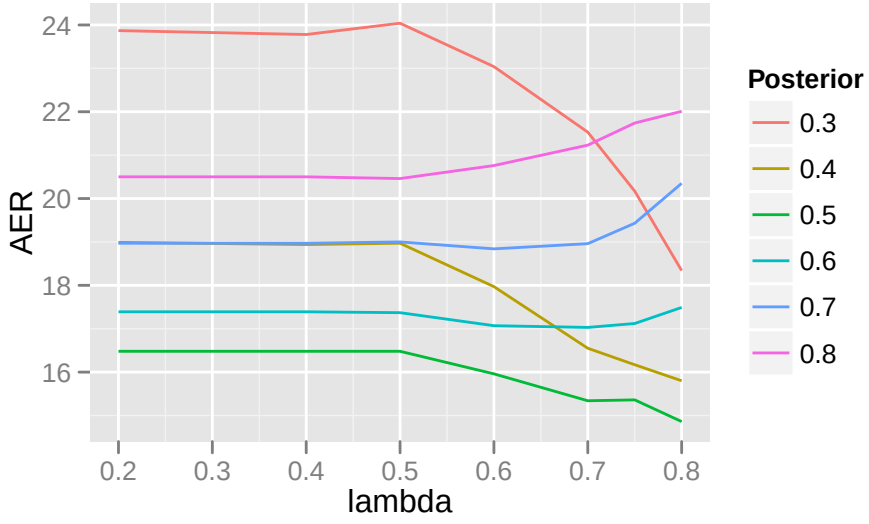

Figure 7.17: Changes in AER on the development set for the large posterior-weighted graph as λ and the posterior threshold vary.

ann(1) : anna(1)
worked(3) : employed(6)
in(4) : in(10)
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) : dollars(35)
is(16) : was(43)
that(17) : that(19)
and(19) : and(45)
she(20) : she(36)
had(21) : had(37)
four(22) : four(38)
children(23) : children(40)
reported(25) : reported(14)
at(26) : at(15)
the(27) : the(16)
some(28) : police(17)
station(31) : station(18)
made(37) : up(61)
made(37) : took(60)
a(38) : a(62)
collection(39) : collection(63)
for(40) : for(64)
her(41) : her(65)
so(42) : woman's(58)
she(44) : she(20)

Figure 7.18: Word alignment for the retelling in Figure 7.1 generated by the small Berkeley model with retelling words italicized.

ann(1) : anna(1)
taylor(2) : thompson(2)
worked(3) : employed(6)
in(4) : in(10)
boston(5) : boston(5)
as(6) : as(7)
a(7) : a(8)
cook(8) : cook(9)
robbed(12) : robbed(32)
of(13) : of(33)
sixty-seven(14) : fifty-six(34)
dollars(15) : dollars(35)
she(20) : she(36)
had(21) : had(37)
four(22) : four(38)
children(23) : children(40)
reported(25) : reported(14)
at(26) : at(15)
the(27) : the(16)
station(31) : station(18)
made(37) : took(60)
a(38) : a(62)
collection(39) : collection(63)
for(40) : for(64)
her(41) : her(65)

Figure 7.19: Word alignment for the retelling in Figure 7.1 generated by the large graph-based model with retelling words italicized.

inter-rater reliability. The scoring accuracy levels reported here are comparable to the levels of inter-rater agreement typically reported for the WLM, and reliability between our automated scores and the manual scores, as measured by Cohen's kappa, is well within the ranges reported in the literature for similar scoring approaches (Johnson et al., 2003).

Table 7.13 shows the classification results for the scores derived from the four alignment models. For comparison purposes, we also reprint the classification results for the five baseline and target models presented in Section 7.1.4, above: (1) examiner-assigned Mini-Mental State Exam score (manual), (2) examiner-assigned standard WLM scores (manual), (3) LSA cosine similarity between the retelling and the source (automatic), (4)

Model	P	R	F	κ
Berkeley-Small	87.2	88.9	88.0	76.1
Berkeley-Large	86.8	90.7	88.7	77.1
Graph-Small	84.7	93.6	88.9	76.9
Graph-Big	88.8	89.3	89.1	78.3

Table 7.12: Scoring accuracy results.

Model	Summ. (s.d.)	Elem. (s.d.)	Elem. subset (s.d.)
Berkeley-Small	73.7 (3.74)	77.9 (3.52)	80.3 (3.4)
Berkeley-Big	75.1 (3.67)	79.2 (3.45)	81.2 (3.3)
Graph-Small	74.2 (3.71)	78.9 (3.47)	80.0 (3.4)
Graph-Big	74.8 (3.69)	78.6 (3.49)	81.6 (3.3)
Manual Scores	73.3 (3.76)	81.3 (3.32)	82.1 (3.3)
MMSE	72.3 (3.8)	n/a	n/a
LSA	74.8 (3.7)	n/a	n/a
Unigram overlap precision	73.3 (3.8)	n/a	n/a
Exact match closed-class	74.3 (3.7)	76.4 (3.6)	n/a

Table 7.13: Classification accuracy results (AUC).

unigram precision (automatic), and (5) story element content word exact match (automatic).

In addition to using summary scores and element-level scores as features for the story-element based models, we also perform feature selection over both sets of features using the chi-square statistic, as described in Section 6.1.4. In all cases, the per-element scores are more effective than the summary scores in classifying the two diagnostic groups, and performing feature selection results in improved classification accuracy. All of the element-level feature sets automatically extracted from alignments outperform the MMSE and all of the alternative automatic scoring procedures, which suggests that the extra complexity required to extract element-level features is well worth the time and effort.

We also see that our automated scores have classificatory power comparable to that of the manual gold scores, and that as scoring accuracy increases from the small Berkeley model to the graph-based models and bigger models, classification accuracy improves. This suggests both that accurate scores are crucial for accurate classification and that pursuing even further improvements in word alignment is likely to result in improved diagnostic

differentiation. We note that although the large Berkeley model achieved the highest classification accuracy when the feature set includes the full list of story elements, this relatively slight margin of difference may not justify its significantly greater computational requirements, particularly since feature selection allows the large graph-based model to outperform the large Berkeley model.

7.8 Application to NEPSY Narrative Memory

The NEPSY Narrative Memory (NNM) subtest is similar to the Wechsler Logical Memory subtest: the subject listens to the examiner read a brief narrative and then must retell the narrative. For a number of reasons, however, the NNM presents many novel challenges. First, the source narrative itself is over 200 words, which makes learning accurate word alignment probabilities more difficult. Secondly, since the subjects are children, some who are neurologically impaired and others who for whatever reason did not feel motivated to participate in the tasks in a meaningful way, the retellings themselves are much more likely than the WLM narratives to be excessively short or completely off-topic. Thirdly, the NNM scoring procedure is entirely different from the WLM scoring procedure, as is discussed at length in Chapter 5. Finally, the amount of available data is more limited, since there is no delayed retelling and since there were many fewer participants in the study. For these reasons, the details of the procedure described above will be altered slightly, but the basic scheme remains the same: we train a word alignment model on parallel data from NNM retellings, align each clinically elicited retelling to a source, extract scores from the resulting alignment, and use those scores for diagnostic classification.

7.8.1 NNM Data

Extensive information about the NNM subtest is presented in Chapter 5. We provide a few details here to remind the reader of the task, the experimental subjects, and the data collected. In the NNM, the subject listens to the examiner read a brief narrative and must retell the narrative immediately. The NNM narrative is shown in Figure 7.20. The retellings are scored in real-time according to how many key story elements were

Jim was a boy whose best friend was Pepper. Pepper was a big black dog. Jim liked to walk in the woods and climb the trees. Near Jim's house was a very tall oak tree with branches so high that he couldn't reach them. Jim always wanted to climb that tree, so one day he took a ladder from home and carried it to the oak tree. He climbed up, sat on a branch, and looked out over his neighborhood. When he started to get down, his foot slipped, his shoe fell off, and the ladder fell to the ground. Jim held onto a branch so he didn't fall, but he couldn't get down. Pepper sat below the tree and barked. Suddenly Pepper took Jim's shoe in his mouth and ran away. Jim felt sad. Didn't his friend want to stay with him when he was in trouble? Pepper took the shoe to Anna, Jim's sister. He barked and barked. Finally, Anna understood that Jim was in trouble. She followed Pepper to the tree where Jim was stuck. Anna put the ladder up and rescued Jim. Wasn't Pepper a smart dog?

Figure 7.20: Text of NEPSY narrative.

Story Element
1. Jim
2. Pepper
3. big
4. black
5. liked to walk in the woods *or* climb trees
6. tree/oak with branches too high for Jim to reach
7. climbed the tree/oak
8. got a ladder *or* carried a ladder to the tree/oak
9. looked out over the neighborhood *or* looked around
10. slipped *or* shoe fell *or* ladder fell *or* got stuck *or* couldn't get down
11. Pepper ran for help *or* went to get help *or* ran away
12. Jim was sad *or* thought Pepper didn't want to stay
13. Anna
14. Jim's sister
15. took her Jim's shoe
16. barked and barked
17. Anna put the ladder back up *or* rescued Jim *or* helped Jim

Figure 7.21: Story element list for the NNM narrative.

used. These story elements are not drawn directly from the source narrative, as they are in the WLM. Rather, there is a distinct list of target elements which has only partial lexical overlap with the source narrative, shown in Figure 7.21. Further details about the administration of the task can be found in Section 5.2.1.

The experimental subjects for the work in this thesis include 45 subjects with typical development (TD), 18 with autism spectrum disorder not meeting the criteria for a language disorder (ALN), 25 subjects with ASD meeting the criteria for a language disorder

(ALI), and 29 subjects with specific language impairment (SLI). Demographic and diagnostic information about the subjects is found in Section 5.2.6. The retellings produced by the experimental subjects were transcribed and manually aligned at the word level, as described above for the WLM in Section 7.1.2. The transcriptions themselves were tokenized and downcased, and all punctuation and pause-fillers were removed in preparation for alignment.

The amount of available data is much more limited for the NNM than for the WLM. There are only 119 experimental subjects, in contrast to the 235 experimental subjects available for the WLM, and there is only one retelling for each subject, as opposed to two WLM retellings for each subject. For this reason, we sought to increase our training data in several ways. First, there are roughly a dozen participants in the larger study of prosody who completed the NNM task but dropped out of the study or were determined to be ineligible. The retellings for these subjects were aligned at the word level and will be used to tune the graph-based system. There is an additional set of 50 children who completed the NNM at a different data collection site but whose diagnoses were not available. Finally, we collected 97 retellings of the story from neurotypical adults. All three of these sources of additional retellings will be used for training the word alignment models.

We also carried out a specialized form of manual scoring in which every phrase from a retelling transcription matching any one of the 17 story elements was identified. From these annotations, we created a supplementary set of parallel data: a phrase-to-phrase corpus, consisting of story elements on the source side and phrases from retellings corresponding to those story elements on the target side. These annotations were also available for the data collected from neurotypical adults. In cases where multiple items were listed as matches for a particular story element in the guidelines, the annotator picked the specific item that best matched the retelling phrase. Only the element-level alignments for the non-experimental subjects were used in the experiments presented here. Examples of these item-specific element-level alignments are shown in Figure 7.22. Table 7.14 summarizes all of the data used to train the word alignment models for the NNM.

Element	Corresponding phrase
tree with branches too high for jim to reach	he couldn't reach the branches
pepper went to get help	pepper was going for help
shoe fell	lost one of his shoes
carried a ladder to the oak tree	got a ladder and put it up against the tree

Figure 7.22: Example element-level alignments.

Corpus	Lines of training
experimental retelling to artificial source	119
experimental retelling to actual source	119
non-experimental subject retelling to artificial source	62
non-experimental subject retelling to actual source	62
neurotypical adult retelling to artificial source	97
neurotypical adult retelling to actual source	97
manual element-level alignments for neurotypical adults	1136
word identity alignments	1052
all retellings to all other retellings	75,000

Table 7.14: Corpora used to build NNM word alignment model

The final difference between the data organization used in the WLM and the data used here is the content of the source narrative. In the WLM, every retelling was aligned to the actual source narrative, since the source narrative contained all and only story elements. In the NNM, the story elements are a subset of the events related in the source narrative. A reteller who is completely faithful to the source narrative will produce a large number of words and phrases that are entirely unrelated to the story elements. (This issue is discussed in detail in Section 5.2.2.) Recall also that the source narrative itself is extremely long, with over 200 words. For these reasons, we created an artificial source narrative composed of all of the story elements concatenated into a single "sentence". From the set of story elements shown in 7.21, this yields an ungrammatical string with multiple references to the same event, but it contains all and only the information necessary for scoring a retelling according to the published guidelines. It is from an alignment of a retelling to this artificial source narrative that the scores will be extracted. Table 7.14 lists each of the parallel NNM corpora used to build word alignment models.

7.8.2 Alignment accuracy

We first built a word alignment model on all of the above corpora except the final retelling-to-retelling corpus, using the Berkeley aligner with 2 iterations of IBM Model 1, one iteration of the HMM, and a posterior threshold of 0.5, which was the best performing EM iteration and threshold configuration for the WLM data. We refer to this model as the small Berkeley model. We then built a word alignment model using the Berkeley aligner with the same configuration on all of the above corpora including the large retelling-to-retelling corpus. We refer to this model as the large Berkeley model. From these alignments, we also generated alignments using the graph-based method with posterior-weighted links described in Section 7.7.

Despite all of the creative data leveraging described above, alignment accuracy was substantially lower on the NNM data than on the WLM data. In Table 7.15, we show precision, recall, and AER for this Berkeley model and several configurations of the graph-based model built on the output of the large Berkeley model. Overall, the graph-based method using posterior-weighted links resulted in improvements almost twice as large in absolute terms than those observed in the WLM data, although the AER remains quite high. In addition to showing the results for the model with the lowest AER, we also show results for a high precision model with AER comparable to the best model, and the highest recall and highest precision models that achieved an AER below the AER of the large Berkeley model. We see that the minimum posterior alignment probability required for an alignment to be included as a link in the graph needs to be lowered from 0.5 in order to yield improvements. In addition, the value of λ appears to have less influence on the eventual AER of the alignment produced. All of the AER values are disappointing, but as we have seen in the early attempts at aligning the WLM retellings, it is possible to achieve reasonable scoring accuracy even when extracting scores from a high AER alignment.

7.8.3 Scoring accuracy

Given the weak alignment accuracy reported above, we might expect to have weak scoring accuracy, as well, and indeed the baseline Berkeley models, particularly the small model,

Model	P	R	AER
Berkeley small (Model1x2, HMMx1)	29.9	30.6	69.8
Berkeley large (Model1x2, HMMx1)	54.8	24.5	63.7
Lowest AER graph-based (posterior=0.3, λ=0.2)	48.3	39.1	55.9
High precision graph-based (posterior=0.4, λ=0.6)	57.6	33.2	56.0
Very high precision graph-based (posterior=0.4, λ=0.8)	69.4	27.8	57.3
Very high recall graph-based (posterior=0.2, λ=0.3)	36.9	44.3	60.4

Table 7.15: Alignment accuracy for NNM.

Model	P	R	F1
Berkeley small (Model1x2, HMMx1)	41.4	49.9	45.2
Berkeley large (Model1x2, HMMx1)	64.5	48.3	55.2
Lowest AER graph-based (posterior=0.3, λ=0.2)	57.7	79.2	66.8
High precision graph-based (posterior=0.4, λ=0.6)	64.5	68.0	66.2
Very high precision graph-based (posterior=0.4, λ=0.8)	72.4	55.1	62.6
Very high recall graph-based (posterior=0.2, λ=0.3)	49.9	86.0	63.2

Table 7.16: Scoring accuracy for NNM.

performed quite poorly, as shown in Table 7.16. The large graph-based models, however, showed a very large improvement in scoring accuracy over the Berkeley models, with the lowest AER model achieving the highest F-measure and the highest precision. Because of the large differences in precision and recall both in terms of AER and scoring F-measure among these four graph-based models, we will continue to explore these models, using their output scores for diagnostic classification.

7.8.4 Classification accuracy

We performed diagnostic classification for language impairment (LI) using these automatically derived scores as features within an SVM. We consider two feature sets: the set consisting of the summary score and the set consisting of 17 binary features representing the 17 story elements. Again, we evaluate the performance of the classifier using leave-pair-out validation. Despite the relatively mediocre scoring accuracy, classification accuracy on scores built from the graph-based alignments are often comparable to that achieved using manual scores. The model with the very high recall in word alignment

Model	Sum. (s.d.)	Elem. (s.d.)
Berkeley large (Model1x2, HMMx1)	52.2 (5.8)	65.7 (5.4)
Lowest AER graph-based (posterior=0.3, λ=0.2)	76.9 (4.6)	75.6 (4.7)
High precision graph-based (posterior=0.4, λ=0.6)	76.2 (4.7)	74.7 (4.8)
Very high precision graph-based (posterior=0.4, λ=0.8)	72.7 (5.0)	73.2 (4.9)
Very high recall graph-based (posterior=0.2, λ=0.3)	77.5 (4.6)	78.0 (4.5)
Manual scores	79.6 (4.4)	79.1 (4.4)

Table 7.17: LI classification accuracy for the NNM (AUC).

produced the best classification results, suggesting that future word alignment strategies for this particular instrument should focus on maximizing recall.

7.8.5 Discussion

Because of the various challenges presented by the NNM, including the length of the narrative, the inconsistency between the source narrative and the story elements, and the sparsity of the available data, we were not able to achieve particularly satisfying alignment or scoring accuracy numbers. We did find, however, that AER could be greatly improved using random walks on graphs to realign the data using the pairwise alignments produced by the Berkeley aligner. In addition, although scoring accuracy was weak, the accuracy of classification performed with the automatically extracted scores was comparable to that performed using manually assigned score features.

7.9 Summary

In this chapter, I presented a novel method for using word alignment to automatically score a narrative retelling according to the published scoring guidelines for two different neuropsychological assessment instruments. I showed that these automatically extracted scores can be used for diagnostic classification at accuracy levels comparable to those achieved using manually assigned scores. In addition, I presented a series of very effective techniques for leveraging the small amount of available retelling data to improve the alignments generated by an existing machine translation word alignment package. The graph-based method in particular decreased alignment error by a large margin on every

set of alignments tested. In the next chapter, I turn to alternate methods of scoring narrative retellings that stray from the published guidelines. I will also explore several automated narrative fidelity evaluation methods that are entirely independent of the standard scoring procedure. These evaluation methods result in very accurate classification, while enabling us to explore more thoroughly the specific narrative deficits associated with different neurological disorders.

Chapter 8

Alternative scoring and narrative features

In the previous chapter, I presented a technique for automatically extracting from narrative retellings scores corresponding to the scores reported under the published scoring guidelines for two different narrative instruments. In this chapter, I will explore several alternatives for analyzing the fidelity of narrative retellings. First, I apply to our retelling data two alternative manual scoring procedures that have been proposed in the psychology literature with the goal of either improving reliability or increasing the sensitivity of the test to particular cognitive deficits and impairments. I then describe how to automate all of these alternative scoring methods using the alignment-based techniques described in the previous chapter. I compare the diagnostic sensitivity of these two methods, which rely on the predetermined story element list, to a scoring method that considers each open-class word as a separate classifier feature. Moving away from automated scoring, I then explore several features that are related to general narrative coherence and topicality. Finally I compare these approaches, many of which rely on a priori knowledge of the identities of the story elements, to three algorithms used in NLP research to measure the similarity of two language samples. I evaluate the effectiveness of all of these narrative fidelity assessment methods in terms of their ability to distinguish the two key diagnostic groups discussed so far: MCI and LI. It is also in this chapter the we will finally see evidence that narrative analysis can be used to distinguish subtypes of autism from language matched controls.

8.1 Alternative scoring

In the standard scoring procedures for both the WLM and the NNM, the examiner counts the number of story elements recalled according to the guidelines in the published manual for the instrument. The standard scoring guidelines for both instruments allow for some degree of lexical substitution and paraphrasing. In the case of the WLM, the guidelines are relatively explicit, with many elements requiring verbatim recall. The guidelines for the NNM, however, are often vague, leaving the permissible set of paraphrases and substitutions for many elements open to interpretation or entirely ambiguous.

8.1.1 Verbatim and veridical scoring

As I discussed in Chapter 5, a number of alternative scoring methods have been proposed for the WLM. I will discuss two of these in the context of both the WLM and NNM: verbatim scoring and veridical scoring. Verbatim scoring, in which the subject must produce a story element word-for-word in order to receive credit for that element, was proposed by Abikoff et al. (1987) as a remedy for the poor test-retest consistency and inter-rater agreement of the original formulation of the test. This scoring technique was found to have higher inter-rater correlations than the WLM scoring in use at the time and to be more sensitive to memory impairment characteristics found in the target population of the study.

The veridical scoring procedure, outlined by Johnson et al. (2003), is similar to verbatim scoring, but it allows for variation in morphology, anaphora, and deletion, insertion, and substitution of function words. As in verbatim scoring, the element *Anna* would only be scored as correct if the subject actually uttered the word "Anna", but *took up a collection* could be rendered as, for instance, "take a collection". The authors' arguments for verbatim scoring include both its high inter-rater reliability and its significantly greater power to distinguish people with mild dementia from typically aging controls.

Recall that we have manual phrase-level alignments available for all of our experimental subjects. In these alignments, every word or phrase corresponding to or containing a story element has been identified manually by a trained labeler and then verified by a second

party. From these phrase alignments, we can easily extract manual verbatim scores without resorting to any additional manual annotation. In order to score the verbatim recall of a retelling, we consider each manual phrase level alignment, and search on the retelling side for the substring that is the verbatim story element. This is preferable to using `grep` since `grep` sometimes finds matches that would not be permissible according to the scoring guidelines. The element *for her*, for instance, would only count as a match for the 25th story element if it actually indicated that the money collected was for her or her children, as outlined in the scoring guidelines. The manual alignments guarantee that the matches found were confirmed by a labeler to be a correct match for the target element.

Manual veridical scores were extracted from the word alignments rather than the phrase alignments. First we applied the Porter stemmer (Porter, 1980) to each content word in the source narrative and every retelling. Then, for each stemmed content word in each story element, we determined whether the manually aligned retelling word-stem was identical. If every stemmed content word in a story element was aligned to an identical stemmed word in the retelling, then that element was considered to be correctly recalled. In this way, we were able to ignore the variations in function word use and anaphora that are permissible under the veridical scoring procedure, as outlined in Johnson et al. (2003). Since both of these methods make reference to the story elements of the source narrative, they yield two types of scores per retelling: a summary score (0-25 for each of the two WLM retellings, 0-17 for the NNM); and a set of binary per-element scores having a value of 0 if the element was not recalled and 1 if the element was recalled. We will use these scores in the same way they were used in Chapter 7 to classify impaired subjects.

Both of these manual scoring procedures can trivially be automated. To automatically generate the verbatim scores, we simply use the Unix utility `grep` to search for each of the story elements within a retelling. Note that this will not result in perfect scoring precision, since a subject can use the words of a story element without actually correctly calling that element, as was illustrated with the case of *for her*, above. In our automated approach to veridical scoring, we generate a word alignment of the type described in Chapter 7. We discard any alignments between two words in which the retelling word is not a morphological variant of the source word. From this reduced set of alignments,

	(a)	(b)	(c)	(d)	(e)	(f)
(a) standard-manual	1.00	0.84	0.81	0.90	0.93	0.89
(b) standard-auto	0.84	1.00	0.68	0.77	0.82	0.91
(c) verbatim-manual	0.81	0.68	1.00	0.91	0.84	0.80
(d) verbatim-auto	0.90	0.77	0.91	1.00	0.90	0.87
(e) veridical-manual	0.93	0.82	0.84	0.90	1.00	0.92
(f) veridical-auto	0.89	0.91	0.80	0.87	0.92	1.00

Table 8.1: Correlations between scoring procedures for the WLM.

	(a)	(b)	(c)	(d)	(e)	(f)
(a) standard-manual	1.00	0.58	0.56	0.58	0.45	0.61
(b) standard-auto	0.58	1.00	0.76	0.79	0.79	0.89
(c) verbatim-manual	0.56	0.76	1.00	0.93	0.88	0.79
(d) verbatim-auto	0.58	0.79	0.93	1.00	0.84	0.80
(e) veridical-manual	0.45	0.79	0.88	0.84	1.00	0.79
(f) veridical-auto	0.61	0.89	0.79	0.80	0.79	1.00

Table 8.2: Correlations between scoring procedures for the NNM.

we then extract story elements just as we did for the original alignment. Each procedure yields the set of summary and element-level scores, described above, which can then be used for diagnostic classification of MCI for the WLM and LI for the NNM.

Both of these scoring alternatives were proposed specifically for the WLM subtest, but we investigate their utility for the NNM, as well. The veridical and verbatim scoring approaches might prove to be especially interesting for the NNM, since scoring under the published guidelines is difficult and prone to unreliability, as discussed in Section 5.2.2. The summary scores derived from both the manual and automatic alternative scoring methods are highly significantly different between groups in both the WLM data and the NNM data (all $p < 0.00001$). In the WLM data, the summary scores derived via the two alternative scoring measures correlate highly with scores derived using the standard scoring procedure, but the correlations are not perfect, as shown in Figure 8.1. The NNM data produce a very different pattern. The verbatim and veridical summary scores do not correlate particularly well with the standard summary scores, which may lead to differences in classification accuracy when using verbatim and veridical scores as features.

Figures 8.1 and 8.2 show, for the WLM and NNM respectively, the classification accuracy for both the manual and automatic versions of each scoring method (standard scoring procedure, veridical scoring, and verbatim scoring) using both the summary scores and elements as features. We continue to use an SVM and evaluate with leave-pair-out cross-validation. In the WLM, veridical scoring provides more discriminative power than the other two scoring methods, which is in line with the results presented in Johnson et al. (2003). The data in the NNM reveal a different pattern. In this case, the verbatim scores, both manually and automatically derived, yield the highest classification accuracy. Interestingly, the automatically derived scores achieve higher classification accuracy than the manually assigned scores. Although the reasons for this are not clear, we do know that the scores extacted from TD retellings are more accurate than those extracted from LI retellings. It is possible that so few elements can be extracted automatically from the LI retellings that the differences in recall performance between the TD group and the LI are more pronounced when derived from automatic alignments.

In neither dataset does the standard scoring procedure offer the best classification performance. Both veridical and verbatim scoring were originally proposed to improve inter-rater reliability. The results presented here provide evidence that the standard scoring procedures for both tests are less than optimal in terms of both reliability of scoring and diagnostic accuracy.

Although I have discussed ASD frequently in this thesis, I have yet to attempt to distinguish the ASD groups from the groups without ASD. Most research shows that when controlling for language ability, there are few differences in ASD and TD narrative recall particularly in terms of narrative content, the target of standard scoring measures (Tager-Flusberg, 1995; Capps et al., 2000; Diehl et al., 2006). The few differences that are reported are often described in impressionistic ways in the literature. For instance, it has been reported that the retellings of children with ASD sound more like recitations than retellings (Loveland et al., 1990; Loveland and Tunali, 1993). Perhaps this impression is caused by a tendency of children with ASD to recall content word-for-word from the source narrative rather than producing approximations of the source content. Although there are no differences between the TD and ALN group in the verbatim summary scores or in the

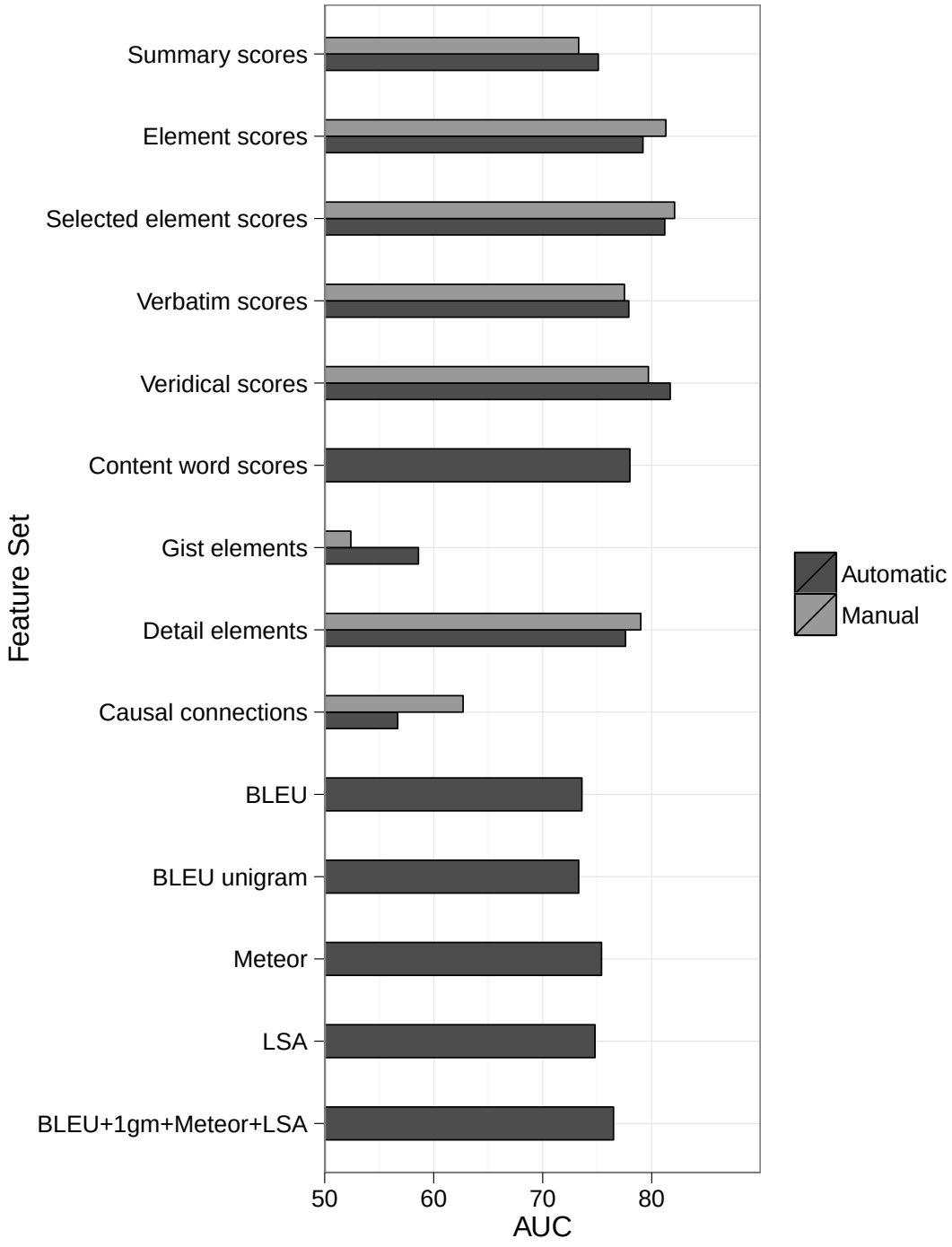

Figure 8.1: MCI classification performance for alternative scoring methods and narrative features extracted from WLM retellings.

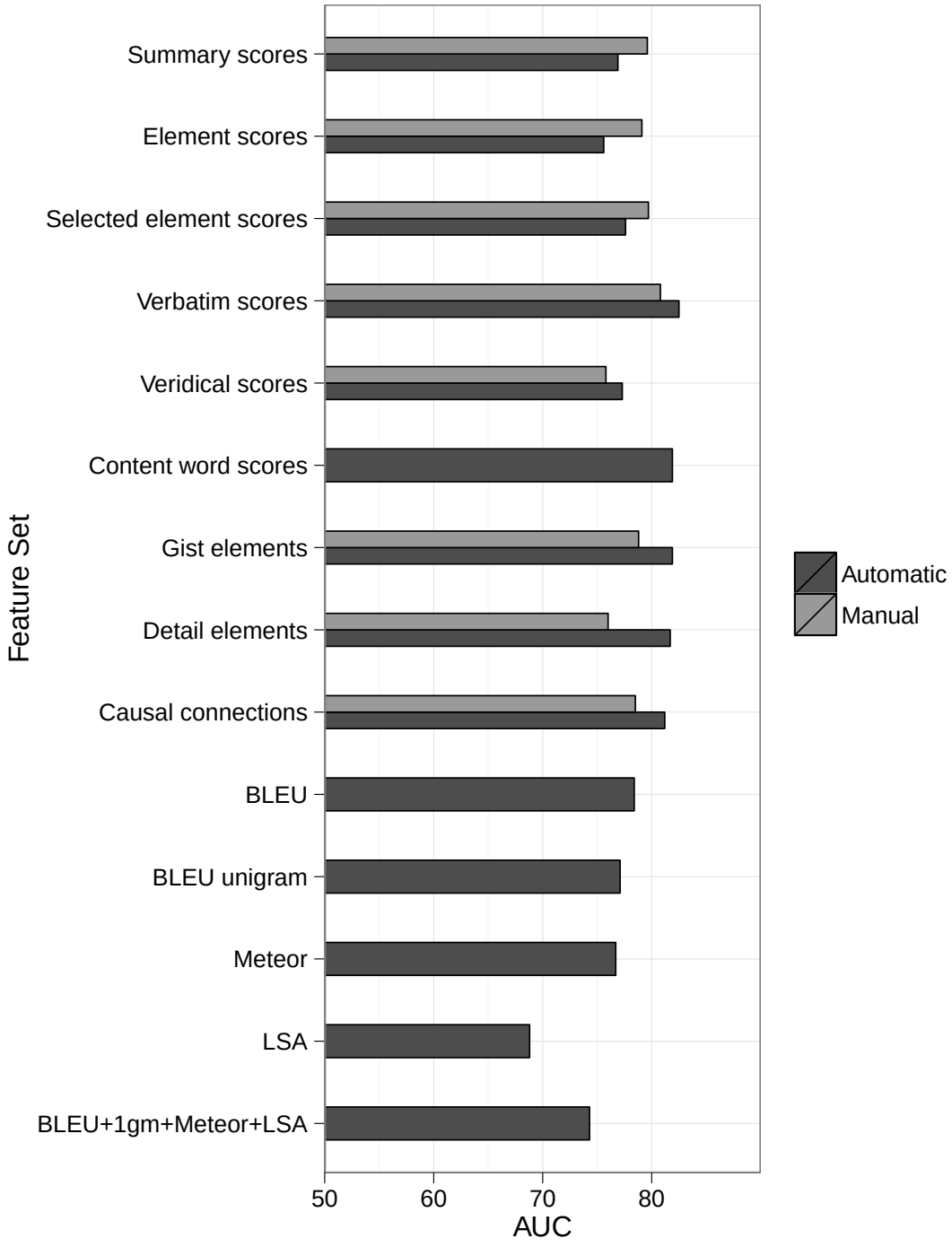

Figure 8.2: LI classification performance for alternative scoring methods and narrative features extracted from NNM retellings.

standard scores, as measured by Student's t, there was a significant difference in manually derived veridical recall scores ($p < 0.01$). In addition, veridical element-level and summary scores were the only SVM scoring features that provided reasonable classification accuracy for ALN (AUC=0.721).

8.1.2 Content word scoring

The above methods of scoring, while varying slightly from the published scoring guidelines for both the WLM and NNM, nevertheless rely directly on the predetermined list of story elements. This limits our ability to extend these scoring approaches to other narrative tests, since many such tests do not include an explicit set of items that must be recalled. Recall that the automated scoring method we propose uses only the open-class or content words in the source narrative. We now propose a scoring technique that considers each content word in the source narrative to be its own story element. Any word in a retelling that aligns to one of the content words in the source narrative is considered to be a match for that content word element. This increases the number of elements, but it allows the scoring method to be easily adapted to other narrative production scenarios that do not have explicit scoring guidlines. In addition, this might resolve some of the problematic aspects of NNM scoring that were discussed in Section 5.2.2. Story element 10 in the standard NNM scoring guidelines, for instance, included several discrete events including *shoe fell, slipped, ladder fell, got stuck,* and *couldn't get down.* If we consider each content word to be a separate element, we might be able to reward subjects who relate the entire sequence of events while penalizing subjects who only remember one or two of the events, which has the potential to improve classification accuracy.

In Figure 8.1, we see that treating each content word in the source narrative as a feature in the classifier yields classification accuracy lower than but within range of the standard element-level scores for the WLM. Applying this scoring method to the NNM data, however, results in an improvement over both the automatic and manual standard element-level scores 8.2.

8.2 Other narrative features

The scoring approaches we have discussed so far are simply variations on the published scoring guidelines. We now turn to other features that characterize the narrative structure and fidelity of a narrative to the source. In particular, we explore features that correspond to two of the phenomena unrelated to scoring that are reported in the literature to characterize narrative retellings in special populations. We will continue also to try to distinguish children with ASD (ALN and ALI) from their language-matched peers (TD and SLI).

8.2.1 Gist, details, and causal coherence

The ability to recall the details of a narrative rather than simply the overall gist seems to be a distinguishing feature for MCI (Haaland et al., 1983). Children with autism, in contrast, tend to produce connections between gist-level events (Diehl et al., 2006). Although some authors use the word "gist" to mean paraphrased rather than verbatim recall (Abikoff et al., 1987), others consider "gist" to mean the parts of the story that, in contrast to details, contribute to the unfolding of events. Following Diehl et al. (2006), we appeal to this latter definition of gist, as it is described in work by Trabasso and colleagues (Trabasso et al., 1984; Trabasso and van den Broek, 1985). This definition relies on the notion of a causal relatedness. Events A and B share a causal relation if event B could not have occurred without event A occurring first. Adapting these ideas to our existing set of manual scores and annotation, we consider each story element to be an event. Gist-level story elements are those elements that stand in a a causal relation with another element. Details are those elements that are off the causal chain; that is, they are the elements that have no causal relations. For each story element in the WLM and the NNM, we used these criteria to manually determine whether the element was a gist element or a detail. We then counted the number of outgoing causal connections for each story element. Using this information and the manual per-element scores for each subject, we then calculate 3 scores for each of the subject's retellings: (1) the number of details recalled; (2) the number of gist elements (i.e., elements in a causal relationship) recalled; and (3) the total number of causal connections in the retelling.

To automate the extraction of the gist and detail scores, we generate a word alignment and extract story element identities from that alignment. Once the identities of the story elements are determined automatically in this way, we simply apply the same method of counting gist elements, detail elements, and causal connection that were used to generate these features from the manual scores.

In the WLM data, the MCI group recalls significantly fewer details but a comparable number of gist elements and causal connections, as was predicted in the literature. This results in high classification accuracy when the raw number of details in a retelling is used as a feature within the SVM but weak performance when using the count of gist elements or causal connections, as shown in Figure 8.1. Thus it seems that the major difference in recall between MCI subjects and typically aging controls is their ability to recall details, which results in a lower overall summary score for the MCI group. The NNM data show a different pattern: the LI group consistently produces fewer details, gist elements, and causal connection, and classification performance is consistently strong regardless of the measure being used, as shown in Figure 8.2. In neither dataset does combining all three features in the classifier result in improvements in classification accuracy over the best-performing of the individual features. Although there was a slight significant difference in the number of causal connections between the TD and ALN group, similar to the results reported in Diehl et al. (2006), this difference was not observed between the ALI and SLI groups.

8.2.2 Intrusions

Subjects with both MCI and autism were reported to produce more intrusions in their narratives (Ulatowska et al., 1988; Chapman et al., 1995; Creamer and Schmitter-Edgecombe, 2010; Loveland et al., 1990; Losh and Capps, 2003; Goldman, 2008). In order to try to capture this phenomenon, we attempted to approximate the amount of off-topic or irrelevant content in each retelling by determining the percentage of words in a retelling that could not be aligned to the source narrative. This feature was not a good predictor of MCI in the WLM data, but it did perform well in the NNM distinguishing TD from ALN (AUC=0.71).

Sentences	Type of intrusions
they took up a collection to help her *feed her children*	true inference
she was held up *on her way home from work*	reasonable inference
this lady evidently *was traveling* and someone had taken her purse	inaccurate inference
she was a cook *so why couldn't she tote some of the food home*	reasonable comment
that happened to a lady here in Clackamas	conversational comment
something about an apple	unrelated and inaccurate

Figure 8.3: Sentences with substrings that cannot be aligned to a source narrative.

We note that although this is in fact a measure of how much of a person's retelling is not related to the source narrative, it fails to distinguish reasonable inferences about the story, questions to the examiner, and socially appropriate personal asides from entirely wrong or irrelevant information. In Figure 8.3 gives examples from our data of the different kinds of strings that would not be able to be aligned to the WLM source narrative. Because we are not differentiating these intrusions, we do not necessarily expect that the intrusion feature will necessarily provide the degree of diagnostic power that is suggested in the literature. We also note that varying results are reported for this phenomena in the literature, often because of coding reliability problems.

8.3 NLP techniques for evaluating text similarity

Many of the evaluation methods so far make reference to the published scoring guidelines and the predefined element boundaries. We now explore three methods of evaluating the WLM that do not rely in any way on knowledge of the predefined list of story elements. These measures, which we refer to as the *unsupervised* automated measures, were originally developed to estimate the similarity of two texts. The technique using latent semantic analysis has been used previously to approximate WLM scoring (Dunn et al., 2002), while the two machine translation-based techniques have not previously been used for neuropsychological evaluation purposes, to our knowledge. All three of these unsupervised methods produce a single score for each retelling, ranging from 0 to 1. A score of 0 for any of these measures indicates that the retelling was poor because it contained no words, stems, synonyms, paraphrases or semantic themes in common with the source.

BLEU (Papineni et al., 2002), which was discussed in some detail in Section 3.4, is an n-gram overlap precision metric commonly used to evaluate the quality of machine translation (MT) output. When calculating BLEU on the retelling data, we assume that in the case of the WLM, the source narrative is the reference sentence and the retellings are candidate sentences. Recall, however, that since we were attempting to automatically score the NNM using the existing scoring guidelines, we aligned the NNM retellings to the list of story elements rather than to the source narrative. (See Chapter 7 for a full discussion.) Since we are now trying to investigate features beyond those defined by the guidelines, we can consider both the list of story elements and the source narrative as references. In addition to considering the full BLEU score, we will also look at unigram precision, as proposed by Hakkani-Tur et al. (2010) in their work on scoring narratives using ASR.

Meteor (Denkowski and Lavie, 2011), which was also discussed in Section 3.4, is similar to BLEU but also searches for overlap between synonyms, paraphrases, and stems. When calculating Meteor, we again treat the source narrative as the reference for the WLM and both the source narrative and the story element list as the references for the NNM retellings.

The third and final approach to unsupervised evaluation of WLM retellings uses latent semantic analysis (LSA) (Landauer et al., 1998). LSA is a way of capturing the *semantic* rather than lexical similarity between two texts by measuring the cosine similarity between the two texts in a high-dimensional semantic space. LSA is commonly used in NLP applications such as information retrieval and automated essay scoring. Previous research by Dunn et al. (2002) found a high correlation between LSA cosine similarity of a retelling and the source narrative and WLM scores assigned manually according to the published guidelines. Following this work, we use the University of Colorado's LSA web interface (available at `http://lsa.colorado.edu/`) to calculate the retelling-to-source cosine similarity for each retelling. For the WLM retellings, we select the 300-factor ninth-grade reading level topic space, while for the NNM, we instead use the third-grade reading level topic space, which was the most developmentally appropriate choice available. The feature

used for classification was the cosine similarity, ranging from 0 to 1, between each retelling and the source narrative.

All four measures yielded high classification accuracy for both MCI and LI, as can be seen in Figures 8.1 and 8.2. We also see that these features do not improve classification accuracy significantly when used collectively as classifier features. Overall, the classification results for these automated measures are often comparable to summary-level manual methods. It seems, however, that the additional power of the element-level scores to classify the groups cannot be adequately captured by these standard NLP text similarity metrics. We also note that these measures were not able to distinguish the ASD groups from their language-matched control groups.

8.4 Summary

In this chapter, I moved beyond automatic scoring of narrative retellings according to the published guidelines to explore other ways of analyzing narrative fidelity both manually and automatically. Some of these approaches, such as the alternative scoring measures and the use of features related to gist and intrusions, were grounded in research in psychology, while others had foundations in the NLP literature. I showed that many of these features, when extracted automatically, can distinguish and classify diagnostic groups as reliably as scores extracted automatically. It seems, however, that many of the most accurate automated measures benefit from knowledge of the identities of the story elements. All of the known previous work in automated analysis of narrative retellings (Dunn et al., 2002; Lautenschlager et al., 2006; Hakkani-Tur et al., 2010) disregards the story element boundaries provided in the scoring sheets for the instruments. The work presented here provides evidence that leveraging knowledge of the identities of the story elements, whether defined in the scoring guidelines or at the content-word level, can result in classification improvements. In the next and final chapter, I describe a few experiments in expanding the system presented here for alignment and scoring of narrative retellings for diagnostic classification.

Chapter 9

Extensions

In this chapter I present a few methods of expanding the techniques I have presented in the preceding chapters for automatically aligning and scoring clinically elicited narrative retellings for diagnostic classification in order to demonstrate the utility of these techniques. First I will demonstrate how to adapt the approach to a very different elicitation paradigm: a picture description task. I then discuss collaborative work on incorporating ASR into the pipeline in order to create an end-to-end system for diagnostic screening.

9.1 Tests with non-linguistic reference

The Boston Diagnostic Aphasia Examination (BDAE) (Goodglass and Kaplan, 1972), a widely used instrument used to diagnose language impairments in adults, includes a language elicitation task that is popularly referred to as the Cookie Theft picture description task. In this test, the subject views a drawing of a lively scene in a family's kitchen and must tell the examiner about all of the actions he sees in the picture. The picture is reproduced below in Figure 9.1. Describing visually presented material is obviously quite different from a task such as the WLM, in which language comprehension and memory play a crucial role. Nevertheless, the processing and language production demands of a picture description task may lead to differences in performance in groups with certain cognitive and language problems. In fact, it is widely reported that the picture descriptions of seniors with dementia of the Alzheimer's type (DAT) differ from those of typically aging seniors in terms of information content (Hier et al., 1985; Gilesa et al., 1996). Interestingly, this reduction in information is not necessarily accompanied by a reduction in the amount

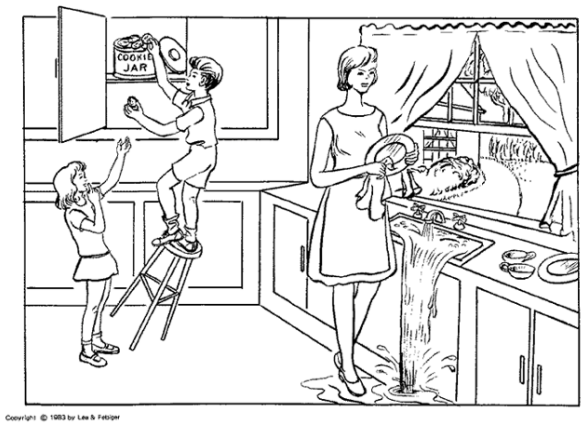

Figure 9.1: BDAE cookie theft picture.

of language produced. Rather it seems that seniors with DAT tend to include redundant information, repetitions, intrusions, and revisions that result in language samples of length comparable to that of typically aging seniors. In this section, I will discuss how I adapted my approach for scoring narrative retellings to the task of analyzing picture descriptions in seniors with and without dementia.

9.1.1 Data

TalkBank (MacWhinney, 2007), the freely available online database of transcribed speech in numerous contexts, has made available a corpus of descriptions of the cookie theft picture by hundreds of individuals, some of whom have one of a number of types of dementia, including MCI, vascular dementia, possible Alzheimer's disease, and probable Alzheimer's disease. From this corpus, I selected a subset of subjects without dementia and a subset with probable Alzheimer's disease (AD). Given the difficulties experienced when aligning the very lengthy NNM narratives, I limited the set of descriptions to those with fewer than 100 words. In addition, I excluded the descriptions of fewer than 25 words, all of which were produced by subjects with AD. This winnowing process yielded 130 descriptions for each diagnostic group. I manually post-processed the descriptions in order to remove all initial and final utterances that were indications that the subject was about to start his description (e.g., *Do you want me to start now?* or *Okay, let me see.*)

or that he had just finished it (e.g., *And I guess that's all.*). I also excluded all information provided after the examiner asked *Anything else?*. There was no significant difference in description word count between the two groups.

9.1.2 Method

The first task was to generate a source description to which all other narratives should be aligned. Working under the assumption that the control subjects would produce good descriptions, I calculated the BLEU score of every pair of descriptions from the control group. The description with the highest average pairwise BLEU score was selected as the source description. I confirmed that it did in fact contain all of the action portrayed in the picture, and I removed all extraneous conversational asides from the description in order to ensure that it contained all and only information about the picture. I then segmented the description into information elements following the style of the WLM. The selected source description is as follows, with slashes indicating boundaries between the 20 elements:

> The boy / is getting cookies / out of the cookie jar. / And the stool / is just about / to fall over./ The little girl / is reaching up / for a cookie. / And the mother / is drying / dishes. / The water / is running / in to the sink / and the sink / is running over / onto the floor. / And that little girl / is laughing./

I then built an alignment model on the full pairwise description parallel corpus (260^2 = 67,600 sentences) and a word identity corpus consisting of each word in each retelling reproduced 100 times. Using this trained model, I then aligned every description to the artificial source description described above. From these alignments I extracted the story elements as described at length in Chapter 7. In addition to using a story-element based scoring system, I also used the content-word scoring system described in Section 8.1.2, in which each content word in the source description is treated as a story element. Using these two kinds of scores as feature sets, I used an SVM to classify the two diagnostic groups, typically aging and probable AD, and evaluated the classifier using leave-pair-out validation.

Feature Set	AUC (s.d.)
Unigram precision	63.0 (3.4)
BLEU	70.1 (3.2)
Story element scoring (summary)	68.2 (3.3)
Story element scoring (elements)	77.0 (2.9)
Content word scoring (summary)	67.6 (3.3)
Content word scoring (words)	83.2 (2.5)

Table 9.1: Classification accuracy using BDAE picture description features.

9.1.3 Results

Table 9.1 shows the classification results using story-element scoring features and content-word scoring features. These can be compared to classification results using the simpler similarity metrics BLEU and unigram precision, described in Section 8.3. We see that using element-level and word-level features results in significantly higher classification accuracy than both the simple similarity metrics and the summary scores. In particular, the content word scoring word-level features yield remarkably high accuracy given the somewhat ad-hoc selection of the source narrative from the set of control retellings.

These results demonstrate the flexibility and utility of our approach to scoring narratives. Not only can it be adapted to other narrative retelling instruments, but it can relatively trivially be adapted to instruments that use non-linguistic stimuli for elicitation. We plan to apply similar bootstrapping techniques for converting a description task into a retelling task for other similar tasks, such as the ADOS Wordless Picture Book activity narrations (Lord et al., 2002) collected as part of our prosody in autism, the picture book narrations included in the CHILDES database (MacWhinney, 2000), and the picture and video description tasks that are used in one of the ongoing data collection projects taking place at CSLU in conjunction with the Oregon Alzheimer's Disease Center.

9.2 Incorporating ASR

One of the long term goals of this research is to create an end-to-end, fully automated screening tool for MCI. The missing component thus far is automatic speech recognition (ASR) to automatically generate the transcripts that are produced manually in the system

as I have described it in Chapter 7. Using ASR for this task is challenging for a number of reasons. First, since the retellings were recorded in a causal, informal environment, using a microphone attached to a laptop or digital recorder, the quality is not optimal, and there are frequent extraneous noises, such as music, chiming clocks, and nearby conversations. In addition, the speakers are elderly, and features of elderly speech might not be well represented in the acoustic data normally used to train an ASR model. The following research in incorporating ASR into the screening pipeline, in which we adapt both the acoustic model and language model to this particular narrative retelling data in order to improve ASR output quality, were carried out in collaboration with Maider Lehr, Zak Shafran, and Brian Roark.

9.2.1 Data

For this experiment, we separated out a group of 72 testing subjects, 35 with MCI and 37 typically aging. We then created a corpus of retellings for 91 other participants in the larger study to be used for acoustic model training and adaptation, 15 of whom had MCI, yielding around 2 hours of recorded audio. The word alignment training data is that which was used to train the alignment models in Chapter 7, and the classifier training data consisted of the manually assigned scores for all of the subjects with a diagnosis of MCI or non-MCI excluding the 72 testing subjects.

9.2.2 Method

The two hours of held-out WLM retelling audio data were insufficient to train an ASR model. We therefore needed to use our in-domain audio to adapt an existing acoustic model built from the publicly available Broadcast News corpus. The baseline acoustic model was trained on 430 hours of transcribed speech from the Broadcast News corpus, with 4000 clustered pentaphone states. The baseline 4-gram language model was built from transcripts of this same corpus, containing a vocabulary of 84,000 words.

We first adapted the acoustic model (AM) of the baseline system to our data in an unsupervised fashion using maximum likelihood linear regression transforms (Legetter and Woodland, 1995) derived from automatic transcriptions of the two hours of in-domain

ASR Model	Total	Control	MCI
Baseline	47.5	45.0	50.6
AM unsupervised	39.8	36.1	44.3
AM supervised	41.7	37.5	47.0
LM	31.2	25.7	38.1
AM supervised + LM	34.9	29.2	42.0

Table 9.2: Average WER under different adaptation schemes.

audio training data (AM unsupervised). We then performed supervised adaptation using the manual transcripts of the in-domain audio training data (AM supervised). In addition to adapting the acoustic models to the WLM domain, we also interpolated the baseline language model with a language model built on our entire corpus of WLM retellings, excluding those of the test subjects (LM). We decoded the test retellings using these four adapted models along with a model that combines both acoustic model and language model adaptation (AM supervised + LM).

9.2.3 Results

ASR output

We first report results on word error rate (WER) for the five model adaptations described above. In Table 9.2, we see that WER generally decreases as the level of adaptation increases, although there seems to be a slight degradation when moving from unsupervised to supervised acoustic models. We note that the automated WLM scoring approach relies only on content words in the source narrative. For this reason, we also measured the WER of content words in the source narrative and those mentioned in the scoring guidelines. Under this definition of WER, shown in Table 9.3 we see that the supervised model outperforms the unsupervised model, as would be expected. Figure 9.2 provides, for a single retelling, the ASR output for each of the 5 ASR models, along with the manual reference transcription.

ASR Model	Total	Control	MCI
Baseline	46.8	43.4	53.3
AM unsupervised	34.0	30.0	41.4
AM supervised	27.3	23.4	34.5
LM	22.7	17.2	33.1
AM supervised + LM	16.9	11.6	26.8

Table 9.3: Average content word WER under different adaptation schemes.

ASR Model	ASR output
Baseline	well and thompson was held at the u. n. robbed an issue reported to the police station and the two collection for her
AM unsupervised	well and thompson was held open and robbed an issue reported to the police station and they took up a collection for her
AM supervised	well anna thompson was held open and robbed and issue reported to the police station and they took up a collection for her
LM	well anne thompson was held up and and robbed a and she reported to the police station and they took up a collection for her
AM supervised + LM	well anna thompson was held up in and robbed and she reported to the police station and they took up a collection for her
Reference	well anna thompson was held up and robbed and she reported to the police station and they took up a collection for her

Figure 9.2: ASR output of the same retelling recording under different model adaptation schemes.

Automatic scoring and classification

We align the ASR output for each of the retellings to the source narrative using a Berkley alignment model built from all of the retellings used to build models in Chapter 7. From these alignments, we extract scores, as described in Chapter 7. Table 9.4 shows that as WER improves, the accuracy of element extraction in terms of precision, recall, and f-measure correspondingly improves, underscoring the importance of accurate ASR output. We see that this fully automated method of extracting story elements is highly accurate, which bodes well for diagnostic classification.

In order to compare the diagnostic sensitivity of the ASR-derived element features to that of element features assigned manually or derived from manual transcriptions, we build a support vector machine classifier. Rather than using the leave-pair-out validation

ASR Model	WER	P	R	F1
Baseline	46.8	82.3	58.5	68.7
AM supervised	27.3	84.6	74.6	79.2
LM	22.7	84.9	84.5	84.7
AM + LM	16.9	84.0	86.4	85.2
manual transcription	n/a	84.0	91.6	87.6

Table 9.4: Content WER and scoring accuracy under different adaptation schemes.

ASR Model	WER	F1	AUC
Baseline	46.8	68.7	75.8
AM supervised	27.3	79.2	80.4
LM	22.7	84.7	80.4
AM + LM	16.9	85.2	82.5
manual transcription	n/a	87.6	82.6
manual scores	n/a	n/a	80.9

Table 9.5: Scoring and classification accuracy under different adaptation schemes.

scheme that we used in previous experiments, we build a single classification model using the manually assigned scores for all of the subjects with a diagnosis of typical aging or MCI who were not included among the 72 test subjects. We then test that model on all of the 72 test subjects. Table 9.5 shows the classification accuracy for MCI as measured by the area under the receiver operating characteristic curve (AUC) for each of the three acoustic models, alongside the accuracies of classifiers built with element features extracted from manual transcripts and manually assigned element features. We see that as ASR quality improves, classification accuracy also improves. In addition, we find that the ASR-derived features yield classification accuracy comparable to manually-derived features. These results demonstrate the feasibility of using ASR within the scoring and classification pipeline described in this thesis.

9.3 Summary

In this brief chapter, I provided two examples of how the approach I have developed for scoring narrative retellings for diagnostic classification can be extended and expanded. First, I applied the same techniques used for evaluating performance on a narrative recall

task to the BDAE Cookie Theft picture description task. I showed how it is possible to select an artificial "source" description from a set of existing control descriptions which can then be segmented like the WLM and used to score a picture description task. These scores yielded accurate classification of subjects with probable Alzheimer's disease. I then described collaborative work in using ASR to generate transcripts of retellings with the goal of building a fully automated end-to-end diagnostic screening tool for MCI. Despite imperfect ASR output, the scoring and classification accuracy for the automatically transcribed retellings was comparable to that for manually transcribed retellings. Both of the extensions to the overall system demonstrate the utility and flexibility of the system developed in this thesis for using narrative retellings for diagnostic classification.

Chapter 10

Conclusions

10.1 Summary

The primary goal of this thesis was to investigate the reliability and diagnostic utility of automatically scoring and analyzing clinically elicited narrative retellings. I first established that manually assigned narrative recall scores of two different instruments can be used to accurately classify subjects with two distinct neurological disorders: mild cognitive impairment (AUC=0.83) and language impairment (AUC=0.80). The performance of the classifier for MCI is particularly remarkable given the much weaker accuracy (AUC=0.73) of a classifier built using scores from the Mini-Mental State Exam, one of the most widely-used dementia screening instruments.

I then presented a method for extracting these scores automatically from an alignment between a retelling and the source narrative. The alignments produced by existing machine translation-based word alignment software packages were insufficient for this task. I proposed multiple improvements to the existing systems, as well as an entirely new graph-based method of word alignment which required no external data resources, could be performed in a matter of minutes, and yielded a lower alignment error rate than a model that took hours and sometimes days to train. The narrative recall scores extracted from these improved word alignments were very accurate (F1=89.3) and achieved inter-rater reliability values (κ=78.3) as high as those reported for human annotaters. In addition, using the automatically extracted scores as features within a classifier achieved classification accuracy comparable to that achieved using manually assigned scores (AUC=81.6).

Next, I explored several alternative methods of automatically scoring narrative retellings and evaluating their fidelity to the source. Some of these features, in particular content word-based scores and veridical scores, yielded classification accuracy higher than the features derived from the standard manual scoring procedure, which suggests that there is room for improvement in the narrative recall scoring approaches and guidelines that are typically used in neuropsychological evaluation. I also showed that scores derived using relatively simple automated language evaluation algorithms, such as those used to evaluate machine translation, did not perform as well as the features derived via word alignment, underscoring the value of the added complexity of the methods proposed in this thesis.

Finally, I presented two extensions to the scoring, evaluation, and classification system. The first was a method for adapting this system to a structured spontaneous language elicitation task that relies on a visual rather than a linguistic stimulus. The second was an exploration of using ASR to generate the transcripts of recordings of narrative retellings for alignment, scoring, and classification. The relative ease with which the system can be expanded and extended, and the very high levels of classification accuracy (AUC>0.8) in both of these applications demonstrates the substantial flexibility and utility of the system proposed in this thesis.

The demand for simple, objective, and unobtrusive screening tools for neurological disorders such as dementia, autism, and language impairment will continue to grow as the prevalence of these disorders increases. Given the differences in performance on narrative tasks that are reported in these diagnostic groups, the analysis of narratives shows great potential as a component of such a screening tool. The approach presented in this thesis overcomes many of the obstacles inherent in manual narrative evaluation. First, narrative recall scoring procedures are not standardized across instruments, and some, such as the NNM, have not been subjected to a rigorous analysis of validity and reliability. Our approach to scoring narratives is consistent for all narrative recall tasks, and can even be applied to other semi-structured spontaneous language evaluation instruments. Secondly, even under the refined guidelines of an established instrument such as the WLM, scoring is inherently subjective. The approach presented here is entirely objective, since it is performed automatically. In addition, if ASR is included in the system, scoring narrative

retellings involves no human intervention at all, making the scoring and screening process more objective and efficient. Finally, none of the existing instruments differentiates between story elements in their scoring procedures. In this thesis, we presented evidence that the identities and characteristics of the individual story elements may hold the key to distinguishing diagnostic groups.

10.2 Future work

Although we were able to achieve high classification accuracy using scores extracted from automatically generated alignments, there is still a great deal of improvement to be made in the word alignments themselves. One possible approach is to explore ideas for enhancing the graph-based method. There is, for instance, currently no way to generate one-to-many alignments within the graph-based system, which could potentially be important for aligning paraphrases and idiomatic expressions. One way to do this could be to select multiple source words over the distribution that is produced by the series of random walks. We would also like to experiment with including non-directional edges and outgoing edges on source words. Another idea for improving the word alignments, which I resisted throughout this thesis, is to begin to incorporate external resources. First, it is likely that we will see improvements at least in the alignment of function words by including external parallel monolingual corpora in the word alignment training data. Another possibility would be to include the sorts of dictionaries, ontologies, and paraphrase corpora that are used in many of the monolingual alignment techniques proposed in the literature for textual entailment, summarization, and machine translation evaluation.

We discussed only a few narrative fidelity features in Chapter 8, but there is more work to be done both in terms of improving the existing features and in engineering new features. We noted, for instance, that the increased use of intrusions in narratives of special populations has been reported in the literature, but that different authors define "intrusion" in different ways. Our method of identifying intrusions in a narrative was accurate in the sense that it did isolate words that were not part of the retelling, but it was unable to distinguish between appropriate comments or inferences and inappropriate

comments or incorrect information. Being able to separate these types of intrusions might shed more light on the ways that different diagnostic groups use intrusions in their narratives. We would also like to determine whether the ordering of events in a narrative might provide some diagnostic information. We had developed a technique for assessing element ordering in previous work, but the feature did not contribute to improved classification. The story elements are not necessarily events in the narrative, and many of them have no natural relative order, which makes measuring the ordering of events based on the set of extracted elements somewhat difficult. Finally, we plan to analyze the use of repetition in retellings to see whether repeating specific elements or content in general might be a useful diagnostic feature.

There is an obvious extension of the research presented here to the common NLP task of word alignment. The graph-based method that we used to align multiple retellings might be able to be adapted to multilingual word alignment of a corpus such as Europarl. The framework as it exists could be used for aligning a small corpus, but we envision instead a way of using the alignment distributions generated by random walks to update the translation and distortion probabilities estimated by baseline alignment model. The updated probabilities could then be used to generate alignments using Viterbi or posterior decoding, potentially resulting in improved AER. Such techniques might be useful for aligning comparable corpora, as well.

Most previous research in applying NLP and computational linguistics algorithms and techniques to the medical domain has focused on general tasks that are already the focus of NLP research, such as named entity extraction or summarization, in the context of biomedical data. In these cases, the problem to be solved is defined in advance as an NLP task. In contrast, I first presented in this dissertation a problem that exists in the medical community, namely the lack of objective and simple screening tools for increasingly common neurological disorders. Having identified this problem, I then proposed methods for modifying existing NLP algorithms that are used for entirely different purposes to the task of diagnostic screening. Although automated analysis of language for diagnostic purposes is not a mainstream NLP research area, it is my hope that the work presented in this

dissertation will serve as an impetus for others to explore the ways in which natural language processing algorithms can be adapted and applied to small datasets of spontaneous, clinically elicited, or otherwise non-canonical language.

Bibliography

Abikoff, H., Alvir, J., Hong, G., Sukoff, R., Orazio, J., Solomon, S. and Saravay, S. 1987. Logical Memory subtest of the Wechsler Memory Scale: Age and education norms and alternate-form reliability of two scoring systems. *Journal of Clinical and Experimental Neuropsychology* 9, 435–48.

Agirre, E. and Soroa, A. 2009. Personalizing PageRank for word sense disambiguation. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 33–41.

Alfonseca, E. and Pérez, D. 2004. Automatic assessment of open ended questions with a BLEU-inspired algorithm and shallow NLP. In *Advances in Natural Language Processing*, volume 3230 of *Lecture Notes in Computer Science*, pages 25–35.

American Psychiatric Association. 2000. *DSM-IV-TR: Diagnostic and Statistical Manual of Mental Disorders*. Washington, DC: American Psychiatric Publishing.

Applebee, A. 1978. *The child's concept of a story*. Chicago: University of Chicago Press.

Artero, A., Tierney, M., Touchon, J. and Ritchie, K. 2003. Prediction of transition from cognitive impairment to senile dementia: a prospective, longitudinal study. *Acta Psychiatrica Scandinavica* 107, 390–393.

Attali, Y., Bridgeman, B. and Trapani, C. 2010. Performance of a generic approach in automated essay scoring. *The Journal of Technology, Learning and Assessment* 10(3).

Attali, Y. and Burstein, J. 2006. Automated essay scoring with e-rater® V. 2. *The Journal of Technology, Learning and Assessment* 4(3).

Baddeley, A. 1992. Working memory. *Science* 255(5044), 556–559.

Baddeley, A. and Hitch, G. 1974. Working memory. *The Psychology of Learning and Motivation* 8, 47–89.

Bamberg, M. and Damrad-Frye, R. 1991. On the ability to provide evaluative comments: further explorations of childrens narrative competencies. *Journal of Child Language* 18, 689–710.

Bannard, C. and Callison-Burch, C. 2005. Paraphrasing with Bilingual Parallel Corpora. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, pages 597–604.

Baron-Cohen, S., Leslie, A. and Frith, U. 1985. Does the autistic child have a theory of mind? *Cognition* 21(1), 37–46.

Barzilay, R. and Lapata, M. 2008. Modeling local coherence: An entity-based approach. *Computational Linguistics* 34(1), 1–34.

Bennett, D., Schneider, J., Arvanitakis, Z., Kelly, J., Aggarwal, N., Shah, R. and Wilson, R. 2006. Neuropathology of older persons without cognitive impairment from two community-based studies. *Neurology* 66, 1837–844.

Bentivogli, L., Dagan, I., Dang, H. T., Giampiccolo, D. and Magnini, B. 2009. The Fifth PASCAL Recognizing Textual Entailment Challenge. *Proceedings of TAC* 9, 14–24.

Berg-Kirkpatrick, T., Bouchard-Côté, A., DeNero, J. and Klein, D. 2010. Painless unsupervised learning with features. In *Proceedings of the Human Language Technology Conference of the NAACL*, pages 582–590.

Bishop, D. 2004. *Expression, Reception and Recall of Narrative Instrument*. London: The Psychological Corporation.

Bishop, D. 2006. What Causes Specific Language Impairment in Children? *Current Directions In Psychological Science* 15(5), 217–221.

Bishop, D. and Donlan, C. 2005. The role of syntax in encoding and recall of pictorial narratives: Evidence from specific language impairment. *British Journal of Developmental Psychology* 23, 25–46.

Bishop, D. and Edmunsson, A. 1987. Language impaired 4-year-olds: Distinguishing transient from persistent impairment. *Journal of Speech and Hearing Disorder* 52, 156–173.

Bishop, D., Whitehouse, A., Watt, H. and Line, E. 2008. Autism and diagnostic substitution: Evidence from a study of adults with a history of developmental language disorder. *Developmental Medicine and Child Neurology* 50, 341–345.

Blackman, J. 2002. Early intervention: A global perspective. *Infants & Young Children* 15(2), 11–19.

Blunsom, P. and Cohn, T. 2006. Discriminative Word Alignment with Conditional Random Fields. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 65–72.

Botting, N. 2002. Narrative as a tool for the assessment of linguistic and pragmatic impairments. *Child Language Teaching and Therapy* 18(1), 1–21.

Brew, C. and Schulte im Walde, S. 2002. Spectral clustering for german verbs. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing*, pages 117–124.

Brin, S. and Page, L. 1998. The anatomy of a large-scale hypertextual Web search engine. *Computer networks and ISDN systems* 30(1-7), 107–117.

Brinkman, S. D., Largen, J., Gerganoff, S. and Pomara, N. 1983. Russell's Revised Wechsler Memory Scale in the evaluation of dementia. *Journal of Clinical Psychology* 39(6), 989–993.

Brookmeyer, R., Johnson, E., Ziegler-Graham, K. and Arrighi, H. 2007. Forecasting the global burden of Alzheimer's disease. *Alzheimer's and Dementia* 3(3), 186–191.

Brown, P., Della Pietra, V., Della Pietra, S. and Mercer, R. 1993. The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics* 19(2), 263–311.

Bschor, T., Kuhl, K.-P. and Reischies, F. 2001. Spontaneous Speech of Patients With Dementia of the Alzheimer Type and Mild Cognitive Impairment. *International Psychogeriatrics* 13(3), 289–298.

Burke, W. J., Miller, J. P., Rubin, E. H., Morris, J. C., Coben, L. A., Duchek, J., Wittels, I. G. and Berg, L. 1988. Reliability of the Washington University Clinical Dementia Rating. *Archives of Neurology* 45, 31–32.

Burstein, K., Claudia, M. C. and Leacock. 2003. CriterionSM: Online essay evaluation: An application for automated evaluation of student essays. In *Proceedings of the Fifteenth Annual Conference on Innovative Applications of Artificial Intelligence*, pages 3–10.

Busse, A., Hensel, A., Ghne, U., Angermeyer, M. and Riedel-Heller, S. 2006. Mild cognitive impairment: long-term course of four clinical subtypes. *Neurology* 67, 2176–2185.

Callison-Burch, C., Koehn, P. and Osborne, M. 2006. Improved Statistical Machine Translation Using Paraphrases. In *Proceedings of the Human Language Technology Conference of the NAACL*, pages 17–24.

Capps, L., Losh, M. and Thurber, C. 2000. "The frog ate the bug and made his mouth sad": Narrative competence in children with autism. *Journal of Abnormal Child Psychology* 28, 193–204.

Chambers, N. and Jurafsky, D. 2008. Unsupervised learning of narrative event chains. *Proceedings of ACL-08: HLT* pages 789–797.

Chambers, N. and Jurafsky, D. 2010. A database of narrative schemas. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation*, pages 1614–1618.

Chang, C.-C. and Lin, C.-J. 2011. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology* 2(27), 1–27.

Chapman, R. M., Mapstone, M., Gardner, M. N., Sandoval, T. C., McCrary, J. W., Guillily, M. D., Reilly, L. A. and DeGrush, E. 2011. Women have farther to fall: Gender differences between normal elderly and Alzheimer's disease in verbal memory engender better detection of AD in women. *Journal of the International Neuropsychological Society* 17, 654–662.

Chapman, S., Ulatowska, H., King, K., Johnson, J. and McIntire, D. 1995. Discourse in early Alzheimer's disease versus normal advanced aging. *American Journal of Speech-Language Pathology* pages 125–129.

Chapman, S., Zientz, J., Weiner, M., Rosenberg, R., Frawley, W. and Burns, M. 2002. Discourse changes in early Alzheimer disease, mild cognitive impairment, and normal aging. *Alzheimer Disease and Associated Disorders* 16, 177–186.

Chenery, H. J. and Murdoch, B. E. 1994. The production of narrative discourse in response to animations in persons with dementia of the Alzheimer's type: Preliminary findings. *Aphasiology* 8(2), 159–171.

Cohen, M. 1997. *Children's memory scale*. San Antonio, TX: The Psychological Corporation.

Cortes, C., Mohri, M. and Rastogi, A. 2007. An alternative ranking problem for search engines. In *Proceedings of the 6th Workshop on Experimental Algorithms*, volume 4525 of *Lecture Notes in Computer Science*, pages 1–21.

Creamer, S. and Schmitter-Edgecombe, M. 2010. Narrative comprehension in Alzheimer's Disease: Assessing inferences and memory operations with a think-aloud procedure. *Neuropsychology* 24(3), 279–290.

Crosson, B., Hughes, C. W., Roth, D. L. and Paul G, M. 1984. Review of Russell's (1975) Norms for the Logical Memory and Visual Reproduction Subtests of the Wechsler Memory Scale. *Journal of Consulting and Clinical Psychology* 52(4), 635–641.

Daumé, H. and Marcu, D. 2005. Induction of word and phrase alignments for automatic document summarization. *Computational Linguistics* 31(4), 505–530.

de Lira, J., Ortiz, K., Campanha, A., Bertolucci, P. and Minetta, T. 2011. Microlinguistic aspects of the oral narrative in patients with Alzheimer's disease. *International Psychogeriatrics* 23(3), 404–412.

DeNero, J. and Klein, D. 2007. Tailoring word alignments to syntactic machine translation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 17–24.

Denkowski, M. and Lavie, A. 2011. Meteor 1.3: Automatic Metric for Reliable Optimization and Evaluation of Machine Translation Systems. In *Proceedings of the EMNLP 2011 Workshop on Statistical Machine Translation*, pages 85–91.

Diehl, J. J., Bennetto, L. and Young, E. C. 2006. Story Recall and Narrative Coherence of High-Functioning Children with Autism Spectrum Disorders. *Journal of Abnormal Child Psychology* 34(1), 87–102.

Dodwell, K. and Bavin, E. L. 2008. Children with specific language impairment: an investigation of their narratives and memory. *International Journal of Language and Communication Disorders* 43(2), 201–218.

Dunn, J. C., Almeida, O. P., Barclay, L., Waterreus, A. and Flicker, L. 2002. Latent Semantic Analysis: A new method to measure prose recall. *Journal of Clinical and Experimental Neuropsychology* 24(1), 26–35.

Egan, J. 1975. *Signal Detection Theory and ROC Analysis*. New York: Academic Press.

Ehrlich, J. S., Obler, L. K. and Clark, L. 1997. Ideational and semantic contributions to narrative production in adults with dementia of the Alzheimer's type. *Journal of Communication Disorders* 30, 79–99.

Erkan, G. and Radev, D. R. 2004. LexRank: Graph-based Lexical Centrality as Salience in Text Summarization. *Journal of Artificial Intelligence Research* 22, 457–479.

Ferri, C., Prince, M., Brayne, C., Brodaty, H., Fratiglioni, L., Ganguli, M., Hall, K., Hasegawa, K., Hendrie, H., Huang, Y., Jorm, A., Mathers, C., Menezes, P., Rimmer, E. and Scazufca, M. 2006. Global prevalence of dementia: A Delphi consensus study. *The Lancet* 366(9503), 2112–2117.

Fivush, R., Gray, J. and Fromhoff, F. 1987. Two-year-olds talk about the past. *Cognitive Development* 2, 393–409.

Folstein, M., Folstein, S. and McHugh, P. 1975. Mini-Mental State - A practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research* 12, 189–198.

Foltz, P., Kintsch, W. and Landauer, T. 1998. The measurement of textual coherence with latent semantic analysis. *Discourse processes* 25(2-3), 285–307.

Fraser, A. and Marcu, D. 2005. ISI's participation in the Romanian-English alignment task. In *Proceedings of the ACL Workshop on Building and Using Parallel Texts*, pages 91–94.

Fraser, A. and Marcu, D. 2007. Measuring word alignment quality for statistical machine translation. *Computational Linguistics* 33(3), 293–303.

Gabani, K., Sherman, M., Solorio, T. and Liu, Y. 2009. A corpus-based approach for the prediction of language impairment in monolingual English and Spanish-English bilingual children. In *Proceedings of the Human Language Technology Conference of the NAACL*, pages 46–55.

Gale, D. and Shapley, L. 1962. College admissions and the stability of marriage. *The American Mathematical Monthly* 69(1), 9–15.

Galvin, J., Fagan, A., Holtzman, D., Mintun, M. and Morris, J. 2010. Relationship of dementia screening tests with biomarkers of Alzheimer's disease. *Brain* 133, 3290–3300.

Gilesa, E., Patterson, K. and Hodge, J. R. 1996. Performance on the Boston cookie theft picture description task in patients with early dementia of the Alzheimer's type: Missing information. *Aphasiology* 10(4), 395–408.

Gillam, R. B. and Pearson, N. A. 2004. *Test of Narrative Language*. Austin, TX: Pro-ed.

Glasgow, C. and Cowley, J. 1994. *Renfrew Bus Story Test - North American Edition*. Centreville, DE: Centreville School.

Glickman, O., Dagan, I. and Koppel, M. 2006. A lexical alignment model for probabilistic textual entailment. In *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Textual Entailment*, volume 3944 of *Lecture Notes in Computer Science*, pages 287–298.

Golden, C. and Freshwater, S. M. 2001. Luria-Nebraska Neuropsychological Battery. In William I. Dorfman and Michael Hersen (eds.), *Understanding Psychological Assessment: Perspectives on Individual Differences*, New York: Kluwer Academic/Plenum Publishers.

Goldman, S. 2008. Brief Report: Narratives of personal events in children with autism and developmental language disorders: Unshared memories. *Journal of Autism and Developmental Disorders* 38, 1982–1988.

Goodglass, H. and Kaplan, E. 1972. *Boston Diagnostic Aphasia Examination*. Philadelphia: Lea and Febiger.

Goodglass, H., Kaplan, E. and Barresi, B. 2001. *Boston Diagnostic Aphasia Examination. 3rd ed.*. Austin, TX: Pro-Ed.

Haaland, K. Y., Linn, R., Hunt, W. and Goodwin, J. 1983. A normative study of Russell's variant of the Wechsler Memory Scale in a healthy elderly population. *Journal of Consulting and Clinical Psychology* 51, 878–881.

Hakkani-Tur, D., Vergyri, D. and Tur, G. 2010. Speech-based automated cognitive status assessment. In *Proceedings of Interspeech*, pages 258–261.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. and Witten, I. H. 2009. The WEKA Data Mining Software: An Update. *SIGKDD Explorations* 11(1), 10–18.

Hanley, J. and McNeil, B. 1982. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 143, 29–36.

Heilmann, J., Miller, J. F., Nockerts, A. and Dunaway, C. 2010. Narrative scoring scheme properties of the narrative scoring scheme using narrative retells in young school-age children. *American Journal of Speech-Language Pathology* 19, 154–166.

Hickl, A. and Bensley, J. 2007. A discourse commitment-based framework for recognizing textual entailment. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, pages 171–176.

Hickl, A., Bensley, J., Williams, J., Roberts, K., Rink, B. and Shi, Y. 2006. Recognizing textual entailment with LCC's GROUNDHOG system. In *Proceedings of the Second PASCAL Challenges Workshop*.

Hier, D., Hagenlocker, K. and Shindler, A. 1985. Language disintegration in dementia: Effects of etiology and severity. *Brain and Language* 25, 117–133.

Higgins, D., Burstein, J., Marcu, D. and Gentile, C. 2004. Evaluating multiple aspects of coherence in student essays. In *Proceedings of the Human Language Technology Conference of the NAACL*, pages 185–192.

Hoops, S., Nazem, S., Siderowf, A. D., Duda, J. E., Xie, S. X., Stern, M. B. and Weintraub, D. 2009. Validity of the MoCA and MMSE in the detection of MCI and dementia in Parkinson disease. *Neurology* 73(21), 1738–1745.

Jing, H. and McKeown, K. 2000. Cut and paste based text summarization. In *Proceedings of the Human Language Technology Conference of the NAACL*, pages 178–185.

Johnson, D. K., Storandt, M. and Balota, D. A. 2003. Discourse analysis of logical memory recall in normal aging and in dementia of the Alzheimer type. *Neuropsychology* 17(1), 82–92.

Johnson, R. E. 1970. Recall of prose as a function of the structural importance of the linguistic units. *Journal of Verbal Learning and Verbal Behavior* 9, 2–20.

Justice, L. M., Bowles, R., Pence, K. and Gosse, C. 2010. A scalable tool for assessing childrens language abilities within a narrative context: The NAP (Narrative Assessment Protocol). *Early Childhood Research Quarterly* 25, 218–234.

Justice, L. M., Bowles, R. P., Kaderavek, J. N., Ukrainetz, T. A., Eisenberg, S. L. and Gillam, R. B. 2006. The Index of Narrative Microstructure: A clinical tool for analyzing school-age children's narrative performances. *American Journal of Speech-Language Pathology* 15, 177–191.

Kanner, L. 1943. Autistic disturbances of affective content. *Nervous Child* 2, 217–250.

Kasari, C. 2002. Assessing change in early intervention programs for children with autism. *Journal of autism and developmental Disorders* 32(5), 447–461.

Kim, Y., Leventhal, B., Koh, Y., Fombonne, E., Laska, E., Lim, E., Cheon, K., Kim, S., Kim, Y., Lee, H., Song, D.-H. and Grinker, R. R. 2011. Prevalence of autism spectrum disorders in a total population sample. *American Journal of Psychiatry* 168(9), 904–912.

King, M. and Bearman, P. 2009. Diagnostic change and the increased prevalence of autism. *International Journal of Epidemiology* 38(5), 1224–1234.

Kjelgaard, M. and Tager-Flusberg, H. 2001. An Investigation of Language Impairment in Autism: Implications for Genetic Subgroups. *Language and Cognitive Processes* 16(2/3), 287–308.

Koehn, P. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the Tenth Machine Translation Summit (MT Summit X)*, volume 5.

Koehn, P., Och, F. and Marcu, D. 2003. Statistical phrase-based translation. In *Proceedings of the Human Language Technology Conference of the NAACL*, pages 48–54.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A. and Herbsts, E. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL Interactive Poster and Demonstration Sessions*, pages 177–180.

Kok, S. and Brockett, C. 2010. Hitting the Right Paraphrases in Good Time. In *Proceedings of the Human Language Technology Conference of the NAACL*, pages 145–153.

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

Labov, W. and Waletzky, J. 1967. Narrative analysis. In J. Helm (ed.), *Essays on the Verbal and Visual Arts*, pages 12–44, Seattle: University of Washington Press.

Landauer, T. K., Foltz, P. W. and Laham, D. 1998. An Introduction to Latent Semantic Analysis. *Discourse Processes* 25, 259–284.

Landauer, T. K., Laham, D. and Foltz, P. 2003. Automatic essay assessment. *Assessment in education: Principles, policy & practice* 10(3), 295–308.

Lapata, M. and Barzilay, R. 2005. Automatic evaluation of text coherence: Models and representations. In *International Joint Conference On Artificial Intelligence*, volume 19, pages 1085–1090.

Lardilleux, A., Gosme, J., Lepage, Y. et al. 2010. Bilingual lexicon induction: Effortless evaluation of word alignment tools and production of resources for improbable language pairs. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation*, pages 252–256.

Lautenschlager, N. T., Dunn, J. C., Bonney, K., Flicker, L. and Almeida, O. P. 2006. Latent Semantic Analysis: An improved method to measure cognitive performance in subjects of non-English-speaking-background. *Journal of Clinical and Experimental Neuropsychology* 28, 1381–1387.

Legetter, C. J. and Woodland, P. C. 1995. Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density HMM. *Computer Speech and Language* 9, 171–185.

Li, Z., Callison-Burch, C., Dyer, C., Ganitkevitch, J., Irvine, A., Khudanpur, S., Schwartz, L., Thornton, W. N. G., Wang, Z., Weese, J. and Zaidan, O. F. 2010. Joshua 2.0: A Toolkit for Parsing-Based Machine Translation with Syntax, Semirings, Discriminative Training and Other Goodies. In *Proceedings of the Joint 5th Workshop on Statistical Machine Translation and MetricsMATR*, pages 133–137.

Liang, P., Taskar, B. and Klein, D. 2006. Alignment by Agreement. In *Proceedings of the Human Language Technology Conference of the NAACL*, pages 104–111.

Liles, B. Z., Duffy, R. J., Merritt, D. D. and Purcell, S. L. 1995. Measurement of Narrative Discourse Ability in Children With Language Disorders. *Journal of Speech and Hearing Research* 38, 415–425.

Lin, C.-Y. 2004. Rouge: A package for automatic evaluation of summaries. In *Proceedings of the Workshop on Text Summarization*.

Liu, Y., Liu, Q. and Lin, S. 2005. Log-linear models for word alignment. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, pages 459–466.

Logue, P. and Wyrick, L. 1979. Initial validation of Russell's Revised Wechsler Memory Scale: A comparison of normal aging versus dementia. *Journal of Consulting and Clinical Psychology* 47, 176–178.

Lopez, A. 2008. Statistical Machine Translation. *ACM Computing Surveys* 40(3), 1–49.

Lord, C. and Paul, R. 1997. Language and communication in autism. In D. Cohen and F. Volkmar (eds.), *Handbook of Autism and Pervasive Developmental Disorders*, pages 195–225, New York: Wiley.

Lord, C., Rutter, M., DiLavore, P. and Risi, S. 2002. *Autism Diagnostic Observation Schedule (ADOS)*. Los Angeles: Western Psychological Services.

Lord, C., Rutter, M. and LeCouteur, A. 1994. Autism Diagnostic Interview-Revised: A revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders. *Journal of Autism and Developmental Disorders* 24, 659–685.

Losh, M. and Capps, L. 2003. Narrative ability in high-functioning children with autism or Asperger's Syndrome. *Journal of Autism and Developmental Disorders* 33(3).

Lovász, L. 1993. Random Walks on Graphs: A Survey. *Combinatorics, Paul Erdos is Eighty* 2(1), 1–46.

Loveland, K., McEvoy, R. and Tunali, B. 1990. Narrative story telling in autism and Down's syndrome. *British Journal of Developmental Psychology* 8(1), 9–23.

Loveland, K. and Tunali, B. 1993. Narrative language in autism and the theory of mind hypothesis: a wider perspective. In S. Baron-Cohen, H. Tager-Flusberg and D. J. Cohen (eds.), *Understanding Other Minds: Perspectives from Autism*, Oxford: Oxford University Press.

Lyons, K., Kemper, S., LaBarge, E., Ferraro, F., Balota, D. and Storandt, M. 1994. Oral language and Alzheimer's Disease: A reduction in syntactic complexity. *Aging and Cognition* 1(4), 271–281.

MacCartney, B., Galley, M. and Manning, C. D. 2008. A phrase-based alignment model for natural language inference. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 802–811.

MacWhinney, B. 2000. *The CHILDES Project: Tools for Analyzing Talk*. Mahwah, NJ: Lawrence Erlbaum Associates.

MacWhinney, B. 2007. The TalkBank Project. In J. C. Beal, K. P. Corrigan and H. L. Moisl (eds.), *Creating and Digitizing Language Corpora: Synchronic Databases, Vol.1*, Houndmills: Palgrave-Macmillan.

Mani, I. and Pustejovsky, J. 2004. Temporal discourse models for narrative structure. In *Proceedings of the 2004 ACL Workshop on Discourse Annotation*, pages 57–64.

Manly, J. J., Tang, M., Schupf, N., Stern, Y., Vonsattel, J.-P. G. and Mayeux, R. 2008. Frequency and course of mild cognitive impairment in a multiethnic community. *Annals of Neurology* 63(4), 494–506.

Marton, Y., Callison-Burch, C. and Resnik, P. 2009. Improved Statistical Machine Translation Using Monolingually-Derived Paraphrases. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 381–390.

McCarty, S. M., Ziesat, H. A., Logue, P. E., Power, D. G. and Rosenstiel, A. K. 1980. Alternate-Form Reliability and Age-Related Scores for Russell's Revised Wechsler Memory Scale. *Consulting and Clinical Psychology* 48(2), 296–298.

McCauley, R. 2001. *Assessment of language disorders in children*. Mahwah, NJ: Lawrence Erlbaum Associates.

McCulla, M., Coats, M., Fleet, N. V., Duchek, J., Grant, E. and Morris, J. 1989. Reliability of clinical nurse specialists in the staging of dementia. *Archives of Neurology* 46, 1210–1211.

Mehdad, Y., Negri, M. and Federico, M. 2011. Using bilingual parallel corpora for cross-lingual textual entailment. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, pages 1336–1345.

Melamed, D. 2000. Models of translational equivalence among words. *Computational Linguistics* 26(2), 221–249.

Meurers, D., Ziai, R., Ott, N. and Kopp, J. 2011. Evaluating answers to reading comprehension questions in context: results for German and the role of information structure. In *Proceedings of the TextInfer 2011 Workshop on Textual Entailment*, pages 1–9.

Mihalcea, R. 2005. Unsupervised large-vocabulary word sense disambiguation with graph-based algorithms for sequence data labeling. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 411–418.

Miller, J. F., Andriacchi, K. and Nockerts, A. 2011. *Assessing language production using SALT software: A Clinician's Guide to Language Sample Analysis*.

Miniscalco, C., Hagberg, B., Kadesjo, B., Westerlund, M. and Gillberg, C. 2007. Narrative skills, cognitive profiles and neuropsychiatric disorders in 7–8-year-old children with late developing language. *International Journal of Language and Communication Disorders* 42(6), 665–681.

Minkov, E. and Cohen, W. 2008. Learning graph walk based similarity measures for parsed text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 907–916.

Mitchell, M. 1987. Scoring discrepancies on two subtests of the Wechsler Memory Scale. *Journal of Consulting and Clinical Psychology* 55, 914–915.

Moore, R. 2005. A discriminative framework for bilingual word alignment. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 81–88.

Moore, R., Yih, W. and Bode, A. 2006. Improved discriminative bilingual word alignment. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 513–520.

Morris, J., Ernesto, C., Schafer, K., Coats, M., Leon, S., Sano, M., Thal, L. and Woodbury, P. 1997. Clinical dementia rating training and reliability in multicenter studies: The Alzheimer's disease cooperative study experience. *Neurology* 48(6), 1508–1510.

Morris, J., Heyman, A., Mohs, R. and Hughes, J. 1989. The consortium to establish a registry for Alzheimer's disease (CERAD): I. Clinical and neuropsychological assessment of Alzheimer's disease. *Neurology* 39(9), 1159–1165.

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

Morris, J., Storandt, M., Miller, J. P., McKeel, D., Price, J., Rubin, E. and Berg, L. 2001. Mild Cognitive Impairment Represents Early-Stage Alzheimer Disease. *Archives of Neurology* 58, 397–405.

Munteanu, D. and Marcu, D. 2006. Extracting parallel sub-sentential fragments from non-parallel corpora. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 81–88.

Myers, S., Johnson, C. et al. 2007. Management of children with autism spectrum disorders. *Pediatrics* 120(5), 1162–1182.

Navigli, R. and Lapata, M. 2010. An experimental study of graph connectivity for unsupervised word sense disambiguation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 32(4), 678–692.

Niehues, J. and Vogel, S. 2008. Discriminative word alignment via alignment matrix modeling. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 18–25.

Nielsen, R. D., Ward, W. and Martin, J. H. 2009. Recognizing entailment in intelligent tutoring systems. *Natural Language Engineering* 15(4), 479–501.

Norbury, C. and Bishop, D. 2003. Narrative skills of children with communication impairments. *International Journal of Language and Communication Disorders* 38, 287–313.

Nordlund, A., Rolstad, S., Hellstrom, P., Sjogren, M., Hansen, S. and Wallin, A. 2005. The Goteborg MCI study: mild cognitive impairment is a heterogeneous condition. *Journal of Neurology, Neurosurgery and Psychiatry* 76, 1485–1490.

Och, F. J. and Ney, H. 2000. A comparison of alignment models for statistical machine translation. In *Proceedings of the 18th Conference on Computational Linguistics*, pages 1086–1090.

Och, F. J. and Ney, H. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics* 29(1), 19–51.

Otterbacher, J., Erkan, G. and Radev, D. R. 2009. Biased LexRank: Passage retrieval using random walks with question-based priors. *Inf. Process. Manage.* 45(1), 42–54.

Page, E. B. 1966. The Imminence of... Grading Essays by Computer. *The Phi Delta Kappan* 47(5), 238–243.

Page, E. B. and Petersen, N. 1995. The computer moves into essay grading: Updating the ancient test. *Phi Delta Kappan* 76, 561–561.

Page, L., Brin, S., Motwani, R. and Winograd, T. 1999. The PageRank Citation Ranking: Bringing Order to the Web. Technical Report 1999-66, Stanford InfoLab.

Pahikkala, T., Airola, A., Boberg, J. and Salakoski, T. 2008. Exact and efficient leave-pair-out cross-validation for ranking RLS. In *Proceedings of AKRR 2008*, pages 1–8.

Papineni, K., Roukos, S., Ward, T. and jing Zhu, W. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318.

Paslawski, T. 2005. The Clinical Evaluation of Language Fundamentals, Fourth Edition (CELF-4): A Review. *Canadian Journal of School Psychology* 20(1-2), 129–134.

Pérez, D., Alfonseca, E. and Pilar Rodríguez, P. 2004. Application of the BLEU method for evaluating free-text answers in an e-learning environment. In *Proceedings of the language resources and evaluation conference (LREC-2004)*, pages 26–28.

Pérez, D., Gliozzo, A., Strapparava, C., Alfonseca, E., Rodríguez, P. and Magnini, B. 2005. Automatic assessment of students free-text answers underpinned by the combination of a bleu-inspired algorithm and latent semantic analysis. In *Proceedings of the Eighteenth International Florida Artificial Intelligence Research Society Conference*, pages 358–362.

Petersen, R., Smith, G., Waring, S., Ivnik, R., Tangalos, E. and Kokmen, E. 1999. Mild Cognitive Impairment: Clinical characterizations and outcomes. *Archives of Neurology* 56, 303–308.

Petersen, R. C. 2011. Mild Cognitive Impairment. *The New England Journal of Medicine* 364(23), 2227–2234.

Peterson, C. 1990. The who, when and where of early narratives. *Journal of Child Language* 17, 433–455.

Plassman, B. L., Langa, K. M., Fisher, G. G., Heeringa, S. G., Weir, D. R., Ofstedal, B., Burke, J. R., Hurd, M. D., Potter, G. G., Rodgers, W. L., Steffens, D. C., McArdle, J. J., Willis, R. J., and Wallace, R. B. 2008. Prevalence of cognitive impairment without dementia in the United States. *Annnals of Internal Medicine* 148, 427–34.

Porter, M. 1980. An algorithm for suffix stripping. *Program* 14, 130–137.

Power, D., Logue, P., McCarty, S., Rosenstiel, A. and Ziesat, H. 1979. Interrater reliability of the Russell version of the Wechsler Memory Scale: An attempt to clarify ambiguities in scoring. *Journal of Clinical Neuropsychology* 1, 343–345.

Price, J. L., McKeel, D. W., Buckles, V. D., Roe, C. M., Xiong, C., Grundman, M., Hansen, L. A., Petersen, R. C., Parisi, J. E., Dickson, D. W., Smith, C. D., Davis, D. G., Schmitt, F. A., Markesbery, W. R., Kaye, J., Kurlan, R., Hulette, C., Kurland, B. F., Higdon, R., Kukull, W. and Morris, J. C. 2009. Neuropathology of Nondemented Aging: Presumptive Evidence for Preclinical Alzheimer Disease. *Neurobiology of Aging* 30(7), 1026–1036.

Prigatano, G. P. . 1978. Wechsler Memory Scale: A selective review of the literature. *Journal of Clinical Psychology* 34, 816–832.

Propp, V. 1968. *Morphology of the folktale*. Austin: University of Texas Press.

Pustejovsky, J., Knippen, R., Littman, J. and Saurí, R. 2005. Temporal and event information in natural language text. *Language Resources and Evaluation* 39(2), 123–164.

Quirk, C., Brockett, C. and Dolan, W. 2004. Monolingual Machine Translation for Paraphrase Generation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 142–149.

Quirk, C., Udupa, R. and Menezes, A. 2007. Generative models of noisy translations with applications to parallel fragment extraction. In *Proceedings of the Machine Translation Summit XI*, pages 377–384.

Rapin, I. and Dunn, M. 2003. Update on the language disorders of individuals on the autistic spectrum. *Brain and Development* 25, 166–172.

Ravaglia, G., Forti, P., Maioli, F., Servadei, L., Martelli, M., Brunetti, N., Bastagli, L., and Mariani, E. 2005. Screening for mild cognitive impairment in elderly ambulatory patients with cognitive complaints. *Aging Clinical and Experimental Research* 17(5), 374–379.

Reynolds, C. and Bigler, E. 1994. *Test of Memory and Learning (TOMAL)*. Austin, Texas: Pro-Ed.

Riesa, J. and Marcu, D. 2010. Hierarchical search for word alignment. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 157–166.

Ritchie, K., Artero, S. and Touchon, J. 2001. Classification criteria for mild cognitive impairment: A population-based validation study. *Neurology* 56, 37–42.

Ritchie, K. and Touchon, J. 2000. Mild Cognitive Impairment: Conceptual basis and current nosological status. *Lancet* 355, 225–228.

Roark, B., Mitchell, M., Hosom, J.-P., Hollingshead, K. and Kaye, J. 2011. Spoken language derived measures for detecting Mild Cognitive Impairment. *IEEE Transactions on Audio, Speech and Language Processing* 19(7), 2081–2090.

Rosen, W., Mohs, R. and Davis, K. 1984. A new rating scale for Alzheimer's disease. *The American Journal of Psychiatry* 141(11), 1356–1364.

Russell, E. 1975. A multiple scoring method for the assessment of complex memory functions. *Journal of Consulting and Clinical Psychology* 43, 800–809.

Rutter, M., Bailey, A. and Lord, C. 2003. *Social Communication Questionnaire (SCQ)*. Los Angeles: Western Psychological Services.

Sagae, K., Lavie, A. and MacWhinney, B. 2005. Automatic Measurement of Syntactic Development in Child Language. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, pages 197–204.

Scarborough, H. 1990. Index of productive syntax. *Applied Psycholinguistics* 11(1), 1–22.

Schank, R. and Abelson, R. 1995. Knowledge and memory: The real story. In R. S. Wyer (ed.), *Advances in Social Cognition. Knowledge and Memory vol. VIII*, pages 1–85, Hillsdale, NJ: LEA.

Schmitt, A. J. and Wodrich, D. L. 2004. Validation of a Developmental Neuropsychological Assessment (NEPSY) through comparison of neurological, scholastic concerns, and control groups. *Archives of Clinical Neuropsychology* 19, 1077–1093.

Schmitt, F., Davis, D., Wekstein, D., Smith, C., Ashford, J. and Markesbery, W. 2000. Preclinical AD revisited: Neuropathology of cognitively normal older adults. *Neurology* 55, 370–376.

Schwartz, M. and Ivnik, R. 1980. Wechsler Memory Scale I: Towards a more objective and systematic scoring system of Logical Memory and Visual Reproduction subtests. In *Proceedings of the American Psychological Association*.

Shankle, W. R., Romney, A. K., Hara, J., Fortier, D., Dick, M. B., Chen, J. M., Chan, T. and Sun, X. 2005. Methods to improve the detection of mild cognitive impairment. *Proceedings of the National Academy of Sciences* 102(13), 4919–4924.

Shermis, M., Burstein, J., Higgins, D. and Zechner, K. 2010. Automated essay scoring: Writing assessment and instruction. *International Encyclopedia of Education* pages 20–26.

Shum, D., Murray, R. A. and Eadie, K. 1997. Effect of Speed of Presentation on Administration of the Logical Memory Subtest of the Wechsler Memory Scale-Revised. *Clinical Neuropsychologist* 11(2), 188–191.

Sinha, R. and Mihalcea, R. 2007. Unsupervised graph-basedword sense disambiguation using measures of word semantic similarity. In *International Conference on Semantic Computing*, pages 363–369.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L. and Makhoul, J. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of Association for Machine Translation in the Americas*, pages 223–231.

Snover, M., Madnani, N., Dorr, B. and Schwartz, R. 2009. TER-Plus: Paraphrase, semantic, and alignment enhancements to Translation Edit Rate. *Machine Translation* 23(2), 117–127.

Solorio, T. and Liu, Y. 2008. Using language models to identify language impairment in Spanish-English bilingual children. In *Proceedings of the ACL 2008 Workshop on Biomedical Natural Language Processing (BioNLP)*, pages 116–117.

Stein, N. 1988. The development of children's storytelling skill. In M. B. Franklin and S. Barten (eds.), *Child language: A Reader*, pages 282–279, New York: Oxford University.

Stein, N. and Nezworski, T. 1978. The effects of organization and instructional set on story memory. *Discourse Processes* 1, 177–193.

Storandt, M., Botwinick, J., Danziger, W. L., Berg, L. and Hughes, C. P. 1984. Psychometric differentiation of mild senile dementia of the Alzheimer type. *Archives of Neurology* 41(5), 497–499.

Storandt, M. and Hill, R. 1989. Very mild senile dementia of the Alzheimer's type: II Psychometric test performance. *Archives of Neurology* 46, 383–386.

Stothard, S., Snowling, M., Bishop, D., Chipchase, B. and Kaplan, C. 1998. Language-impaired preschoolers: A follow-up into adolescence. *Journal of Speech, Language and Hearing Research* 41, 407–418.

Strassel, S. 2004. *Simple metadata annotation specification v6.2*. Linguistic Data Consortium.

Sullivan, K. 1996. Estimates of interrater reliability for the Logical Memory subtest of the Wechsler Memory Scale - Revised. *Journal of Clinical and Experimental Neuropsychology* 18, 707–712.

Tager-Flusberg, H. 1995. Once upon a ribbit: Stories narrated by autistic children. *British journal of developmental psychology* 13(1), 45–59.

Tager-Flusberg, H. 2001. Understanding the language and communicative impairments in autism. *International Review of Research in Mental Retardation* 23, 185–205.

Taskar, B., Lacoste-Julien, S. and Klein, D. 2005. A discriminative matching approach to word alignment. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 73–80.

Tiedemann, J. 2003. Combining clues for word alignment. In *Proceedings of the Tenth Conference of the European Chapter of the Association for Computational Linguistics*, pages 339–346.

Tierney, M., Yao, C., Kiss, A. and McDowell, I. 2005. Neurospsychological tests accurately predict incident Alzheimer disease after 5 and 10 years. *Neurology* 64, 1853–1859.

Titley, J., D'Amato, R., D'Amato, R. and Hartlage, L. 2008. Understanding and using the NEPSY-II with young children, children, and adolescents. *Essentials of neuropsychological assessment: Treatment planning for rehabilitation* pages 149–172.

Titley, J. E. and D'Amato, R. C. 2008. Understanding and using the NEPSY-II with young children, children, and adolescents. In Rik Carl D'Amato and Lawrence Hartlage (eds.), *Essentials of Neuropsychological Assessment: Treatment Planning for Rehabilitation*, New York: Springer.

Tomblin, B. 2011. Co-morbidity of autism and SLI: Kinds, kin and complexity. *International Journal of Communication and Language Disorders* 46(2), 127–137.

Tomblin, J. B. 1996. Genetic and environmental contributions to the risk for specific language impairment. In Mabel Rice (ed.), *Toward a Genetics of Language*, pages 191–211, Hillsdale, NJ: Lawrence Erlbaum Associates.

Tomblin, J. B., Records, N. L., Buckwalter, P., Zhang, X., Smith, E. and O'Brien, M. 1997. Prevalence of Specific Language Impairment in Kindergarten Children. *Journal of Speech Language and Hearing Research* 40, 1245–1260.

Toutanova, K., Ilhan, H. and Manning, C. 2002. Extensions to HMM-based statistical word alignment models. In *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing*, volume 10, pages 87–94.

Toutanova, K., Manning, C. and Ng, A. 2004. Learning random walk models for inducing word dependency distributions. In *Proceedings of the 21st International Conference on Machine Learning*, page 103.

Trabasso, T., Secco, T. and van den Broek, P. 1984. Causal cohesion and story coherence. In H. Mandl, N. Stein and T. Trabasso (eds.), *Learning and Comprehension of Text*, pages 83–111, Hillside, NJ: Erlbaum.

Trabasso, T. and Sperry, L. L. 1985. Causal relatedness and importance of story events. *Journal of Memory and Language* 24, 595–611.

Trabasso, T. and van den Broek, P. 1985. Causal thinking and representation of narrative events. *Journal of Memory and Language* 24, 612–630.

Tractenberg, R. E., Schafer, K. and Morris, J. C. 2001. Interobserver Disagreements on Clinical Dementia Rating Assessment: Interpretation and Implications for Training. *Alzheimer Disease and Associated Disorders* 15(3), 155–161.

Turney, P. D. and Pantel, P. 2010. From Frequency to Meaning: Vector Space Models of Semantics. *Journal of Artificial Intelligence Research* 37, 141–188.

Ulatowska, H., Allard, L., Donnell, A., Bristow, J., Haynes, S., Flower, A. and North, A. 1988. Discourse performance in subjects with dementia of the Alzheimer's type. In H.A. Whitaker (ed.), *Neuropsychological Studies of Non-focal Brain Damage*, New York: Springer-Verlag.

United Nations. 2002. *World Population Ageing 1950-2050*. New York: United Nations.

Vogel, S., Ney, H. and Tillmann, C. 1996. HMM-based word alignment in statistical translation. In *Proceedings of the 16th conference on Computational linguistics-Volume 2*, pages 836–841.

Vuorinen, E., Laine, M. and Rinne, J. 2000. Common Pattern of Language Impairment in Vascular Dementia and in Alzheimer Disease. *Alzheimer Disease and Associated Disorders* 14(2), 81–86.

Wang, Q.-S. and Zhou, J.-N. 2002. Retrieval and encoding of episodic memory in normal aging and patients with mild cognitive impairment. *Brain Research* 924, 113–115.

Wang, R. and Callison-Burch, C. 2011. Paraphrase Fragment Extraction from Monolingual Comparable Corpora. In *Proceedings of the 4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web*, pages 52–60.

Wechsler, D. 1945. *Wechsler Memory Scale*. New York: The Psychological Corporation.

Wechsler, D. 1987. *Wechsler Memory Scale - Revised manual*. San Antonio: The Psychological Corporation.

Wechsler, D. 1997. *Wechsler Memory Scale - Third Edition*. San Antonio: The Psychological Corporation.

Wechsler, D. 2002. *Wechsler Primary and Preschool Scale of Intelligence - Third edition (WPPSI-III)*. San Antonio: Harcourt Assessment.

Wechsler, D. 2003. *Wechsler Intelligence Scales for Children - Fourth Edition (WISC-IV)*. San Antonio: The Psychological Corporation.

Wechsler, D., Holdnack, J. and Drozdick, L. 2009. *Wechsler Memory Scale - Fourth Edition Technical and Interpretive Manual*. San Antonio: Pearson.

Woloszyn, D., Murphy, S., Wetzel, L. and Fisher, W. 1993. Interrater agreement on the Wechsler Memory Scale - Revised in a mixed clinical population. *The Clinical Neuropsychologist* 7, 467–471.

Woodcock, R., KS, K. M. and Mather, N. 2001. *Woodcock-Johnson Test of Achievement. III*. Itasca, IL: Riverside Publishing.

Yerkes, R. 1921. *National Academy of Sciences vol. XV: Psychological Examining in the U.S. Army*. Washington, D.C.: Government Printing Office.