



# The Next Wave

The National Security Agency's Review of Emerging Technologies  
Vol 18 No 1 • 2009

## Assisting the Language Analyst:

## Multi-disciplinary Research at CASL

divergent thinking

morphological parser

cognitive task analysis

language aptitude

working memory



### Letter from the Editor

*The Next Wave* (TNW) is known for its coverage of computer science and engineering. In this issue, however, we take a slightly different path to present research ongoing at the University of Maryland Center for Advanced Study of Language (CASL).

In 2003, NSA/CSS was designated Executive Agent for the University Affiliated Research Center (UARC) at the University of Maryland Center for Advanced Study of Language. CASL was created as the first and only Department of Defense (DoD) UARC dedicated to research in human skills—language, cognition, and culture. This unique status came about because DoD and the Intelligence Community (IC) recognized the need for a sustained focus on language readiness.

In 2008, CASL was integrated into the NSA/CSS Research Directorate. This integration is expected to provide CASL with multidisciplinary knowledge, helping CASL address all dimensions of the language analyst's tasks.

CASL's mission is to provide empirical data from solid research to support management decisions involving language analysts. For example, how can we identify people who can learn languages more easily? How can we increase a language analyst's ability to problem solve? How can we use technology to assist language analysts in finding unfamiliar words? These questions, and more, are addressed in CASL's research as presented in this issue of TNW.



The Next Wave is published to disseminate technical advancements and research activities in telecommunications and information technologies. Mentions of company names or commercial products do not imply endorsement by the US Government.

For more information, please contact us at  
TNW@tycho.ncsc.mil





# CONTENTS

## INTRODUCTION

- 4 The Mind of the Language Analyst:  
A Peek Inside the Black Box

## FEATURES

- 6 A Working Memory Workout  
for Language Practitioners
- 14 “Thinking Out of the Box”  
Through Cognitive Neuroscience
- 20 For a Better Dictionary, Build a Better Parser
- 33 A Talent for Language

## FOCUS

- 42 What’s the Bottom Line?
-

# The Mind of the Language Analyst: A Peek Inside the Black Box

As Jane the language analyst (LA) listens to voice material and composes her report, she is faced with a cognitively, linguistically, and technologically challenging job. Although we can hear the voice material and read her report, we can't see what's going on inside Jane's head as she works. The University of Maryland Center for Advanced Study of Language (CASL) is devoted to finding out what's going on inside Jane's head and to giving her the linguistic, cognitive, and technological support she needs to do her job more effectively and efficiently.

Considerable resources have boosted the Intelligence Community's capabilities to capture information, and technologically advanced language tools have been created. However, optimizing the performance of LAs facing challenging tasks also requires considering insights and methods from the behavioral and social sciences, particularly the cognitive, language, and brain sciences. CASL's multidisciplinary research brings together scientists from these diverse fields to help Jane, the human in the equation, work more productively. CASL scientists begin with complicated tasks such as text comprehension or translation and decompose them into more elementary cognitive functions and more detailed information processing mechanisms such as

- working memory,
- divergent thinking and creative problem solving,
- information evaluation and selection,
- analysis of word meaning and its relation to word form, and
- speech/sound discrimination and its contribution to general language aptitude.

The articles in this issue of *The Next Wave* describe CASL research on these five processes in more detail. The articles illustrate how CASL scientists decompose Jane's mental processes to improve language analyst selection, training, and on-the-job performance while highlighting the importance of the partnership between LAs and CASL researchers.



## A working memory workout

Because Jane needs to hold many bits and pieces of information in mind as she listens, she needs a good working memory. Working memory is the cognitive mechanism that allows us to hold disparate information in mind. It is dedicated to the temporary processing, maintenance, and integration of information. Michael Bunting and his collaborators have worked to alleviate some working memory limitations for Jane. Because limitations in working memory are related to language processing difficulty and misinterpretation, training that improves working memory should improve memory capacity. Additionally, such training should also generalize to broader sets of cognitive tasks and processes that rely on working memory—for example, resolving ambiguity, overcoming misleading context, and filling in missing or corrupted information.

## “Thinking out of the box”

Jane is also faced with the problem of linking the information that she hears with other knowledge that she has. In other words, she needs to make connections between what she hears and what she knows. Henk Haarmann and Jared Novick discuss how insights and methods from cognitive science may be used to improve the divergent thinking of LAs and all-source analysts so that they can more easily connect information. Divergent thinking is the ability to think of many useful solutions to open-ended problems—not just common solutions, but also creative ones. For example, LAs are often required to think of possible causes, consequences, and conceptualizations of actors and events they report on. Divergent thinking can be enhanced by changing Jane's neurophysiological and cognitive states. Providing analysts with approaches from



cognitive neuroscience to improve their divergent thinking is important for improving their ability to generate multiple plausible interpretations for texts with missing information, massive linguistic ambiguity, and active distortion or intended deception.

### What's the bottom line?

Jane and her supervisor need to ensure the quality of Jane's report. As part of a quality evaluation, Jane's supervisor needs to assess the difficulties Jane faced with the material and to evaluate it along critical dimensions. Erica Michael and her colleagues have identified key component cognitive processes of what's going on inside Jane's head as she accomplishes her complex language analysis task. Based on this cognitive task analysis, CASL has developed a related quality control (QC) tool for assessing the quality of the summary product. The QC tool's evaluation dimensions (e.g., completeness, significance, accuracy) identify important elements for evaluation, engender thorough quality assessment, and provide a common language for supervisors and LAs to discuss the reports.

### For a better dictionary, build a better parser

When analysts have less-than-perfect language skills, finding words in the dictionary can be difficult due to the complicated structure, or morphology, of many word forms, as well as the complexities of spelling in some languages. As a result, junior analysts working in morphologically complex languages like Pashto and Urdu may be spending too much time looking up unfamiliar words, making it difficult for them to arrive at a thorough understanding of the overall text. Anne David and her associates describe work that makes dictionary search both faster and more accurate by means of a computational tool called a morphological parser. Their innovative approach offers two improvements over traditional parser-building methodologies: (1) it safeguards against software obsolescence thanks to grammatical formalisms that are written at a linguistic, rather than a computational, level; and (2) it provides a template for developing grammars—and hence parsers—that is easily portable to other languages, thus addressing a 30-year-old limitation.

### A talent for language

On the other hand, if Jane is an expert, experienced LA, she will not have many of the problems faced by junior LAs. How can we identify those LAs who have the potential to become expert LAs? Cathy Doughty and her colleagues are developing an aptitude battery to assist with the selection and hiring of future top analysts. The battery aims to identify individuals with high language aptitude so that training and hiring resources can be focused on those who are most likely to succeed. Currently, used language aptitude tests predict success in the early lower-level stages of language learning. A unique feature of the new battery—Hi-LAB—is that it predicts high-level language learning success. The Hi-LAB cognitive behavioral tasks decompose language learning, a multi faceted and complicated cognitive language task, into component processes that support it, such as verbal working memory capacity, processing speed, and sound discrimination, in order to assess a candidate's aptitude. Doughty and her team currently are validating the Hi-LAB battery for operational use.

So what's going on inside Jane's head—the black box—is neither simple nor transparent. We often think of language analysis as simple, because using language in our everyday lives proceeds so easily without conscious thought. However, research shows that the problems faced by Jane on the job are complex and difficult. CASL's research is intended to alleviate some of those problems and make Jane's job a little easier. 📌





# A Working Memory Workout for Language Practitioners

**T**he brain stops developing after childhood. At least that was the commonly held belief—until recently. A person may continue to learn throughout life, but core cognitive abilities—such as the speed with which information can be processed, the ability to focus attention, the capacity for remembering and tracking information when distracted—supposedly remain fixed and even decline with age.

Scientific research is changing that perception, however, suggesting now that certain types of mental workouts, also known as “brain training,” can actually improve core mental abilities and protect against cognitive decline. Just as exercising a muscle increases physical strength, exercising the mind can increase mental fitness in terms of how much information can be processed and maintained at once, and the ability to focus attention on a task. A thriving brain training industry is growing around this research. Brain training software is now available for video game systems and even cell phones [1].

The possibility of improving mental abilities into adulthood and even old age is indeed exciting, as is the prospect of mitigating the effects of clinical disorders such as Alzheimer’s disease and attention deficit hyperactivity disorder—ADHD. But normal populations stand to benefit from cognitive training as well. Researchers at CASL are currently testing how intensive mental workouts benefit healthy participants and who is likely to gain most from such interventions.

Within the Intelligence Community (IC), language analysts have been identified as likely candidates for brain training. Understanding language in one’s first and foreign languages requires the temporary maintenance of linguistic input because words, phrases, and sentences are not spoken (and therefore not heard) all at once. Instead, they unfold sound by sound, word by word. The same is true for written material: we make successive eye fixations across text on a page as we read, rather than viewing an entire sentence all at once. Until additional information reaches the ears or the eyes, a listener, reader, or translator must hold provisional interpretations in mind so that newly encountered input can be updated and integrated into one’s developing representations of sentence meaning. Thus, comprehension proceeds incrementally.

Even under ideal listening or reading conditions, focusing attention is crucial for an accurate interpretation. Just a brief mental lapse can cause important information to be missed. Mental focus is even more important when interpreting material that is either incomplete or ambiguous, as this article explains later. And when details of a story unfold over long periods of time, or when the full account comes from multiple sources bit by bit, mental focus is all the more important.

People differ naturally in their core cognitive abilities, and these differences have an impact on language processing and comprehension tasks. What is not clear, however, is whether performance on language tasks can be improved through a relevant brain training regimen that targets general cognitive abilities and, if so, under what conditions and by how much.

Research at the University of Maryland Center for Advanced Study of Language (CASL) has targeted understanding core cognitive abilities. This research, particularly in memory and attentional control, applies to language practitioners and their efforts to understand a foreign language.



### Which mental resources enable quick and efficient information processing?

In order to quickly and efficiently process information, we rely on the working memory system—that is, the attention and memory systems that enable what we commonly refer to as “multi-tasking.” Working memory can also be thought of as our mental workspace: the small amount of memory that holds and manipulates information for ongoing use. This dynamic working memory system guides behavior and allows us to maintain conscious awareness of goal-relevant information.

At the same time, the working memory system prevents potentially irrelevant or distracting information from gaining access to our consciousness. By deliberately focusing or dividing attention, we can pay attention, make and maintain plans, and engage in goal-directed behavior.

The working memory system is different from the long-term memory system, which is the repository of information that can be stored for days, weeks, and years. Information such as our personal experiences (e.g., what we ate for lunch yesterday), facts (e.g., the date the Declaration of Independence was signed), and how to

perform a certain task (e.g., riding a bicycle) are all saved in long-term memory, not working memory.

#### The limitations of working memory

As we know from the experience of our own memory limitations, working memory is time and capacity limited. For example, many people have had the experience of almost immediately forgetting the name of a new acquaintance. Clearly, there are constraints on how much information can be processed, manipulated, and integrated effectively all at once. Although everyone has had the experience of losing (i.e. forgetting) information they had been holding in working memory (like your new acquaintance’s name), individuals vary in their working memory capacity and in how susceptible they are to short-term forgetting. Furthermore, people are not always aware of how divided attention affects their ability to process information from one task to another. For example, when a conversation is abruptly interrupted by another event (a waiter asking to take your order), many people experience an inability to remember exactly what they had intended to say.

Attentional control is important to the function of the working memory system. Here, we define attention as a cen-

tral, limited-capacity resource that can be voluntarily applied to holding and manipulating information in memory [2-5]. The *central* aspect of attention indicates that the resource is shared between all modalities (vision, hearing, etc.) and types of information coding (phonological, orthographic, spatial, etc.). The *limited-capacity* aspect indicates that one type of manipulation or storage can be increased only at the expense of other types.

The allocation of attention can be modified voluntarily, as is indicated when participants adjust their attention allocation according to variable instructions or payoffs. However, attention is not completely voluntary: flashing lights, loud noises, motion (or “pop-ups”) on a website, and sudden movements can involuntarily grab attention away from an intended task. Such a distraction can cause a lapse in attention and, by extension, a failure to remember critical components of the task that was being completed (e.g., the name of your new acquaintance; the context of the newspaper article you were just reading).

#### The neuroscience of working memory

The working memory system is located diffusely throughout the brain and appears to rely mainly on the same brain

Testing attentional control: Flashing lights grab people’s attention, and advertisers often exploit this as a way to get us to look at their ads. It is also used in laboratory tasks to measure attentional control. The anti-saccade task [6] is one such laboratory task. (A saccade is a quick eye movement that we make from one spot to another in order to look at an object.) In this task, individuals are shown a screen with only a single fixation point in the center to look at (see Figure 1). Then an object suddenly appears on either the right or left side of the screen. Participants are instructed ahead of time to either look toward the new object (pro-saccade) or away from it (anti-saccade). The natural response is to look toward the newly appearing object because it captures our attention; looking in the other direction takes attentional control. People with greater working memory capacity are better able to voluntarily control attention and direct it away from the flashing object [7].

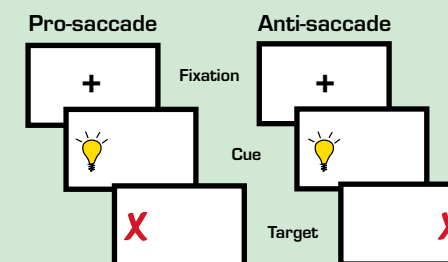


Figure 1: Attentional capture in the pro-saccade and anti-saccade tasks. In the pro-saccade task, participants are always told to look toward the flashing light in anticipation of the red “X”, which flickers only briefly on the same side of the screen. But in the anti-saccade task they must look away from the flashing light in order not to miss the red “X”’s brief presentation on the opposite side of the screen. The flashing light captures attention, but resisting the urge to look in the direction of the red “X” in the anti-saccade task takes attentional control.



regions as general fluid intelligence. The *control* center of the working memory system is located in the prefrontal cortex (PFC), the area at the front of the brain above the eyes. The primary function of the PFC is to orchestrate, coordinate, and guide our thoughts and actions in concert with current task goals and intentions. Some individuals have greater control over their thoughts and actions than others.

Studies of patients with traumatic brain injury to the PFC, Alzheimer's dementia, Parkinson's disease, and regular aging have shown that interruptions and impairment of the PFC make multi-tasking difficult or impossible. In a classic example of this, Baddeley, Bressi, Della Sala, Logie, and Spinnler [8] compared the dual-tasking performance of young and elderly adults to that of patients with Alzheimer's dementia. The first of two tasks tested participants' memory for sequences of digits. The second of the two tasks tested motor control and the ability to track a moving light with a hand-held stylus. The Alzheimer's patients and the young and elderly adults were equally good in their performance when these tasks were completed separately. However, only the Alzheimer's patients were significantly impaired when they tried to combine the tasks by remembering digits while tracking the light at the same time.

Consistent with previous findings, the Alzheimer's patients had some unimpaired memory and motor function, but their dual-tasking ability was severely compromised. Baddeley et al.'s results suggest that age does not necessarily influence the capacity for dual tasking, but that the loss of neurons and synapses in cortical and sub-cortical regions associated with Alzheimer's dementia does have a significant impact.

### **CASL's working memory workout**

Whether *sweatin' to the oldies* or *tackling Tae-Bo*, science has long extolled the benefits of physical fitness training.

While you have probably heard of the aerobic exercise *spinning*, you probably have never heard of *spanning*, which is a term CASL researchers coined for their new memory exercise program, designed not for physical fitness but for memory fitness. CASL researchers are developing a working memory workout designed to increase working memory *span*—the length of the information sequence one can hold in immediate memory—through attentional control training.

CASL's attention and working memory training program is designed to stimulate brain activity and promote the sort of higher-level cognitive functioning required for text comprehension and language analysis. Research in cognitive psychology and cognitive neuroscience, including work done at CASL, suggests that the attentional processes that make multi-tasking possible may be critical for first and foreign language text analysis and reading comprehension. Indeed, the same regions of PFC that are known to be involved in the *control* aspects of working memory are also involved in language processing tasks. Brain-imaging studies have demonstrated an overlap in the location of brain activity when healthy adults perform anti-saccade-like tasks (Figure 1) and certain types of language processing tasks that are complex or ambiguous—tasks that also require attentional control (see examples that follow). Likewise, patients with brain lesions to the PFC exhibit a failure to correctly perform working memory tasks that require attention control (like the anti-saccade task) as well as a failure to correctly understand grammatically complex or ambiguous sentences.

### **How do working memory training programs work?**

Working memory training programs capitalize on the brain's enormous capacity to adapt to changes in the environment. People differ in their working memory abilities, and these differences have a direct impact on their daily lives and job performance. Individuals with

greater working memory resources have better attentional focus and better short-term memory; Individuals with fewer working memory resources are more easily distracted and are more likely to forget information they were trying to use. Recall the previous example of the difficulty people sometimes have when trying to pick up a conversation following brief and unexpected interruptions.

CASL researchers hypothesize that the proposed working memory/attentional control training program stimulates the PFC. As the control center of the working memory system, the PFC is the brain region responsible for fast and efficient working memory performance. CASL's research scientists are challenging the conventional wisdom that working memory capacity is innate and cannot be enhanced through training.

A critical assumption of this work is that cognitive abilities, and the underlying brain structures that support them, are amenable to change. CASL's researchers hypothesize that engagement in cognitively demanding tasks leads to functional changes in the neurological pathways that sustain effective and efficient working memory processes.

One of the most important recent findings to emerge from the cognitive neuroscience literature is that of neural and cognitive plasticity in adults [9,10]. Previously, scientists believed only children had such plasticity. However, recently scientists have found that adults' neurological pathways and structures also appear to change in response to repeated engagement in particular tasks [11-13]. Importantly, changes in neural structure can result from the day-to-day engagement in routine activities or through extensive training or practice [14-19].

CASL's *working memory workout* is still in development, so we cannot convey all the details of the research in this article. However, we can mention that the program consists of multiple attention and memory exercises, some of which are based on traditional behavioral laboratory measures of working memory capacity.

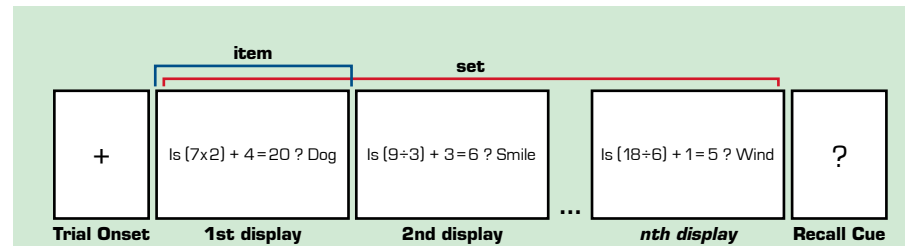


Figure 2. Graphical depiction of an  $n$ -item set in the operation span task, where  $n$  typically ranges from 2 to 6. At trial onset the participant reads the string aloud beginning at “is.” At “?” in the string, the participant says “yes” if the equation is correct or “no” if it is incorrect. The participant reads aloud and retains each word until recall is prompted.

### How do researchers develop working memory training tasks?

Researchers rely on traditional laboratory measures of working memory to develop working-memory training tasks. A group of tasks known collectively as *complex span tasks* are typically used to measure verbal working memory span. Complex span tasks are tasks that actually involve performing two separate tasks concurrently, combining a memory requirement with an attentional control requirement.

A commonly used complex span task is the *operation span task* (O-Span) [20]. Figure 2 depicts the traditional O-Span task for a single O-Span item (top) and a single trial (bottom). Each *item* consists of a mathematical equation paired with a to-be-remembered word. Each *trial* consists of a series of such items. A trial begins with the presentation of a fixation point followed by a mathematical equation and the to-be-remembered word. For each item, the participant reads the equation aloud, indicates if the equation is correct or incorrect by saying yes or no, and then reads the word aloud. After reading the word, the experimenter advances to the next equation-word pair, and the procedure repeats. At the end of the trial, the participant receives a cue to recall all the words in serial order; the number of words correctly reported reveals the participant’s span.

To transform a behavioral laboratory task such as the O-Span task into a

working-memory *training* task, researchers must make several modifications. The task needs to progress in difficulty from easy to difficult, adapting dynamically as an individual’s performance improves (i.e., as their span increases), and the task needs to offer users feedback on their performance. Repeated practice at regular intervals is also critical.

### Does working memory training improve performance in real-world tasks?

Yes. Recent work in the field of cognitive neuroscience demonstrates the effectiveness of extensive attention, perception, and memory training for improving overall cognitive functioning in both children and adults. For example, performance on tests of general fluid intelligence, as well as on mathematical reasoning, has been shown to increase (compared to relevant control conditions) following as little as 10 to 20 hours of cognitive training [21].

Strikingly, cognitive training has been shown to reverse age-related declines in cognitive performance, decrease automobile accidents in elderly adults, and even reduce behavioral symptoms associated with childhood ADHD. These results, and others, speak to the efficacy of cognitive training for improving functioning across a wide variety of tasks and across a wide variety of individuals.

Further, these results suggest that by increasing attention allocation and work-

ing memory capacity, people may also be able to improve their performance in more general areas that rely on attention and working memory, including language. Although prior work has shown that cognitive training can improve performance on psychometric tests of ability, no work has tested the hypothesis that it results in real gains in real-time language comprehension or the interpretations readers assign to sentences.

### Why is working memory training relevant to language?

Language processing, comprehension, and translation require attention. During reading and spoken language processing, readers and listeners have to extract the meaning of the individual words they encounter, and also package them into phrases to form a coherent sentence that accurately reveals who is doing what to whom, that is, the general content of what the writer or speaker is intending to communicate.

Many factors can adversely affect reading comprehension, including textual inconsistencies within and across sentences; complex or unusual vocabulary within the text; the reader’s unfamiliarity with the subject matter; infrequent words or unexpected propositions; and the reader’s cognitive skills, including memory capacity and attentional control. These factors affect reading comprehension in a person’s first *and* second language.

Prior research has found that working memory supports the comprehension process of resolving ambiguities in text, which arise routinely during language interpretation. Evidence from language processing studies indicates that readers and listeners interpret sentence meaning in real-time as words and phrases are perceived. In other words, readers and listeners do not wait until the end of an utterance to begin assigning an interpretation to what they just read or heard [22-26]. While economical, one important consequence of real-time processing is that it routinely comes at the cost of

having to deal with temporary ambiguity. Consider for instance sentence (A):

*(A) The agents finally discovered the crucial evidence could not be located with GPS.*

As the words in example A are encountered, a reader begins to interpret the input in a way that is consistent with his or her prior knowledge of the verb *discovered*, namely that it is highly likely to appear with some discoverable object, reflecting that something was indeed discovered (e.g., *The agents discovered the evidence and sent it to the lab.*). However, such knowledge is insufficient to be an exact guide to how the rest of the sentence will unfold, as illustrated in (example A), where new input, *could not be located with GPS*, is encountered and unambiguously signals that the agents actually did not discover the evidence! In other words, at the point in the sentence when readers encounter *the evidence*, the sentence is temporarily ambiguous.

When encountering later-arriving information—*could not be located*—readers tend to exhibit processing difficulty reflected by elevated reading times and/or several re-reads, presumably because they have to “slam on the brakes”—they must override their initial interpretation of *the evidence* as something the agents discovered, and instead recover an alternative meaning that allows *the evidence* to refer to something that could not in fact be located [27]. This processing difficulty is known as the *garden-path effect*: readers (or listeners) are led down one path of interpretation in light of strong developing information that points toward one likely interpretation, only to find that this initial interpretation commitment is ultimately wrong once new information is encountered. CASL researchers are currently recording readers’ eye movements as they make their way through ambiguous sentences of the garden-path variety,

to collect fine-grain, real-time data about where in a sentence difficulty is encountered (and how eye movement patterns change after working memory training, i.e., how real-time processing difficulty is alleviated).

Moreover, under such ambiguous conditions, readers tend to make more errors when answering comprehension questions that probe whether initially misunderstood elements of an ambiguous phrase linger by continuing to influence readers [28]. For instance, when asked questions such as, “Did the agents discover the evidence?” many people (with lower working memory capacity) will answer yes, even though the correct answer is no. Thus, the incremental processing of sentences can place high demands on the language processing and attentional control systems, especially when the initial interpretation of a temporary ambiguity must be overridden in light of later arriving linguistic material.

The ability to deal with ambiguous and complex language has long been tied to individuals’ working memory capacity and attentional control abilities [28-32]. The kind of ambiguity resolution scenario described in example A and the cognitive processes needed to resolve it have also been equated with those needed to successfully complete general non-linguistic attentional control tasks (e.g., the difficulty most people have when directing their gaze in the opposite direction of an attention-grabbing image; see Figure 1). Neurologically and cognitively speaking, working memory, ambiguity resolution, and non-linguistic attentional control appear to engage many of the same neural systems within the PFC [33-36].

**So, working memory helps readers and listeners avoid misleading assumptions?**

A failure to consider alternative sentence meanings and fully resolve ambiguities in real-time could result in a mischaracterization of sentence meaning that ultimately has important implications. Consider: a foreign language practitioner

encounters complex or ambiguous material in her foreign language to translate, such as the example in example A, and thus has to rely on working memory and attentional control. A lapse in attention may cause a foreign language practitioner to arrive first at the most reliable or frequent interpretation of sentence meaning (that the evidence was discovered), but not override it (e.g., in example A) above, translating the text to reflect that certain evidence was discovered when indeed it was not.

A large-scale, intensive training regimen that targets the attentional control aspects of working memory may therefore have important effects on real-time language processing and the ultimate resolution of ambiguity at multiple levels. We predict that after cognitive training, readers will experience less on-line reading difficulty and fewer overall comprehension errors linked to complex or ambiguous language—in other words, we predict that the effects of complexity and ambiguity will diminish. Streamlining reading processes by “greasing” the cognitive wheels that support them has obvious implications for language and intelligence analysts.

## Conclusion

The study of language comprehension has a long and rich history, and much is known about how people process linguistic information. However, despite the many theoretical and empirical advances that have been made, relatively little work has been devoted to developing training procedures that improve first and foreign language processing. By increasing trainees’ working memory capacity and attentional control through CASL’s spanning *workout*, language practitioners may be able to improve their analysis performance, particularly when texts are corrupt, incomplete, ambiguous, or otherwise difficult to process. Because analysts typically work with materials that are not in their first language, dealing with these challenges is especially hard and places significant burdens on their work-



**If you think the structure of the universe is complex, consider how the interconnectedness of billions of neurons self-organize to produce a single spoken word, a sentence, or a lengthy conversation.**

Language comprehension is a multifaceted and complex process. It requires access to a mental lexicon (the repository of words stored in your mind) and implicit knowledge of syntactic rules so we know how to correctly package words and phrases together in order to figure out who is doing what to whom in a sentence. The working memory system is only one small piece of this process, but it is a crucial piece. Example A (*The agents finally discovered the crucial evidence could not be located with GPS.*) is just one of many that illustrate the need for careful attention and working memory so as not to experience reading difficulty or ultimately misinterpret a sentence.

Working memory and attentional control also help readers maintain information across long spans within (and across) sentences. Consider the following:

*The little girl wearing a green hat and grey coat, who was sitting next to an old lady with a tiny umbrella, fell asleep on the train.*

A reader here has to hold *the little girl* in memory for a long time before he or she gets to know what happened—that *she fell asleep on the train*. Having a smaller working memory span and attentional control can cause someone to quickly forget (or at least be confused about) the subject of the sentence (the little girl rather than the old lady). Thus, the reader with lower working-memory capacity may be required to re-read or backtrack, which will result in a slower comprehension process and a greater susceptibility to erroneous interpretation (e.g., thinking that the old lady fell asleep, due to the physical proximity of the phrase *an old lady with a tiny umbrella* to the phrase *fell asleep on the train*).

In addition, having a smaller working memory capacity can also impede the rate of learning a new language, delay reading time, decrease comprehension, and lead to errors in linguistic inferences. Low working memory capacity has been implicated as a reason for reduced comprehension of metaphoric expressions such as *the masked hijackers were politicians* [37], in which the literal meaning of a sentence cannot be taken at face value. With metaphoric expressions, readers must suppress the literal meaning to promote an alternative, contextually relevant meaning (perhaps that the hijackers were not really senators, for instance, but rather maintained politician-like qualities including negotiation, posturing, etc.).

Misinterpreting such a phrase could have negative consequences. Because readers and listeners, including foreign language practitioners, frequently deal with this type of material, the psycholinguistic finding that readers with lower working memory capacities tend to provide incorrect analyses of metaphoric expressions ought to inform how we think about improving language analysis. This finding is especially relevant given that processing and comprehending input in someone's second language taxes that person's working memory more than first language processing, especially at lower proficiency levels—understanding in such cases is less *automatic*. We believe, therefore, that training working memory is fertile experimental ground.

Attentional control and working memory are also required for many other aspects of language use and comprehension. *Code switching*—a term linguists use to describe the use of two or more languages in a conversation—is a prime example of having to control the shifting of attention among multiple tasks. Likewise, translation and interpretation from one language to another requires engaging multiple languages simultaneously.

Other research shows that working memory capacity co-varies with the rates of first and foreign language vocabulary learning, with reading and listening comprehension, and with writing proficiency [38-41]. In vocabulary learning, for example, working memory makes it possible to hold a newly learned word in memory while learning to associate it with known words and meanings.

Taken together, prior work has demonstrated a compelling correspondence between working memory ability and the ability to engage in successful language understanding at multiple levels, including resolving ambiguity and mitigating against its lingering effects, metaphor comprehension, and switching between one's first and foreign language within a single conversation. Consequently, pairing this research with the research on working memory training may have important implications: enhancing the working memory system through adaptive training may have valuable applied outcomes for language comprehension, including improved performance on tasks of ambiguity, metaphor generation and comprehension, and code switching—all tasks routinely undertaken by foreign language practitioners.

ing memory and attentional control systems. By training this system—which is critically relevant for resolving complex, incomplete, or ambiguous language—CASL researchers hope to improve the performance of language practitioners. 🌱

## References

- [1] Clarke T. Brain fitness seen as hot industry of the future. Reuters [Internet]. 2008 Mar 12. Available from: <http://www.reuters.com/article/technologyNews/idUSN1218668920080312?feedType=RSS&feedName=technologyNews&pageNumber=2&virtualBrandChannel=0&sp=true>
- [2] Baddeley AD, Hitch GJ. Working memory. In: Bower GA, editor. Recent advances in learning and motivation. New York (NY): Academic Press; 1974. vol. 8, p. 47-90.
- [3] Cowan N. Working memory capacity. New York (NY): Psychology Press; 2005.
- [4] Engle RW, Kane MJ, Tuholski SW. Individual differences in working memory capacity and what they tell us about controlled attention, general fluid intelligence, and functions of the prefrontal cortex. In: Miyake A, Shah P, editors. Models of working memory: Mechanisms of active maintenance and executive control. New York (NY): Cambridge University Press; 1999. p. 102-134.
- [5] Kane MJ, Hambrick DZ, Tuholski SW, Wilhelm O, Payne TW, Engle RW. The generality of working memory capacity: A latent-variable approach to verbal and visuospatial memory span and reasoning. *Journal of Experimental Psychology: General*. 2004;133:189-217.
- [6] Everling S, Fischer B. The antisaccade: A review of basic research and clinical studies. *Neuropsychologia*. 1998;36:885-899.
- [7] Unsworth N, Schrock JC, Engle RW. Working memory capacity and the antisaccade task: Individual differences in voluntary saccade control. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. 2004;30:1302-1321.
- [8]

- working memory in Alzheimer's disease: A longitudinal study. *Brain*. 1991;114:2521-2542.
- [9] Greenwood PM. Functional plasticity in cognitive aging: Review and hypothesis. *Neuropsychology*. 2007;21:657-673.
- [10] Mercado E 3rd. Neural and cognitive plasticity: From maps to minds. *Psychological Bulletin*. 2008;134:109-137.
- [11] Huttenlocher PR. Synaptogenesis in human cerebral cortex. In: Dawson G, Fischer KW, editors. *Human behavior and the developing brain*. New York (NY): Guilford Press; 1994. p. 137-152.
- [12] Katz LC, Shatz CJ. Synaptic activity and the construction of cortical circuits. *Science*. 1996;274:1133-1138.
- [13] Kolb B, Whishaw IQ. Brain plasticity and behavior. *Annual Review of Psychology*. 1998;49:43-64.
- [14] Garlick D. Understanding the nature of the general factor of intelligence: The role of individual differences in neural plasticity as an explanatory mechanism. *Psychological Review*. 2002;109:116-136.
- [15] Garlick D. Integrating brain science research with intelligence research. *Current Directions in Psychological Science*. 2003;12:185-189.
- [16] Nyberg L, Sandblom J, Jones S, Stigsdotter Neely A, Petersson KM, Ingvar M, Backman L. Neural correlates of training-related memory improvement in adulthood and aging. *Proceedings of the National Academy of Sciences, USA*; 2003. vol. 100, p. 13728-13733.
- [17] May A, Hajak G, Ganßbauer S, Steffens T, Langguth B, Kleinjung T, Eichhammer P. Structural brain alterations following 5 days of intervention: Dynamic aspects of neuroplasticity. *Cerebral Cortex*. 2007;17:205-210.
- [18] Olesen P, Westerberg H, Klingberg T. Increased prefrontal and parietal activity after training of working memory. *Nature Neuroscience*. 2004;7:75-79.
- [19] Temple E, Deutsch GK, Poldrack RA, Miller SL, Tallal P, Merzenich MM, Gabrieli JDE. Neural deficits in children with dyslexia ameliorated by behavioral remediation: Evidence from functional MRI. *Proceedings of the National Academy of Sciences, USA*; 2003. vol. 100, p. 2860-2865.
- [20] Turner ML, Engle RW. Is working memory capacity task dependent? *Journal of Memory and Language*. 1989;28:127-154.
- [21] Jaeggi SM, Buschkuhl M, Jonides J, Perrig WJ. Improving fluid intelligence with training on working memory. *Proceedings of the National Academy of Sciences, USA*; 2008. vol. 105, p. 6829-6833.
- [22] Altmann GTM, Kamide Y. Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition*. 1999;73:247-264.
- [23] Ferreira F, Clifton C. The independence of syntactic processing. *Journal of Memory and Language*. 1986;25:75-87.
- [24] Rayner K, Carlson M, Frazier L. The interaction of syntax and semantics during sentence processing: Eye movements in the analysis of semantically biased sentences. *Journal of Verbal Learning and Verbal Behavior*. 1983;22:358-374.
- [25] Trueswell JC, Tanenhaus MK, Garnsey SM. Semantic influences on parsing: Use of thematic role information in syntactic ambiguity resolution. *Journal of Memory and Language*. 1994;33:285-318.
- [26] Van Berkum JJA, Brown CM, Zwitserlood P, Kooijman V, Hagoort P. Anticipating upcoming words in discourse: Evidence from ERPs and reading times. *Journal of Experimental Psychology: Learning, Memory and Cognition*. 2005;31:443-466.
- [27] Garnsey SM, Pearlmutter NJ, Myers E, Lotocky MA. The contributions of verb bias and plausibility to the comprehension of temporarily ambiguous sentences. *Journal of Memory and Language*. 1997;37:58-93.
- [28] Christianson K, Williams CC, Zacks RT, Ferreira F. Younger and older adults' "good-enough" interpretations of garden-path sentences. *Discourse Processes*. 2006;42:205-238.
- [29] Daneman M, Carpenter PA. Individual differences in working memory and reading. *Journal of Verbal Learning & Verbal Behavior*. 1980;19:450-466.
- [30] Fedorenko E, Gibson E, Rohde D. The nature of working memory capacity in sentence comprehension: Evidence against domain-specific working memory resources. *Journal of Memory and Language*. 2006;54:541-553.
- [31] Kemper S, Liu C. Eye movements of young and older adults during reading. *Psychology and Aging*. 2007;22:84-93.
- [32] Miyake A, Just MA, Carpenter PA. Working memory constraints on the resolution of lexical ambiguity: Maintaining multiple interpretations in neutral contexts. *Journal of Memory and Language*. 1994;33:175-202.
- [33] Fiebach CJ, Vos SH, Friederici AD. Neural correlates of syntactic ambiguity in sentence comprehension for low and high span readers. *Journal of Cognitive Neuroscience*. 2004;16:1562-1575.
- [34] Mason RA, Just MA, Keller TA, Carpenter PA. Ambiguity in the brain: What brain imaging reveals about the processing of syntactically ambiguous sentences. *Journal of Experimental Psychology: Learning, Memory, & Cognition*. 2003;29:1319-1338.
- [35] Ye Z, Zhou X. Involvement of cognitive control in sentence comprehension: Evidence from ERPs. *Brain Research*. 2008;1203:103-115.
- [36] Novick JM, Trueswell JC, Thompson-Schill SL. Cognitive control and parsing: Reexamining the role of Broca's area in sentence comprehension. *Cognitive, Affective, & Behavioral Neuroscience*. 2005;5:263-281.
- [37] Chiappe DL, Chiappe P. The role of working memory in metaphor production and comprehension. *Journal of Memory and Language*. 2007;56:172-188.
- [38] Atkins PWB, Baddeley AD. Working memory and distributed vocabulary learning. *Applied Psycholinguistics*. 1998;19:537-552.
- [39] Baddeley AD. Working memory and language: An overview. *Journal of Communication Disorders*. 2002;36:189-208.
- [40] Daneman M, Hannon B. What do working memory span tasks like reading span really measure? In: Osaka N, Logie RH, D'Esposito M, editors. *The cognitive neuroscience of working memory*. New York (NY): Oxford University Press; 2007. p. 21-41.
- [41] Engle RW. What is working-memory capacity? In: Roediger HL 3rd, Nairne JS, editors. *The nature of remembering: Essays in honor of Robert G. Crowder*. Washington (DC): American Psychological Association; 2001. p. 297-314.

### Further Reading

Rakauskas ME, Ward NJ, Boer ER, Bernat EM, Cadwallader M, Patrick C J. Combined effects of alcohol and distraction on driving performance. *Accident Analysis & Prevention*. 2008; 40: 1742-1749.



# “Thinking Out of the Box” Through Cognitive Neuroscience

Imagine you are interpreting an utterance in a foreign language. The utterance contains an ambiguous word—let’s pretend the word in the foreign language is *dax*—and you can only glean the gist of what the sentence means: “onto an airplane, load kitchen and food supplies in the *dax*.” You ask yourself, *what does dax mean in this context?*

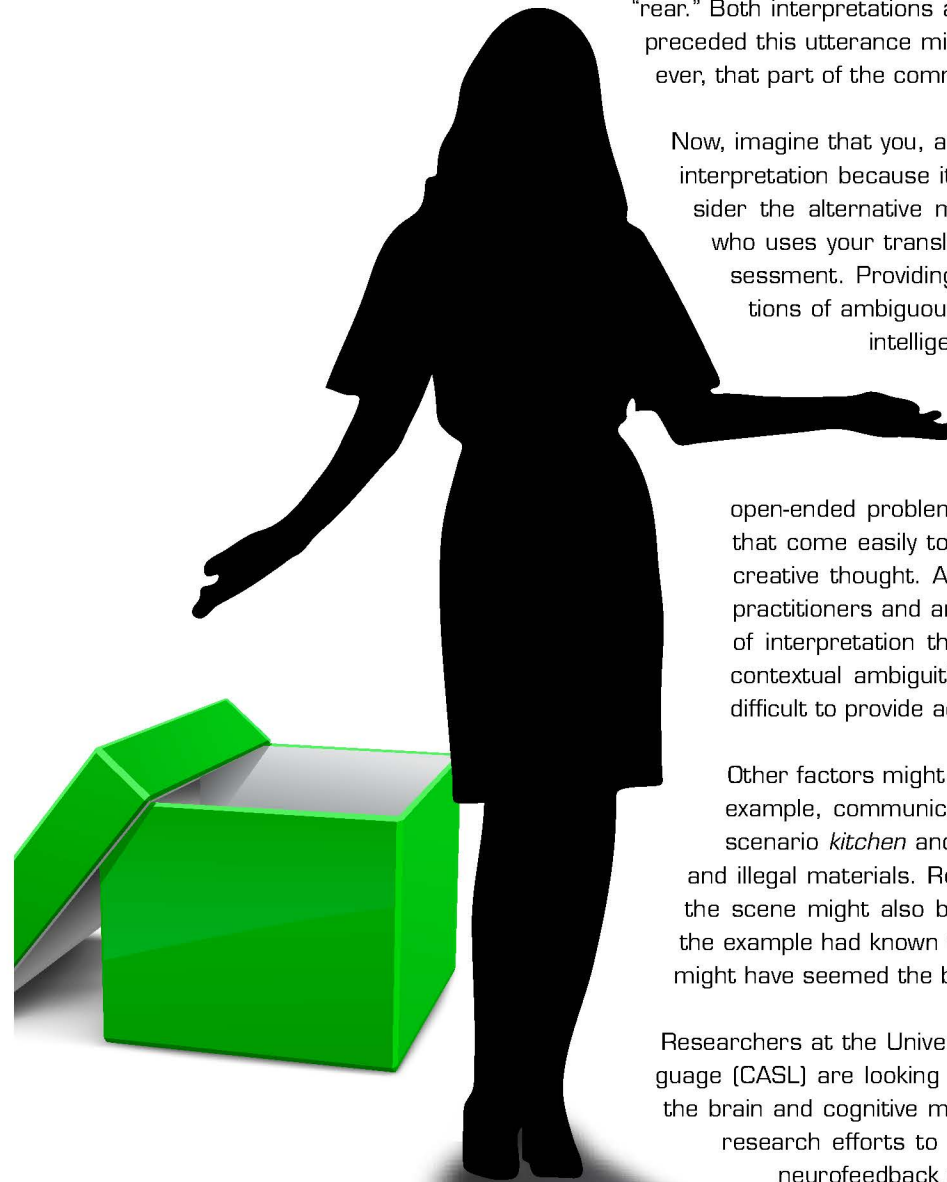
You recognize that *dax* usually means “afternoon.” However, *dax* can also mean “rear.” Both interpretations are plausible in this case. The linguistic context that preceded this utterance might have pointed toward the location meaning; however, that part of the communication is missing.

Now, imagine that you, a foreign language practitioner, select the “afternoon” interpretation because it is the more common usage of *dax*. Failure to consider the alternative meaning—“rear”—could lead the intelligence analyst who uses your translation to draw an incorrect and possibly perilous assessment. Providing scenarios that account for all possible interpretations of ambiguous terms requires foreign language practitioners and intelligence analysts to apply their best *divergent thinking* skills to the task.

Divergent thinking (DT) is the cognitive ability to think of as many useful solutions as possible for open-ended problems. These solutions include both common solutions that come easily to mind, and less common ones that require flexible, creative thought. As the opening scenario illustrates, foreign language practitioners and analysts require good DT skills because the problems of interpretation they encounter are often open-ended. Linguistic and contextual ambiguities, combined with incomplete transcripts, make it difficult to provide accurate interpretations.

Other factors might contribute to the interpretive challenge, as well. For example, communications can be intentionally distorted. In the opening scenario *kitchen* and *food supplies* could be euphemisms for dangerous and illegal materials. Relevant conditions such as weather or personnel on the scene might also be in flux. If the linguist translating the utterance in the example had known that a storm was predicted to clear soon, *afternoon* might have seemed the better word choice [1].

Researchers at the University of Maryland Center for Advanced Study of Language (CASL) are looking into ways to improve DT. In this article we discuss the brain and cognitive mechanisms involved in DT. We then describe CASL’s research efforts to improve analysts’ DT using two methods: (1) alpha neurofeedback training and (2) changing mental set.



### Brain and cognitive mechanisms

Cognitive neuroscience is aimed at understanding the neurobiological bases of human thought processes, including problem solving and attention. Recently, a panel from the National Academy of Sciences concluded that cognitive neuroscience methods are crucial for both assessing and inducing brain states that reflect expert performance of Intelligence Community (IC) and military professionals [2]. Two advances in cognitive neuroscience have proven to be particularly important for our understanding of the neurocognitive mechanisms underlying creative problem solving, including DT ability: (1) recent insights about the kinds of brain waves associated with good DT and (2) elucidation of the role of cognitive control mechanisms in the frontal parts of the brain, near the forehead.

### Brain waves and DT

Alpha brain waves have been associated with DT. Brain activity varies in intensity and, like radio waves, can be analyzed in terms of signals in various frequency ranges (Figure 1).

Alpha waves are regular sinusoidal brain waves. When a person is in a calm and alert state of mind, alpha waves are more prominent than theta or beta waves.

Hans Berger, the discoverer of the human EEG (electroencephalograph), observed that when a person closes his eyes, alpha waves become easily visible in recordings focused over the vision centers in the back of the brain. However, under certain conditions researchers also see an increase in alpha brain wave activity when subjects have their eyes open. Three states in which a person's alpha wave activity is elevated are (1) while trying to ignore a visible distracting stimulus, (2) while attending to an imagined stimulus, and (3) while holding information in the part of memory that stores word meaning. What these different conditions have in common is an inward direction of attention that frees the mind from actively processing potentially distracting stimuli

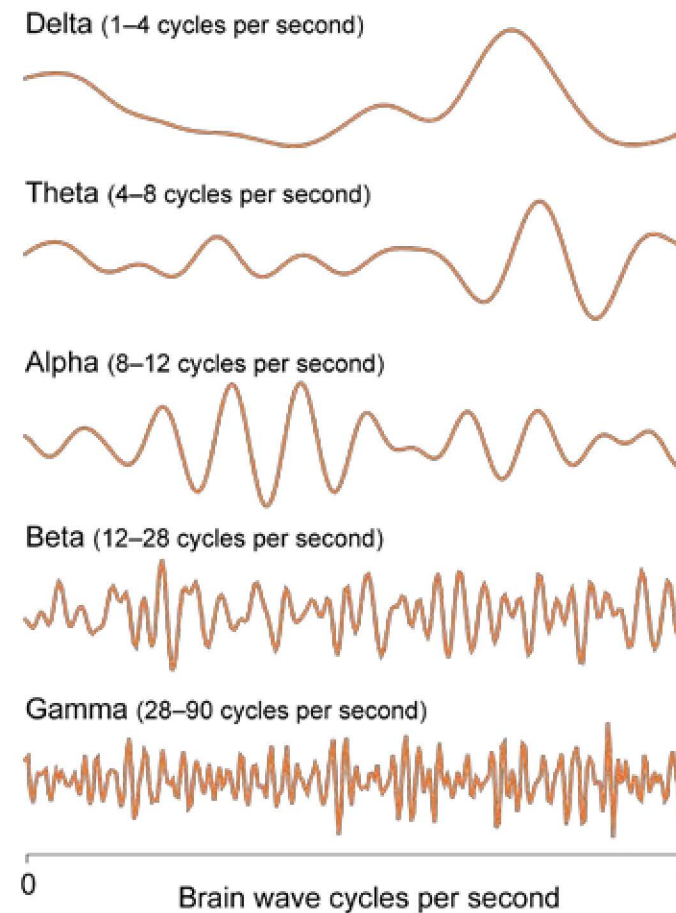


Figure 1: Brain waves

in the environment.

Recent cognitive neuroscience studies have found evidence to support the hypothesis that a problem solver in an elevated alpha state is much more receptive to becoming aware of his inner trains of thought. This heightened awareness extends to weakly associated thoughts, which are otherwise difficult to detect. Conceiving such remotely associated thoughts, or *weak associates*, is central to DT.

Jung-Beeman et al. [3] tested participants in EEG and fMRI (functional magnetic resonance imaging) studies for compound remote associates (CRA). During the CRA test, participants must think of a word that forms a compound word or phrase when associated with each of three other words. The test is designed to add

difficulty by including *solution words* that have a relatively weak association with each of their three related *stimulus words*. In the test a participant might be given the stimulus words *shoe*, *car*, and *French*, and have to come up with *horn* for the solution word. Immediately after answering, the participant reports whether he arrived at his answer more with insight—defined as a sudden awareness of the solution—or sheer mental processing effort.

Just before participants solved a problem with insight, the researchers observed an increase in alpha waves over the visual-spatial regions in the back of the right brain (specifically over a sensor location known as PO8). Heightened alpha states were followed by a burst of gamma waves. Jung-Beeman and colleagues reasoned that alpha waves reflected a recep-

tive state in which the visual-spatial part of the brain was not overwhelmed with processing potentially distracting stimuli, thus making linked regions more receptive to detect the weakly associated answer word. The subsequent gamma wave was hypothesized to reflect the sudden emergence of the insight solutions into awareness.

Further evidence for a role of alpha brain waves in creative problem solving was reported in an EEG study on DT by Grabner, Fink, and Neubauer [4]. Participants in the study were given an alternate uses (AU) task in which they had to generate as many uses for a given object as they could think of. The subjects were encouraged to focus on creative and original applications. For example, in response to the test word *brick* a participant might answer *paper weight, door stop, weapon*, and so forth. Just before the participant was ready to give another answer, he would press a button. After the test, the participant judged on a rating scale<sup>a</sup> the originality of his answers.

The study found that alpha activity over the right visual-spatial regions of the brain (including PO8) was greater when participants generated ideas they rated as more original. Results from the AU study were found to be similar to those obtained by Jung-Beeman and colleagues in terms of brain region and cognitive process. Weak meaning associates in the CRA test and original answers on the AU test both had to compete with more dominant answers for conscious access, indicating a relationship between alpha brain waves and receptivity to detecting less obvious answers. The observed increase in alpha activity is hypothesized to be due to the subjects' inwardly turned attention during an alert, distraction-free state.

### Executive control

Executive control processes enable goal-directed thought and action. Ori-

ginating in the prefrontal cortex, executive control processes enlist other parts of the brain, including the areas for memory and language, in the service of a person's immediate goals [6]. Attention control plays a crucial role in assigning meaning for directing these cognitive resources. Attention control allows the brain to override biased responses in the face of ambiguity and reconcile competing alternatives.

We believe that this meaning-based attention control is crucial for selecting weak associates over dominant ones and thereby promotes the production of useful, original solutions during DT. The same control system likely helps analysts think of uncommon interpretations. However, attention control has a limited capacity. If the cognitive demands of a task surpass the capacity for attention, then not enough attention resources remain to cope with effortful meaning selection. As a result, weak semantic associates may be overlooked in favor of stronger associates, which can lead to incorrect translations and false interpretations.

*CASL research aimed at improving analysts' divergent thinking*

### Alpha neurofeedback training

Alpha neurofeedback training (NT) is seen as one way to improve analysts' DT. In NT, brain waves are continuously monitored with one or more electrode sensors placed, noninvasively, on the trainee's scalp. Information about whether targeted brain waves are increasing or decreasing is continuously fed back to the learner. Using this feedback information, the trainee can learn to increase or decrease brain waves in selected frequency ranges, including alpha brain waves. Previous studies have shown that NT improves performance on working memory [7] and

mental rotation tasks in healthy adults [8]. In these studies, the NT was aimed at increasing brain waves in frequency ranges that were believed to be important for the assessed cognitive function.

Given the finding of an association between increased alpha brain waves and good creative problem solving over brain region PO8 [3,4], we hypothesize that NT aimed at teaching people how to increase their alpha brain waves over the same brain region should improve DT. Recall that PO8 is a region over the right posterior part of the brain.

### CASL research aimed at improving analysts' divergent thinking

A related reason for increasing alpha brain waves over a portion of the right (but not left) hemisphere of the brain is that the right hemisphere is superior in finding connections among unrelated ideas, which is one of the hallmarks of creative thinking. This capability probably stems from two sources: (1) the right hemisphere's ability to detect overlapping meanings for words that are not strongly associated and (2) its ability to integrate word meanings that have been derived from different contexts.

As a first step in our NT research, CASL recently completed a study with university students that demonstrates an effective NT method for increasing alpha brain waves over PO8 within a single training session [9]. Participants received three NT sessions of about half an hour each. The feedback signal for the test consisted of a moving bar on a display that rose higher as alpha brainwaves grew in intensity. Trainees were instructed to try to increase the height of the feedback bar as much as possible during 2.5-minute training segments. There were also one-minute baseline segments during which the bar did not move and no feedback was given.

We hypothesized that this alpha brain wave training would improve the

<sup>a</sup> Originality can also be determined more objectively as the inverse of the frequency with which a useful idea is generated among a large group of problem solvers [5]



trainees' ability to block out the alpha-decreasing effect of a moving visual stimulus, allowing brain regions necessary for creative problem solving to become more sensitive to internal associations. The major finding of the study was that our alpha NT protocol allowed most trainees—ten of thirteen—to learn to increase the intensity of their alpha brain waves over PO8 in just a single session. Twelve of the thirteen trainees showed improvement overall. Participants in a no-neurofeedback control group did not increase their alpha levels within a session. See Figure 2 for study results.

As the next step in our research, we will use our alpha NT protocol in an empirical study with government analysts and University of Maryland students. The study is designed to test the hypothesis that increasing alpha brain waves will improve performance on tests requiring DT. Participants will complete a DT test after NT in which they are presented short texts conveying scenarios that pose an open-ended problem that is difficult to solve, and for which they must generate as many solutions—common and uncommon—as possible. Before the test, participants will

be shown a small set of common solutions. The participants will be instructed that they cannot generate solutions from this set of well known solutions. We hypothesize that this experimental manipulation will induce a state of *mental fixation*, in which subjects will find it difficult to think of uncommon solutions in addition to the common ones already provided. These well-known solutions should make the participants think of closely related solutions that merely reinforce their awareness of the common solutions.

Mental fixation is likely to arise frequently in the workplace during both individual and group brainstorming, especially when the analysts have already generated several solutions—typically the more obvious ones—and realize that they must also consider further solutions that are less obvious but still potentially relevant. Indeed, in a study on DT with government analysts, we found that people generate more of the less obvious, original solutions during a later phase of an eight-minute problem-solving task posing scenario-based problems [5].

### Incubation

Incubation will be a critical manipulation in our next study with alpha NT. Incubation occurs during a period of time that a person has been distracted from a task before returning to the problem. The premise underlying incubation is that when mental fixation occurs, people solve problems better after some time has elapsed. This expected improvement in problem solving is due to the redirection of attention to something else. Redirection helps reduce the influence of the mental state that caused the mental fixation, making it easier to access the weak thought associations required for uncommon solutions.

Furthermore, to optimize the accessibility of weak thought associations, our next study will have participants engage in alpha NT during the incubation period. Incubation effects are well established in studies with convergent thinking tasks that present closed insight problems for which there is only a single solution, after changing from a dominant to a less obvious perspective. However, only a few behavioral studies have examined effects of

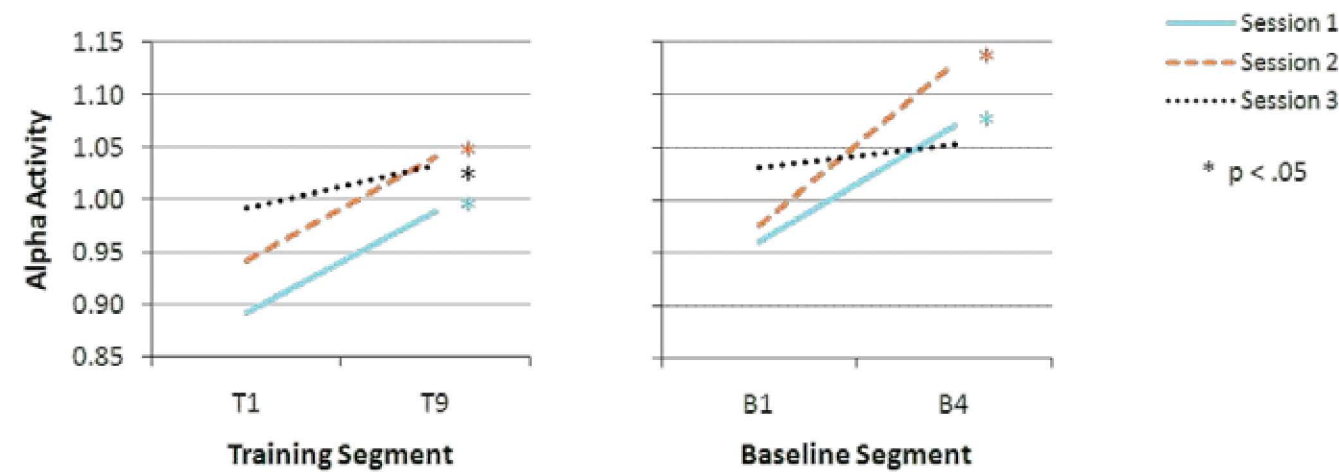


Figure 2: Alpha power increased within and across sessions

incubation on DT, and with mixed results. Moreover, the application of alpha NT as an incubation method is new.

An interesting aspect of alpha NT as a method for improving DT is related to the fact that there are systematic differences in creativity as an individual trait. For example, we assessed such individual differences in a study with university students and government analysts with a standard and novel task of DT and found that performance is predicted by the participant's ability to comprehend written texts. Alpha NT might allow people of average or low creative ability to become more creative in their problem solving. If after a few training sessions of alpha NT people can be taught to change their alpha waves up or down merely on cue without the alpha NT technique being used, the benefits could be easily adopted in the work place.

Developing methods for inducing brain states associated with good DT could increase the number of candidates the IC has to choose from for assignments that require higher levels of DT. With a larger pool of candidates to select from, the IC could more easily identify people with a difficult-to-find combination of knowledge and skills needed for analysis of technically specific assignments.

### Changing mental set

Mental set refers to the tendency to repeatedly think about and approach a problem in a habitual manner. Changing a person's *mind set* to one that is more effective provides another way to improve performance. DT can be used to train a person to see a problem through an alternative lens, thus bringing new candidate solutions in view and altering the prevailing mental set.

CASL researchers are currently adopting this approach by designing behavioral intervention tasks aimed broadly at overcoming fixation and generating alternative solutions. This work is motivated in part by results in the cognitive

creativity literature that show that people are more likely to solve insight problems after they have performed a variant of an AU task [10].

Previous studies have shown that AU techniques can be used to generate new and creative solutions to insight problems. AU research has tested whether insight problems were better solved by people who first completed an alternative categories test (ACT) compared to those who did not [10]. The ACT required generating uses for objects in alternative meaning categories, such as thinking of a shoe as a hammering tool instead of footwear. This test was completed prior to attempting to solve insight problems. The results reliably demonstrated improved problem solving performance in the ACT group, regardless of whether participants were explicitly aware that the training was relevant to the subsequent problem-solving tasks.

Successful problem solving may therefore hinge on the process of building *goal-derived categories* [11]—categories constructed on the fly to achieve a particular objective. In fact, further research revealed that ACT training enabled people to form a greater number of, and more variability in, goal-derived categorizations while problem solving across six different tasks [12]. Presumably, the ACT implicitly trained people to construct novel alternative categories based on the given problems, which in turn enhanced problem solving performance and enabled the problem solver to detect distant-meaning relations.

At CASL we are currently extending this behavioral intervention approach to test its effect on an ecologically valid DT test in which government analysts must generate as many common and uncommon solutions to an open-ended, scenarios-based problem in which mental fixation is induced. We predict that the most useful and original solutions will be provided by people who are trained to *escape* from a mental set by first completing

a set of intervention tasks such as the ACT. This intervention would be relatively easy to implement in the workplace.

### Conclusion

The cognitive neuroscience approach has been fruitful in uncovering the brain's information processing mechanisms that support creative problem solving, including DT and its component functions. The detailed functioning of these mechanisms remains to be further elucidated. Nevertheless, the time is ripe for a systematic exploration of ways to facilitate DT through various neuroscience and behavioral techniques for changing neurocognitive states, including alpha neurofeedback training and changing mental set. These approaches can be leveraged to ensure exceptional language analysis through better DT, which will help IC professionals anticipate the kinds of low-probability yet high-impact events that threaten security. 📌

Emerging cognitive neuroscience and related technologies. Washington (DC): National Academic Press; 2008.

[3] Jung-Beeman M, Bowden EM, Haberman J, Frymiare JL, Arambel-Liu S, Greenblatt R, et al. Neural activity when people solve verbal problems with insight. *PLoS Biol.* 2004;2(4):E97.

[4] Grabner RH, Fink A, Neubauer AC. Brain correlates of self-rated originality of ideas: Evidence from event-related power and phase-locking changes in the EEG. *Behavioral Neuroscience.* 2007;121(1):224-230.

[5] Haarmann HJ, Platt K, Donavos D, Fox L, Bowles A. Divergent thinking in intelligence analysts: an empirical study of incubation and predictors. College Park (MD): University of Maryland Center for Advanced Study of Language; 2008. Final Technical Report No.: E.3.1.

[6] Miller EK, Cohen JD. An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience.* 2001;24:167-202.

[7] Vernon DJ, Egner T, Cooper N, Compton T, Neilands C, Sheri A, et al. The effect of training distinct neurofeedback protocols on aspects of cognitive

performance. *International Journal of Psychophysiology.* 2003;47(1):75-85.

[8] Hanslmayr S, Sauseng P, Doppelmayr M, Schabus M, Klimesch W. Increasing individual upper alpha power by neurofeedback improves cognitive performance in human subjects. *Appl Psychophysiol Biofeedback.* 2005;30(1):1-10.

[9] Haarmann HJ, George T, Smaliy A, Grunewald K, Novick JM. A method for quickly increasing alpha brain waves through neurofeedback. College Park (MD): University of Maryland Center for Advanced Study of Language; 2008. Technical Report No.: E.3.2.

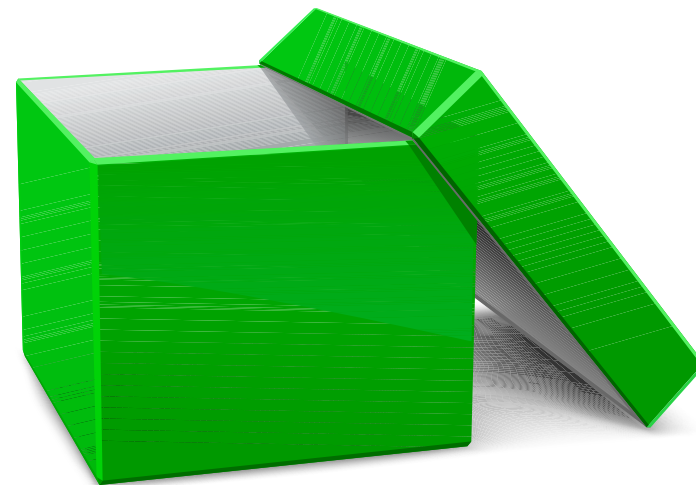
[10] Chrysikou EG. When shoes become hammers: Goal-derived categorization training enhances problem-solving performance. *J Exp Psychol Learn Mem Cogn.* 2006a;32(4):935-942.

[11] Barsalou LW. Ad hoc categories. *Mem Cognit.* 1983;11(3):211-227.

[12] Chrysikou EG, Alvarez K, Clarke JN. Do alternative categorizations improve problem solving? Evidence from a verbal protocol analysis. *Proceedings of the 28th Annual Conference of the Cognitive Science Society;* 2006b; Vancouver, BC (Canada).

Haarmann HJ. Cognitive neuroscience guidelines for improving divergent thinking. College Park (MD): University of Maryland Center for Advanced Study of Language; 2007. Final Technical Report No.: M.1.

[2] National Research Council (US), Committee on military and intelligence methodology for emergent neurophysiological and cognitive/neural research in the next two decades.





# For a Better Dictionary, Build a Better Parser



### The problem: dictionary lookup

People working with a language in which they're not fluent must often resort to finding words in the dictionary. Depending on the language, dictionary lookup can be problematic. For example, the language might have complex orthography—spelling rules—as with English, where the correspondence between letter and sound is notoriously far from simple. Upon hearing a word that begins with an *f* sound, new English learners may wonder whether to look it up under *f* or *ph*. And how do they know whether the long *a* sound they hear in a word is spelled *ay*<sup>a</sup> as in *lay, may, day*; *ai* as in *pain, maid, bait*; *ei*<sup>b</sup> as in *rein, skein, veil*; *ey* as in *they, hey, whey*; *eiht*<sup>c</sup> as in *eight, weight, neigh*; *a* (+ one consonant)<sup>d</sup> as in *nature, bacon, haven*; or *a* (+ silent *e*) as in *take, save, ace*?

Another common problem with dictionary lookup is determining what *form* of the word to look for. In general, dictionaries list only one head word, called the *citation form*, for a given word. For example, an English dictionary will have *rake* as the citation form in its entry for that verb, while the verb's other forms, *raked* and *raking*, if they appear at all, will be included only under the information for *rake*, rather than having their own entries further down the page.

The study of related word forms such as *rake/raked/raking*, their component parts, or *morphemes* (for example, *rak-*, *-ed*, and *-ing*), and the rules for putting those morphemes together is called *morphology*. We refer to *raked* and *raking* as *inflected* forms of the verb *rake*. *Rak-* is a stem; *-ed* and *-ing* are suffixes<sup>e</sup>. Automatically producing those inflected verbs as a speaker and analyzing them as a reader or listener are easy for even a beginning user of English, because the verb *rake* is regular—or *weak*, in English grammatical terminology—and weak English verbs follow very simple, regular rules of inflection. But some more examples from English will illustrate how even simple morphology is more complicated than a native speaker may realize.

<sup>a</sup> Which can also represent the long *e* sound, as in *quay*

<sup>b</sup> Which can also represent the short *i* sound, as in *weird*

<sup>c</sup> Which can also represent the long *i* sound, as in *height*

<sup>d</sup> Which can also represent short *a* sound, as in *satın*

<sup>e</sup> Most of the time in English, the stem and citation form are the same, as in the verbs *reach* (*reach-ed, reach-ing*), *jump* (*jump-ed, jump-ing*), *play* (*play-ed, play-ing*)

### Interagency Language Roundtable Language Skill Level Descriptions for Reading, Writing, Listening, and Speaking (from <http://www.govtilr.org/>)

Shaded area indicates range of expected proficiency upon completion of full-length DLI course.

PROFICIENCY LEVEL	DESCRIPTION
0	No Proficiency
0+	Memorized Proficiency
1	Elementary Proficiency
1+	Elementary Proficiency, Plus
2	Limited Working Proficiency
2+	Limited Working Proficiency, Plus
3	General Professional Proficiency
3+	General Professional Proficiency, Plus
4	Advanced Proficiency
4+	Advanced Proficiency, Plus
5	Functionally Native Proficiency

Most speakers of English are not explicitly aware that the English plural morpheme actually comes in three versions, pronounced *-s*, *-z*, and *-əz*. Thus we have, for example, *aardvark/aardvark-s*,

## Morpheme: the smallest unit of meaning in a language

**Stem:** Roughly, that part of a word that carries the main meaning. The stem of the English verb *rains*, for example, is *rain* (which also happens to be a word on its own). For the noun *raincoats*, the two morphemes *rain* and *coat* together constitute the stem.

**Prefix:** A morpheme that precedes a stem, as in English *re-*, *un-*, *bi-*.

**Suffix:** A morpheme that follows the stem, such as the two suffixes *-s* in *rains* and *raincoats*. While these two suffixes are spelled the same, they have different meanings: one is a present tense ending on a verb, while the other is a plural ending on a noun.

**Infix:** A morpheme that goes inside a stem. English does not have infixes, but later on we will talk about a language that does.

**Affix:** A prefix, suffix, or infix.

Example: *re-educat-ed*

*re-*prefix (= *again*); *educat-*stem; *ed-*suffix (past tense)

1. *-z* after vowels or the consonant sounds *b*, *d*, “hard” *g*, *l*, *m*, *n*, *ng*, *r*, *th*, or *v*
2. *-s* after the consonant sounds *f*, *k*, *p*, or *t*
3. *-əz* after the consonant sounds *ch*, “soft” *g*, *s*, *sh*, *z*, or *zh*

Similarly, the English past tense morpheme, spelled *-ed*, has three variants in spoken English, pronounced *-t*, *-d*, and *-əd*, as in *rake-d* (pronounced *rake-t*), *live-d*, and *wait-ed*. Which form is used depends, as with the plural morpheme, on the sound at the end of the word to which it is added:

1. *-d* after vowels and the consonant sounds *b*, (“hard” and “soft”) *g*, *l*, *m*, *n*, *ng*, *r*, *th*, *v*, or *z*
2. *-t* after the consonant sounds *ch*, *f*, *k*, *p*, *s*, *sh*, or *th*
3. *-əd* after the consonant sounds *d* or *t*

In writing, the plural is spelled *s* or *es*, while the past tense is nearly always spelled *ed*. If English were written closer to the way it is pronounced, we would instead have three spellings of each of these suffixes, reflecting their different spoken forms. This situation of morphemes with multiple forms is called *allomorphy*, and each variant form is called an *allomorph*. English also has many cases of allomorphy that do not follow rules and are therefore labeled *irregular*, such as *goose/geese*, *mouse/mice*, *louse/lice*,<sup>g</sup> etc. among nouns and *build/built*, *run/ran*, *bring/brought*, etc., among verbs. With these words, the allomorphy lies in the stem of the word, rather than in any affixes. Speakers must simply learn these forms, and non-fluent speakers, including young native speakers, often get them wrong until they become more fluent. Fortunately in the case of English, there are comprehensive dictionaries that list all the irregular forms and alphabetize them where one would expect them to be. But for many languages there are no comprehensive dictionaries, and the irregular forms are only listed in the main entry, or not at all.

### Inflectional morphology

The set of inflected forms of a word is referred to as its paradigm. Paradigms of nouns are often called declensions and paradigms of

*dog/dog-s* (pronounced *dog-z*), and *ostrich/ostrich-es* (pronounced *ostrich-əz*). Without thinking about it, native speakers choose an ending based on the following rules (with some exceptions):<sup>f</sup>

<sup>f</sup> Note that the rules are based on the final *sound*, not letter, of the word

<sup>g</sup> The two pairs *mouse/mice* and *louse/lice* appear to represent a regular pattern that speakers could learn to predict, but compare the forms *house/house-s*, *spouse/spouse-s*



verbs, conjugations. For English, the paradigm of a noun consists of the singular and plural forms (and perhaps the possessive forms); the paradigm of most English verbs consists of the bare form (like *walk*), the third person singular present tense form (*walks*), the present participle form (*walking*), and the past tense form (*walked*).<sup>h</sup>

Believe it or not, English morphology is relatively simple. In languages with complex morphology, the paradigms may be much more complicated, running into thousands of forms. Moreover, the citation forms of words with perfectly regular morphology are often not obvious to a non-fluent user of the language, because the inflected forms can be very different from the citation forms. For example:

- **Prefixes and suffixes:** In Swahili (a Bantu language of east Africa), verbs often have a large number of prefixes and suffixes. The inflected verb *amewanunulia*, for example, contains three prefixes and two suffixes: the prefix *a-*, meaning a class 1 subject (nouns in Swahili belong to one of a dozen or more classes, similar to gender in Romance languages, and verbs agree in class with their subjects); the prefix *me-*, meaning perfective aspect (similar to *has done* in English); the prefix *wa-*, meaning a class 2 object; the suffix *-li*, marking the “applicative” form of the verb (this has to do with how the direct and indirect objects appear in the sentence); and the suffix *-a*, whose meaning is somewhat difficult to pin down. The verb at the heart of this long word is *nunu*, meaning *to buy*.
- **Infixes:** Besides prefixes and suffixes, some languages employ *infixes*, affixes that go inside stems. In Tagalog, a language of the Philippines, the infix *-um-* can be used to create a noun from a verb. For example, the verb *sulat*, meaning *to write*, can take this infix to become *sumulat*, meaning a *writer*.
- **Reduplication:** Many languages employ a process called *reduplication*, in which all or part of a word is repeated. In Indonesian,

for example, *kira* means *guess*, while *kira-kira* means *approximately*. In this case, the entire word is repeated.

Tagalog gives us an example of partial repetition: from the verb *sulat* again, a sort of future tense is derived by reduplicating the first syllable to give *susulat*. Reduplication is not always this simple. Some Tagalog verbs, borrowed from Spanish, start with two consonants. For these verbs, reduplication involves only the first of the two consonants, plus the following vowel: *trabaho*, *to work* (from Spanish *trabajo*, *I work*) becomes *tatrabaho*, *I will work*.

- **Stem allomorphy:** As if prefixes, suffixes, infixes, and reduplication were not enough, languages often modify the forms of stems. Sometimes this pertains only to irregular words; in Spanish, for example, removing the *-es* suffix from the word *tiene*s, meaning *you have*, gives a stem *tien-*. But the form listed in the dictionary, *tener* (an infinitive) has a different stem *ten-*.

In other languages, stem changes are more or less regular. In Bangla, a language of Bangladesh and India, stems which contain two consonants display an alternation between the vowels *o* and *u*, depending on the following consonant: *shono*, *you hear*, but *shuni*, *I hear*.

#### Foreign language dictionaries

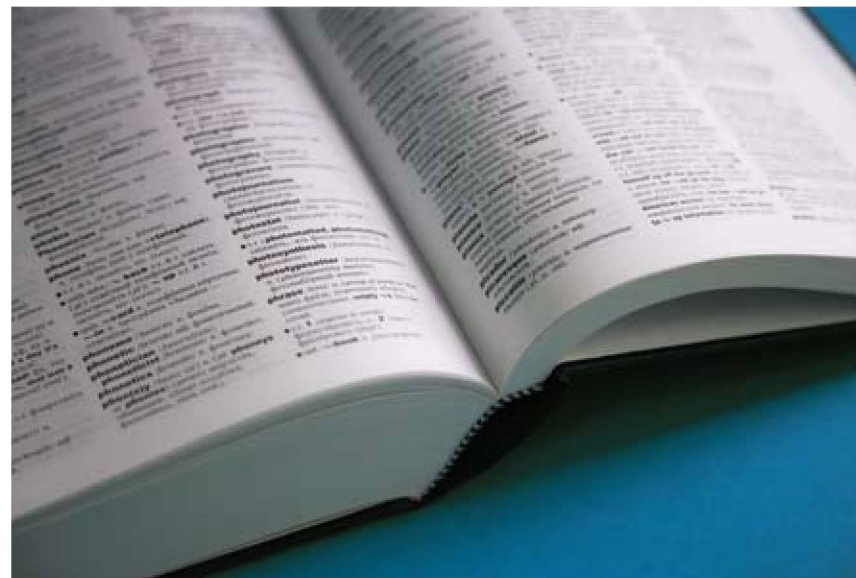
Significant inflectional morphology such as these examples illustrate can present serious difficulties for finding words in the dictionary. Since dictionaries do not normally list all such forms of words, morphological operations may make it difficult to look up a word. Affixes must be removed, and often other affixes must be attached. Take the Swahili example above, where the word *amewanunulia* contains the verb root *nunu*: Swahili dictionaries usually list this verb with the *-a* suffix attached to the root (but without the intervening applicative suffix *-li*), giving *nunua*. In other situations, the word may begin with prefixes, contain

<sup>h</sup> Irregular verbs often have a separate past participle form as well, as in *sung*, *given*, *gone*, and so on, although for other verbs, the past participle is the same as the past form

infixes, or the form of the stem may be modified.

Linguists who are fluent and experienced in a language can surmount these problems. Often, however, linguists work with languages in which they aren't sufficiently expert to know—or at least to be able to determine quickly—a word's citation form. They may have completed language training recently or been called on to work in a related language they don't know well. The examples above demonstrate how much detail a user of a morphologically complex language must know for efficient dictionary lookup, and that sort of knowledge comes only with time.

Difficulties with dictionary lookup can lead to significant loss of time and efficiency in translating. Linguists are often forced to spend their time and



energy searching through the dictionary, guessing at the citation form of an intricately inflected word. They need a tool to help them.

### **The solution: a morphological parser**

Fortunately, tools that help us find the dictionary form of a word exist; they are called morphological parsers. A morphological parser is an automatic tool that breaks up an inflected word into its morphemes. One of these morphemes will be the stem, from which the form of the word as it is listed in the dictionary can be produced and looked up in an electronic dictionary. The English

word *reruns*, for example, would parse into the prefix *re-*, the stem *run*, and the suffix *-s*. The words being analyzed may be all the words of a text, or individual words that are typed or pasted in to the computer.

Researchers at the University of Maryland Center for Advanced Study of Language (CASL) are building a morphological parser by developing a new, more efficient methodology that also answers the problem of software obsolescence. Our project, “Dual-Use Grammars in Related Languages,” has thus far focused on building morphological parsers for two South Asian languages: Bangla and Urdu. We are now beginning research on the Pashto language, which will pose new challenges due to dialectal variation and a scarcity of written resources.

Three things are needed to create a morphological parser for a language:

1. a list of the language's words, as they appear in a dictionary;
2. a list of the language's grammar rules, particularly the morphological rules; and
3. a morphological parsing engine.

### **The dictionary**

Obtaining an adequate dictionary can be a challenge. At a minimum, the entries must be labeled for part of speech (noun, verb, adverb, etc.). For some languages, additional information may be required:

- Noun class (All nouns in Spanish are either masculine or feminine; nouns in Bantu languages belong to one of a dozen or more classes.)
- Conjugation or declension class (These tell how a verb, noun, or adjective is inflected; for some languages, this can be inferred from the form of the word listed in the dictionary, while for others it cannot.)
- Irregular forms (such as *oxen* and *wept*)

Perhaps surprisingly, information about the meaning of a word is not needed for purposes of morphological parsing—although of course helping the user find its meaning is why we're doing the parsing in the first place!

Dictionaries of well-known languages such as French or German contain the information a parser requires, but dictionaries of less commonly taught languages are less likely to be complete.

### Example of a parsed Bangla verb

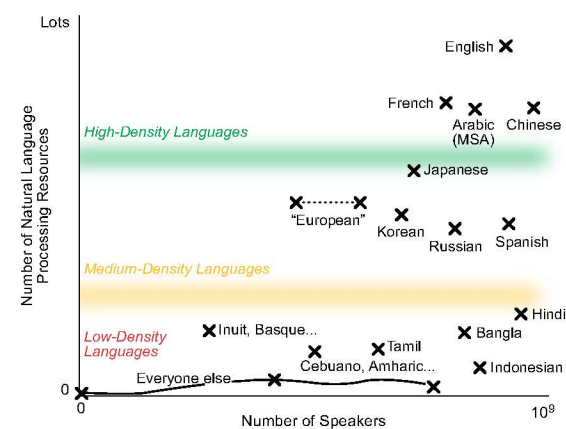
শিখিয়েছিস  
 /shikh-iyech-ish/  
 teach-**PRS.PRF-2SG.IF**  
*you have taught*

citation form: শেখানো /shekha-no/ (-no is a suffix that appears on the citation form; notice that the stem is spelled differently in the original word and in the citation form)

The morpheme glosses mean that this word is the present perfect tense (**PRS.PRF**), second person (**2SG**) informal (**IF**) form of the verb meaning *teach*.

Additionally, dictionaries for many lesser known languages do not contain a sufficient number of words on which to base a parser. Dictionaries of better known languages often contain thirty or forty thousand words, which is adequate for most purposes. But for many smaller languages it may be difficult to find an electronic dictionary with more than ten thousand words—or for some written languages, any electronic dictionary. And of course the thousands of unwritten languages have no dictionaries at all.

Borrowed words present yet another challenge. For many dictionary compilers, it is a matter of pride to exclude loanwords from other languages, even when such words may be used frequently in speech or texts (particularly texts on technical topics). A non-native speaker may need to look for such words in dictionaries of related languages. But because the spelling (and pronunciation) of loanwords is often quite different from the accepted spelling in the word's original language, these words can be difficult to find in the other languages' dictionaries. The Tagalog word *iskor*, for example, is borrowed from the English word *score*—but you will not find *iskor* in an English dictionary! Chinese loanwords in Tagalog, or Russian loanwords in Uzbek, can likewise be hard to find in dictionaries.



### The grammar rules

As our examples show, morphological rules can be quite complex. But for a morphological parser, we need a description of the language's morphology that is clear and unambiguous. Ambiguity, however, is inherent to natural language. We therefore also need a grammar written in a formal—that is, computer-readable—language. So for each language, we require two grammars: an old-fashioned descriptive grammar in straightforward English prose *and* a formal grammar that can be



## Spelling standardization

In English, as well as languages like Spanish and French, we are accustomed to one “correct” spelling, although different countries may have different spellings for the same word (think of *color* and *colour*). But many written languages lack such standards, and words are simply spelled as they sound, making it difficult to look words up. There may be different ways to write words depending on pronunciation, which in turn depends on dialect, or how fast one is speaking. And of course there is no accepted writing system for some languages. This is particularly true for many minority languages of Africa or Asia and for many sign languages.

automatically converted to serve as the grammar rules for the parser.<sup>1</sup>

### The parsing engine

The third component, the parsing engine, is a computational tool that incorporates information about how languages work in general, such as the fact that prefixes precede a stem, and suffixes follow a stem. This generic quality is why linguists need only provide a dictionary and the grammar of the language—what the particular affixes and stems of that language are, for example, and how they fit together. There are several advantages to keeping language-specific information (the dictionary and the grammar) separate from the information that is true for all languages (information which is built directly into the parsing engine). First, separating the two kinds of information in this way makes it easier and faster to build the parser. It also makes later modifications to the parser easier; for example,

correcting errors in the grammar, adding a new dictionary, or adapting the grammar to a related language without starting from scratch.

## Challenges of building parsers

Bringing these pieces together to build a morphological parser is problematic, for several reasons. First, rarely does one person combine the necessary knowledge of both the target language linguistics and the computational technology for building parsers. Second, parsers, once built, are limited by the life of the software used for implementation, a life that is often short. In addition, morphological parsers—and the grammatical resources that lie behind them—are currently hand-built, so creating them is extremely labor-intensive. Finally, we need a way of testing for accuracy: How do we know that the grammar on which the parser is based is an accurate description of the language?

CASL’s principled approach to building a parser addresses these problems in the following ways:

### Problem: finding the expertise

Writing software requires two kinds of expertise: knowledge of the problem to be solved, and knowledge of how to program software. For parsers, the problem-specific knowledge requires understanding the grammar of the target language. It might seem that finding someone who understands the grammar of any particular language is as easy as finding a fluent native speaker. Unfortunately, as generations of field linguists have discovered, a native speaker’s knowledge of a language is notoriously implicit—just recall the complexity of the unconscious rules for forming plurals and past tenses in English! Converting that knowledge into explicit rules is no simple task.

What about using extant written resources? During the initial stages of research into our first target language, we were surprised to discover that no thorough and reliable descriptive grammar of modern colloquial Bangla existed, despite over two hundred million native speakers (it is one of the ten largest languages in the world). Instead, we had to rely on descriptions of Bangla morphology from a variety of sources, including grammars of

<sup>1</sup> Dual grammars are also solutions to other problems, discussed at the end of this article

varying quality, journal articles, and dissertations. The descriptions in these sources were sometimes vague, did not always agree, and even contained a few gaps.

The difficulties we encountered in understanding grammatical descriptions, reconciling different grammatical accounts, and filling in gaps in coverage underline that we could not have simply picked up an existing grammar and written our formal grammar. For languages with any degree of inflectional complexity, the complexities prevent such a simple approach. Clearly creating a morphological parser requires linguistic as well as computational expertise. But finding one person who combines a detailed, sophisticated understanding of a given language's grammar with the computational expertise necessary to write a formal grammar is difficult, particularly with languages for which we do not already have large amounts of resources—precisely the languages that need tools like morphological parsers.

#### **Solution: collaborative grammar writing**

Combining a formal grammar with the descriptive grammar gives us an answer to the problem of ambiguity in traditional grammars: a formal grammar is neither ambiguous nor vague. In our methodology, one member of our team works as a descriptive grammarian, collecting grammars and other resources on the language of interest. The first step is to research existing descriptive and pedagogical grammars. Because some of these sources turn out to be relatively useless, we simultaneously search for potential consultants to fill in gaps in coverage, clarify ambiguities, proofread examples, and check the accuracy of our grammatical description, as well as help annotate text for test purposes.

The second step is determining how to represent the words of the language. For many languages, representation is not an issue, as existing orthographies are adequate. For the languages of CASL's project, however, representation has been and continues to be an issue. Urdu, for example, uses a right-to-left writing system (related to that used for Arabic), which makes writing and glossing examples difficult. In addition, short vowels are not written in the Urdu writing system, so a non-fluent speaker has difficulty knowing whether the vowel following a consonant is *a*, *i*, or *u*. Jst pctr hw dffclt rdng wtht vwls cn b whn y dn't knw th lngg!

The Bangla script presents its own orthographic peculiarities; for example, one of its vowels is not represented when following a consonant. Called the *inherent vowel*, it is simply understood as being present—*except when it's not*. Consonants *not* pronounced with the inherent vowel after them are written exactly the same as consonants pronounced *with* it, so the words *kon* (*which?*) and *kono* (*any*) are spelled the same in the Bangla script, with no explicit representation of the final *o* in *kono*. Knowing whether to pronounce such consonants with or without the inherent vowel requires familiarity with the language. As a result, we have chosen to write our Urdu and Bangla examples using both the native script and a transcription, that is, a system that writes all the necessary vowels.

After gathering resources and addressing logistical questions of script and transcription, we then begin writing our descriptive grammar. This process involves understanding the grammars we have collected, describing in our terms their collected wisdom, augmenting where necessary by help from our consultants, and ultimately vetting for accuracy—again assisted by consultants. Each grammar contains a chapter on the phonology and writing system of the language and chapters for the various parts of speech. Each chapter describes the inflectional affixes each part of speech takes and how the resulting inflected forms define the paradigms. The way these forms are used in sentences is also described, with illustrative examples.

The third step in the grammar-writing process is writing our formal grammar, using the descriptive grammar as a guide. In theory, this means simply taking the description condensed by the descriptive grammar writer from multiple source grammars, and turning it into a formal description in our formal XML language (see below for more on XML). In practice, the nature of formal grammars is that no matter how careful the descriptive grammarian is, the formal grammar brings to light gaps and ambiguities. When ambiguities surface, the descriptive grammar is clarified, either by referring back to the source grammars or by asking our consultants. (It is worth noting that when we created a formal grammar for Bangla, our consultant frequently commented that no one had asked these particular questions before. We take this to be in part a result of using computers, which, unlike humans, tolerate no ambiguity; but it also underscores the state of investigation of these languages.)

In summary, our division of labor, together with simultaneously developing the two kinds of grammar using our computational tools and incorporating immediate feedback, has made possible a much better result than writing the descriptive grammar, and then writing the formal grammar.

#### Problem: obsolescence

Unfortunately, software does not last forever. The development of computer-processable morphological grammars is often tied to the programming language of a particular morphological parsing engine or to a general purpose computer programming language. If the parsing engine were guaranteed to be around forever, or if there were an agreed-on descriptive language that all such parsers would use, this might not be problematic. But neither of these conditions is true. In fact, we estimate the average lifetime for language-based computational tools to be five or 10 years. In the past 25 years, at least a dozen mutually incompatible morphological parsing tools have been created, each with its own programming language.

Of course, there is little doubt that future parsing engines will be improvements on current parsing engines. We do not need to build parsing engines that will continue to be used decades from now. Rather, the language-specific information that goes into a parser—the grammar—should be written in such a way that it can be easily reused with future parsing engines.

One of our motivations for wanting to build parsing tools with a longer life is the part they play in the enterprise of linguistics, a major goal of which is the documentation and description of human languages. Parsing tools constitute a description of a language in two senses: first, the grammar that the parser uses is a formal description of the language's morphology; and second, the parser can be used to analyze language texts and to produce paradigms.<sup>j</sup> That is, a parser is an active description, not a static one.

While scholars of today can read with understanding grammars and corpora penned thousands of years ago, the use of digital technologies

such as parsers means that archived language data can become unusable much more quickly than printed descriptions.<sup>k</sup> If we wrote formal grammars using the programming languages understood by today's parsing engines, the grammars would soon need to be rewritten. Thus, although a parser constitutes a valuable description of a language, it is—until now—an ephemeral description.

#### Solution: a generic way to write computer-readable grammars

We expect that English will be understood for a long time, just as Latin is still understood by experts today. So our descriptive grammars, written in English, should be understandable for centuries, perhaps even millennia from now. Not so our formal grammars. Trying to write formal grammars in English, however, is not a solution due to the ambiguity of English and other natural languages. What we need is to define the formal grammar such that it can be unambiguously translated into the programming language of both today's parsing engines and future engines. We can't know what those future engines will be, but we can assume that they will be at least as capable as today's engines.

We have, therefore, chosen to write our formal grammars using the objects that linguists have discovered: prefixes, suffixes, infixes, phonemes, phonological rules, and so forth. These objects are described in our formal grammars using XML, that is, using descriptive tags to indicate what each linguistic object is. For example, a simplified representation of the English *-ing* suffix might look like this:

```
<suffix>
  <form>-ing</form>
  <gloss>-PresentParticiple</gloss>
</suffix>
```

The tags are named in ways that make sense to linguists, and should make sense even if XML itself becomes obsolete some day. Moreover, we embed these formal grammar objects directly inside the descriptive grammar, so that we provide future readers of our formal grammar even more information about the meaning of these objects. For example, the formal grammar description of the

<sup>j</sup> Technically, a parser can analyze inflected words, and a generator can produce inflected words, i.e., a paradigm

Today, these two functions are usually combined in a single tool, called a transducer

<sup>k</sup> See Bird and Simons 2003



present participle suffix would be embedded into the descriptive grammar next to a human-readable description of what a present participle is and how it is used, including examples of its use.<sup>1</sup>

In order to be used as a parser, the XML structures of our grammars need to be extracted from the descriptive grammar and converted into the programming language of a parsing engine. Although this conversion could be done by hand, we have built a program that does this automatically for our current parsing engine. This program can easily be re-targeted to future parsing engines. In fact, we predict that future parsing engines will incorporate more knowledge of linguistics—for example, they will probably have more built-in capabilities for handling the complexities of morphology that we described above. This improved capability should make it easier to convert our linguistic descriptions into the programming languages of those future parsers.

While the use of XML and a converter program helps us avoid the Scylla of writing grammars in programming languages that will soon become obsolete, we also need to avoid the Charybdis of linguistic theories that will likewise become obsolete. The future of linguistic theories is just as hard to see as the future of parsing engines and programming languages. We therefore need to ensure that our XML-based formalism is simultaneously understandable, simple, and sufficiently powerful.

Fortunately, over the last hundred years, linguists have investigated a wide variety of languages—wide enough that we can say with a degree of confidence that we know the range of operations that languages carry out in morphology and phonology (if not in areas such as syntax and semantics). Linguistic mechanisms powerful enough to handle nearly all such operations were developed fifty years ago; much linguistic research since then has been aimed at issues irrelevant to parsing. Therefore, we can write our descriptions using the older linguistic mechanisms, informed in a few cases by newer discoveries. And in fact, descriptions using these older formalisms are, in many cases, easier to understand than they would

## Endangered languages

Some languages have hundreds of millions of speakers, such as English or Mandarin Chinese. But most of the seven thousand languages in the world today are much smaller—some with only a few speakers. Most of these small languages (and even some with thousands of speakers today) will probably die out by the end of this century. Such languages are referred to as “Endangered Languages”; they can be found around the world, but are especially common in Africa, South America, and parts of Asia.

Language death is of course not new; Latin more or less died with the end of the Roman Empire, although the modern Romance languages are descended from it. Today we know Latin from documents written in Latin, and from descriptions written by Latin grammarians. Linguists today are trying to document endangered languages by writing down stories in those languages, and by describing the languages with dictionaries and grammars.

be if they were written using the most up-to-date theoretical fashions.

### Problem: labor-intensive nature of grammar production

Building a grammar is difficult, time-consuming, and expensive. In fact, out of the thousands of languages of the world, linguists have described the grammars of less than a thousand, and many of these descriptions are far from complete—precisely because grammars are difficult to write.

<sup>1</sup> Some readers may recognize that this is an implementation of a technique known among computer scientists as Literate Programming

### **Solution: efficient grammar production and portability to other languages**

The division of labor we have created among the descriptive grammar writer, the consultants, and the formal grammar writer could be duplicated in quite different situations. Our methodology and templates are adaptable by other teams of traditional and computational linguists.

For example, a potential grammar-writing team would include a grammar writer contracted to write the descriptive grammar, together with an in-house computational linguist who would attempt to write a corresponding formal grammar from the descriptive grammar and then test that formal grammar by converting it into a parser. By staging this process in logical sections (for example, the case marking of nouns might be one module), the descriptive grammarian can be evaluated throughout the process and given feedback on how to improve the description. Quality could thus be controlled not only in terms of whether the description is understandable, but whether it is objectively correct. In sum, we would view our work as being a segue into ways of improving the out-sourcing of grammar development.

Grammars might also be built by adapting a grammar of one language to serve another closely related language. Most languages of the world are related to some set of other languages; Spanish, Italian, Portuguese, and French, for example, are all descended from Vulgar Latin and therefore exhibit similarities in vocabulary and grammar. Other groups of more or less closely related languages are the Turkic languages (Turkish and Uzbek, among others); Indo-Aryan languages such as Hindi, Urdu, Bangla; Dravidian languages like Tamil and Malayalam; still other groups include most Philippine languages, or the Bantu languages of mid-Africa. Within each of these groups, it is often the case that some languages have been described more completely and correctly than others. A promising approach would be to write grammars for the well-described languages, and then to adapt the formal grammars (and perhaps the dictionaries) to the less well-described languages. This process could be done by hand, but automatic adaptation is another possibility.

Automatic adaptation has not been a subject of great interest among researchers; most methods for automatic morphology learning developed thus

far approach learning without reference to existing grammars of related languages. This limitation is at least in part due to the fact that until now there have been no standard formal ways of writing grammars. Without such standards, there are no formal grammars from which adaptation could be done. We hope to change that through this project. The formal grammar standards developed here, and the actual formal grammars being developed, mean that there is now a concrete basis for automatic grammar adaptation.

### **Problem: how do we test the accuracy of the grammar?**

One of the authors was involved twenty years ago in editing other people's grammars. He was continually frustrated because he could usually tell whether a grammatical description made sense, but it was difficult, if not impossible, to tell whether the description was correct—that is, whether it actually covered the language data, including the examples in the grammar. It was too hard to work through all the grammar rules by hand for each example, and at that time the technology to test the grammar on the computer was not readily available.

### **Solution: incorporating verifiability**

Our project makes technology for testing the grammar available—and moreover, makes it usable during the grammar writing, rather than later. Therefore, we can verify that our grammar works using the very examples and paradigm charts that it contains, as well as any additional electronic texts. We can demonstrate that our grammars make sense and that they work on real language data.

### **Conclusion**

Linguists working with a language in which they lack fluency often face a daily challenge with dictionary lookup. Using our principled approach of building a morphological parser based on parallel grammars, we are creating tools that will help linguists while developing methods that will make such linguistic tool building more efficient and cost-effective. The payoffs include not only more efficient and capable linguists, but also a methodology that can be used to build grammars for other languages in the future.

There are many other potential benefits from our work. For instance, the example sentences and paradigm charts contained in our descriptive grammars may help language analysts now,

## Dialectal variation


All languages exhibit dialectal variation, most often based on geographical or social differences. For example, English speakers use a wide variety of second person plural pronouns: *you* (standard American and British English), *y'all* (Southern American English, African American vernacular English, some Western American varieties, and South Asian English), *you-uns* (Western Pennsylvania and Appalachians), *you guys* (colloquial non-Southern American English, Australian English), *yous(e)* (Irish English and mid-Atlantic American English), and so on. For an even wider range of dialectal variation within English, listen to some of the sound files on the International Dialects of English Archive (<http://web.ku.edu/idea/>).

Similarly, some words in the languages of our project vary according to region, and we have tried to address this phenomenon in our grammars by listing alternate forms when we hear of them from our informants or in our written sources.

although they were intended to clarify usage issues for future computational linguists. In this regard, using XML formats allows us to extract the relevant portions of our grammars and turn them into “help” documents for language analysts. For example, we could easily extract an explanation of the use of the “locative-instrumental” case in Bangla, together with examples of its use.

Furthermore, a morphological parser is a necessary tool for machine translation of languages with complex morphology. While higher quality machine translation requires resources available for only a few languages, there are applications that do not require such high quality—indeed, word-for-word “translation” may be adequate for such tasks as sifting through a large number of documents to find the few that may be worth translating by hand.

Finally, our methodology can be used to document languages in an enduring way. There are more than seven thousand human languages in the world, and most of them have little to no documentation. Important knowledge of how human languages work is being lost as many of

these languages die out with the last speakers (whose children are turning to other languages). A goal of many linguists is to preserve this knowledge. Grammars of endangered languages could be written using this process, so that a hundred or a thousand years from now texts in the language can be studied and the grammatical system reconstructed more reliably than is possible with purely descriptive grammars. 



## References

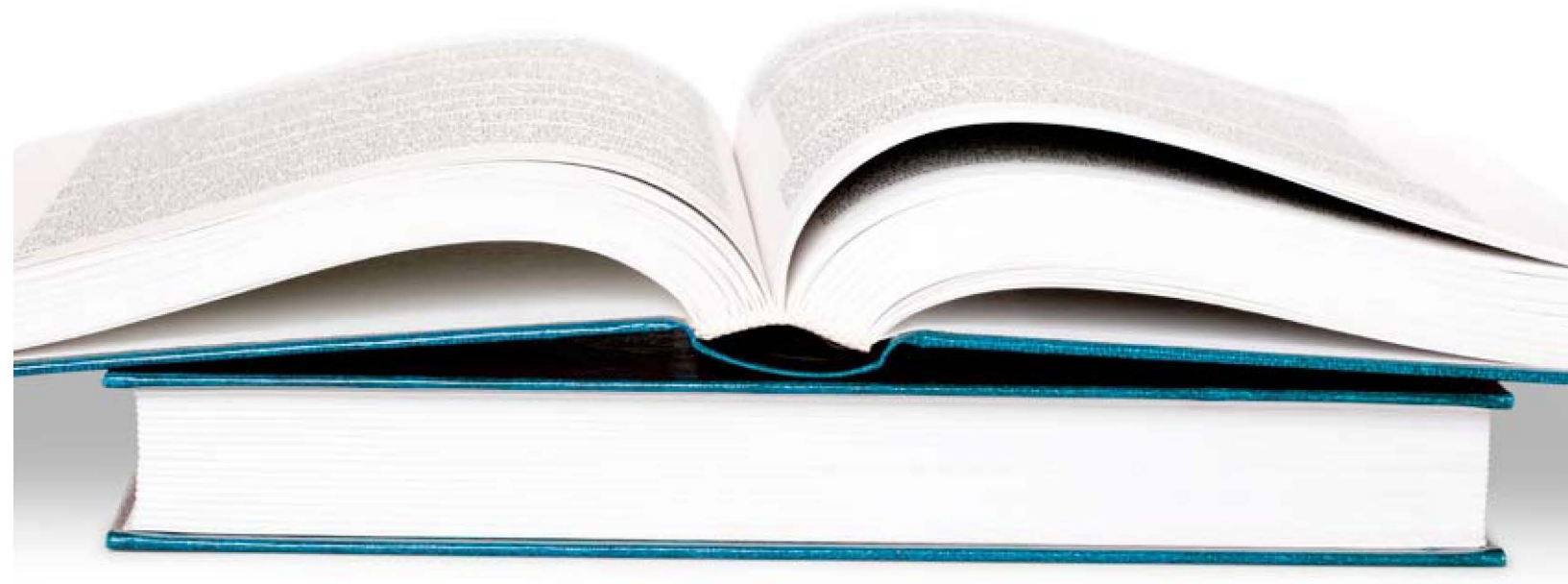
Bird S, Simons G. Seven dimensions of portability for language documentation and description. *Language*. 2003;79:557-582.

Knuth DE. *Literate Programming: CSLI Lecture Notes*. Stanford (CA): Center for the Study of Language and Information; 1992.

Maxwell M, David A. Interoperable grammars. *Proceedings from First International Conference on Global Interoperability for Language Resources (ICGL 2008)*; Hong Kong; 2008.

Maxwell M, David A. Joint grammar development by linguists and computer scientists. *Proceedings from Workshop on NLP for Less Privileged Languages, Third International Joint Conference on Natural Language Processing*; Hyderabad, India; 2008.

Poser WJ. Making Athabaskan dictionaries usable. In: Holton G, editor. *Proceedings of the Athabaskan Languages Conference*; Fairbanks (AK). Alaska Native Language Center, University of Alaska; 2002. p. 136-147.



# A Talent for Language

What makes a sports champion? Is Tiger Woods—the only golfer to hold all four professional major championships at the same time—the sport’s greatest because he was born to golf, because he trained hard every day for years, or because he had the help of an expert coach? Most people would claim that athletes need all three of these factors for success: *talent, hard work, and good coaching*. Is the same true for successful linguists? What makes a linguistic champion? Are hard work and good coaching sufficient, or is a special talent for language necessary? Researchers at the University of Maryland Center for Advanced Study of Language (CASL) are trying to answer these questions through a set of related studies on *foreign language aptitude*.

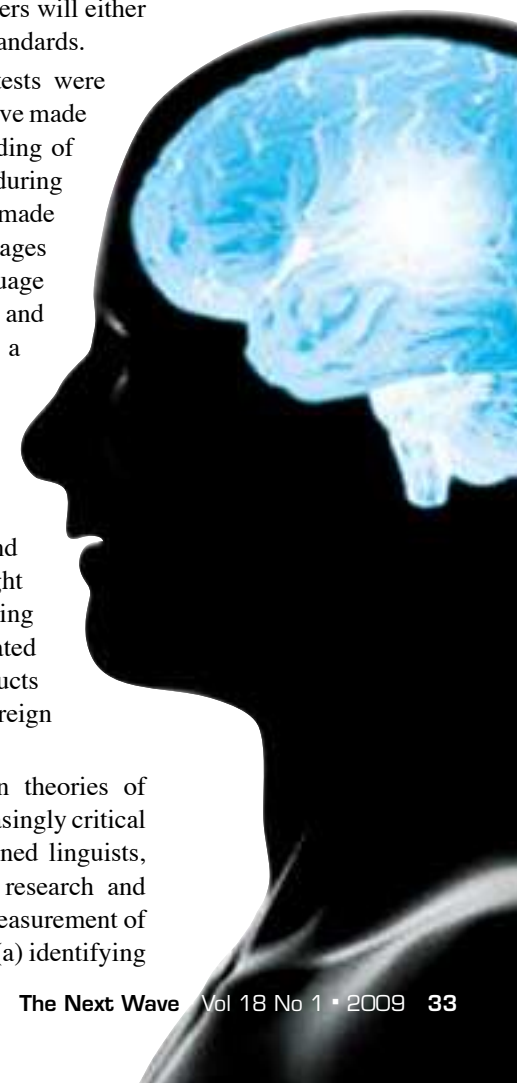
*Aptitude* refers to a person’s inherent capability or underlying talent. If a classroom has ten students, all with the motivation to work hard and the same opportunities to benefit from good teaching or coaching, students’ results will still vary based upon aptitude. John B. Carroll, a distinguished psychologist and co-developer of the Modern Language Aptitude Test (MLAT), defined foreign language aptitude as an “...individual’s initial state of readiness and capacity for learning a foreign language, and probable degree of facility in doing so” [1]. To study aptitude scientifically, researchers must determine how to rigorously define concepts such as *state of readiness* and *capacity for learning* and how to measure these abilities reliably.

Attempts to define foreign language aptitude and determine its subcomponents had begun by the 1920s in the United States [1]. During subsequent decades, several tests were developed, with the support of the military, for purposes of personnel selection and placement into foreign language training. These include the Army Language Aptitude Test (ALAT), the Defense Language Aptitude Battery (DLAB), the Modern Language Aptitude

Test (MLAT), and the VORD [2]. (VORD requires test takers to apply rules to artificial language segments presented in context. *VORD* is the word for word in the artificial language, which is based on Turkish, a language typologically different from Western European languages.) Currently, the DLAB and the MLAT are the most widely used tests for US government (USG) purposes. Although these tests are useful for establishing cutoff scores for basic language study and for supporting hiring decisions, they may not predict with enough accuracy which learners will succeed and which learners will either drop out or fail to meet proficiency standards.

During the years since these tests were developed, cognitive psychologists have made substantial progress in the understanding of human memory and learning. Also, during this period, significant advances were made in understanding how foreign languages are learned. The field of Second Language Acquisition (SLA) was established, and SLA research findings prompted a paradigm shift in language teaching methodology. Contemporary approaches emphasize developing communication rather than the study of language as object and, thus, incorporate interactive and experiential learning tasks. In light of these scientific advances, existing aptitude tests have become outdated in terms of their underlying constructs and their congruence with how foreign languages are learned.

Based upon these advances in theories of language learning, and upon the increasingly critical need for a large number of well-trained linguists, the USG has asked CASL to lead research and development efforts to improve the measurement of foreign language aptitude in terms of (a) identifying



## A New Aptitude-Screening Test

The Pre-DLAB is a short screening test that predicts who is likely to attain the DLIFLC entry cut-off score on the DLAB. The Pre-DLAB can be administered in less than 30 minutes, in any setting, to thousands of people (e.g., each year over 500,000 people take enlistment tests at Military Entrance Processing Stations). Since the DLAB takes 90 minutes, it will be more efficient if only those who pass the Pre-DLAB screening test move on to take the DLAB during the recruitment process.

Using the test specifications that were developed during previous research, new items were written at Second Language Testing, Inc. for certain sections of DLAB, taking care to match the new items to the aptitude constructs underlying the original test. CASL then conducted a study with 500 university volunteers who took both the DLAB and the prototype Pre-DLAB. Findings showed that the initial form of the Pre-DLAB accurately predicts a cut score of 95 on the DLAB. Moreover, the Pre-DLAB test does not require audio equipment, thus it is adaptable to administration in a wide range of settings. This pre-testing greatly increases the military's or USG's ability to identify talented language learners.

potential for language expertise, (b) mitigating attrition from language training, and (c) casting the net more widely to increase the number of individuals tested for language aptitude. In response, CASL has undertaken research on language aptitude:

- To predict the ability to reach high-level proficiency in a foreign language through the creation of the High-Level-Language Aptitude Battery (Hi-LAB), an innovative test that measures the ceiling on potential and predicts the maximum level of language-learning success
- To update aptitude testing in and mitigate attrition from basic language programs at the Defense Language Institute Foreign Language Center (DLIFLC) by updating DLAB to DLAB 2, comprising measures of aptitude, personality, and motivation
- To manage the tremendous increase in the amount of aptitude testing that must be done to identify candidates for foreign language training at DLIFLC by developing and validating Pre-DLAB, a short screening test that can be easily administered to thousands of military recruits

### Hi-LAB – finding language experts

Existing aptitude tests were designed to predict early rate of learning and successful attainment of intermediate level language proficiency. Such tests can be very useful when managers must decide who should be selected for basic language training. However, many learners perform well in early training, but never reach the higher levels of proficiency necessary for many jobs. The scientific explanation is known as the *critical period effect*. Time, money, and effort could be saved if testing were

able to identify which beginning or intermediate level learners would be most likely to overcome the critical period effect and achieve high-level competence in a foreign language, given extensive training.

Hi-LAB, a new test currently under development at CASL, is designed to identify individuals with the capacity to reach advanced levels of foreign language proficiency. High-level language aptitude is defined as a measurable *ceiling* on language-learning ability, holding equal all other factors such as motivation, stable personality characteristics, and opportunities for instruction or immersion. Hi-LAB constructs (Table 1) were motivated by theories of learning and memory

Constructs		Brief Definitions
<b>MEMORY</b>		
Working Memory	Short-term Memory Capacity	the capacity to process and store input with active trade-offs among these components: the small amount of information that can be kept in an accessible state in order to be used in ongoing mental tasks
	Executive Capacity & Control	a set of processes that, collectively, regulate and direct attention and control voluntary processing
Long-term Memory	Rote Memory	explicit, intentional long-term memory storage that results from rehearsal
<b>ACUITY</b>		
	Perceptual Acuity	an individual's capacity to detect difficult-to-perceive auditory or visual information
<b>SPEED</b>		
	Processing Speed	speed of an individual's perceptual, motor, or decision responses
<b>PRIMABILITY</b>		
	Priming	the extent to which prior experience facilitates subsequent processing
<b>INDUCTION</b>		
	Implicit Induction	the process of reasoning from the specific to the general, i.e., noticing similarities among several instances and drawing a generalization based on these similarities acquiring the statistical patterns contained within complex input without the learner's conscious awareness
	Explicit Induction	acquiring the patterns in input through conscious awareness and reasoning
<b>PRAGMATIC SENSITIVITY</b>		
	In Research & Development	the ability to hypothesize connections between context and use; registering and tracking salient context cues; detecting miscommunication
<b>FLUENCY</b>		
	In Research & Development	the ability automatically to plan and articulate speech

Table 1: Hi-LAB Components



## Age and Foreign Language Learning

Studies reveal a tight correlation between age of first exposure to a foreign language and overall success to Level 4 in that language throughout the neurological critical period [3]. At around puberty, the age-success correlation falls off sharply, as shown schematically in Figure 1. Nonetheless, some individuals do appear to attain near-native expertise [4]. What characterizes the individuals who become experts in a foreign language as adults and, more importantly, can they be identified at the outset of learning?

derived from research in cognitive psychology and by theories of second language acquisition. Multiple measures of each construct, developed or adapted by CASL, are undergoing usability, reliability, and validity testing with government, military, and university populations. During this iterative testing process, factor analytic procedures are used to identify the most useful measures for inclusion in the operational version of the test [5].

An important innovation in Hi-LAB is that abilities are measured behaviorally, by computerized testing. This direct-measurement methodology provides several advantages over more traditional testing formats, such as paper and pencil tests. Computer-delivery ensures that both accuracy and reaction time (in milliseconds) are available for each cognitive or perceptual task included in the battery. Reaction time is a good indicator of automaticity, a hallmark of foreign language expertise.

Each of the Hi-LAB constructs is hypothesized to be important for moving from intermediate to professional or high levels of foreign language proficiency. Thus, the test is designed to measure “what is left to learn” at this relatively advanced stage of language learning. Some constructs, such as perceptual acuity, are clearly important for the beginning stages of learning as well. For example, beginning learners must learn to hear and produce the sounds of a new language (*perceptual acuity*). Other Hi-LAB constructs may reflect aspects of aptitude that are not predictive of early language success, but that are critical for moving beyond intermediate stages. One example of this type of construct is *pragmatic sensitivity*, which reflects a learner’s ability to learn and use the aspects of language that depend on context. For example, a person high in pragmatic sensitivity would be more likely to accurately learn the correct forms of address for persons of differing social ranks or to notice nonverbal signals indicating that the listener has misunderstood something.

Another construct of this type is *implicit induction*, which refers to an individual’s ability to

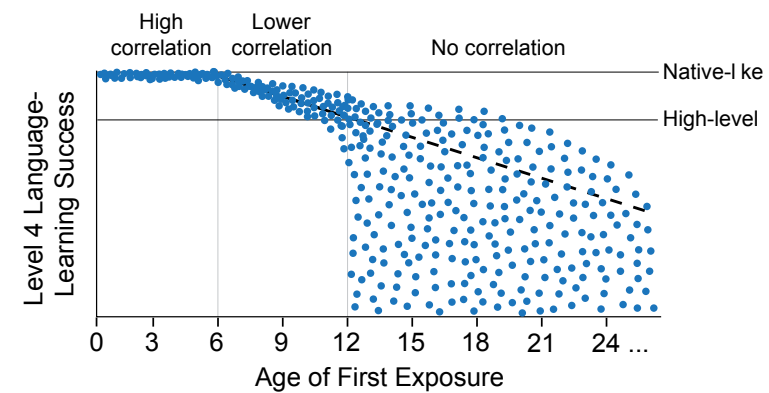
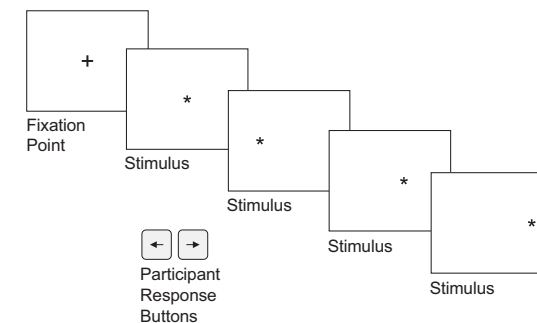


Figure 1: Schematic correlation scatter plot for age/level 4 language learning success



The **Serial Reaction Time** task asks examinees to respond to the location of an asterisk on the computer screen. The asterisk can appear in one of a number of locations, each with a corresponding response button. Unknown to the examinee, the locations follow a designated 10-trial pattern, in which one location was never followed by the same location more than once. This means that to know the location of the next asterisk, one would need to know the entire 10-trial string and current serial position in the string. After a number of string repetitions, the order is switched to a randomized order which does not follow the pattern before returning to the repeating strings. Scoring of the Serial Reaction Time task depends on the average reaction times during the sequence repetition condition versus the average reaction times during the random condition. The longer reaction times after switching to a random sequence are an indication of the facilitation the examinee experiences when they have implicitly learned, or induced, the sequence of the repeating trials.

Figure 2: The Serial Reaction Time task measure of implicit induction

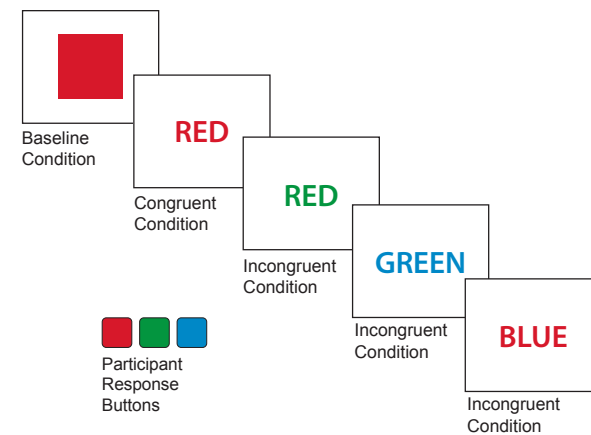
learn the statistical regularities in a set of complex data, without conscious awareness or explicit reasoning. Research has shown, for example, that human infants can use this sort of statistical learning to learn where word boundaries are in a string of incoming syllables produced by adults [6]. By deriving these word boundaries from statistical regularities in the language they hear, babies are able to “bootstrap” the process of learning their first language. Psychological researchers have not yet resolved to what extent adults are able to make use of this type of statistical learning. It is possible, however, that individuals who maintain this ability in adulthood will be better able to learn complex, probabilistic aspects of a foreign language grammatical system that are necessary for “native-like” performance. The serial reaction time task, shown in Figure 2, measures implicit induction.

### Advances in memory

Hi-LAB is the first foreign language aptitude test to incorporate important advances in the

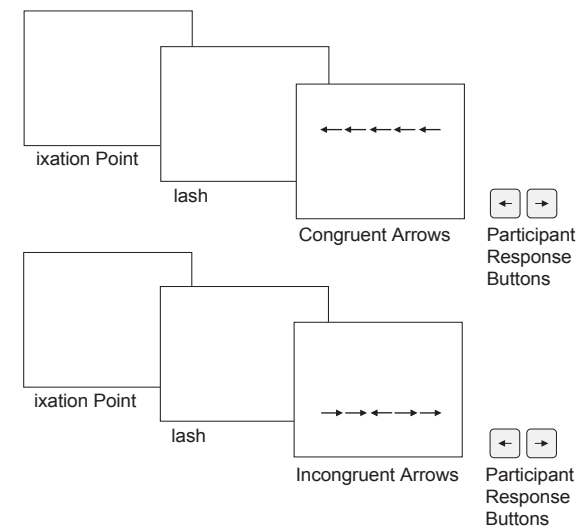
understanding of the human memory system. Memory comprises several sub-systems. Previous aptitude tests primarily measured long-term memory, or the ability to store information in a permanent form and recall it after a delay. Like these previous tests, Hi-LAB contains measures of long-term memory, but unlike these older tests, Hi-LAB also includes measures of working memory (WM) with its component systems, *short-term memory* and *executive capacity and control*. The *short-term memory* (STM) system allows a person to hold information in an accessible state for a few seconds, such as when rehearsing a phone number. Hi-LAB specifically taps one STM sub-component, namely verbal-acoustic STM, which aids in the rehearsal or maintenance of unfamiliar words, such as vocabulary in a foreign language [7].

The *executive capacity and control* system controls the focus of attention and includes three separable sub-constructs [8, 9]—*updating*, *inhibition*, and *task-switching*. *Updating* refers to



**The Standard Stroop Task** The original Stroop task is a measure of an individual's ability to inhibit the highly automatic skill of reading. Examinees view a series of color words (red, green, or blue) and color swatches (also red, green, or blue). The task is to name the “ink” color in which the color word or the color swatch is printed. For some of the color words, the color in which the word is displayed matches the word (example “R” printed in red ink). This is referred to as the “congruent condition.” For other color words, the ink color does not match the word (example “R” printed in blue ink). This is referred to as the incongruent condition. When the colors are presented as swatches, this is the “baseline condition.” The degree of slowing produced by incongruent stimuli, rather than baseline or congruent stimuli, is measured and can be reported either absolutely (number of milliseconds slower for incongruent stimuli) or relatively (proportion of average response time slower for incongruent stimuli). Individuals who are faster to respond to incongruent stimuli are more effective at inhibiting automatic processes that would otherwise interfere.

Figure 3: The Stroop task measure of inhibition [13, 14]



**The Attention Network Test (ANT)** is a measure that combines cued reaction time and the flanker task. The ANT requires participants to determine whether a centrally presented arrow points to the left or to the right. The arrow appears above or below a fixation point, following a brief flash of asterisks that may or may not appear where the arrow is displayed. The central arrow may be accompanied by flanker arrows to the right and left. When flanker arrows are present, they appear under one of two conditions: *congruent*, where the flanker arrows point in the same direction as the target arrow, and *incongruent*, where the flanker arrows point in the opposite direction. Efficiency of attentional networks is assessed by measuring how response times are influenced by alerting cues, spatial cues, and flankers.

Figure 4: The Flanker task measure of inhibition [15, 16, 17]

the process of refreshing the contents of working memory with new, more relevant information [10]. *Inhibition* is the ability to ignore a dominant or automatic response when necessary, a skill that may be particularly necessary for effective bilingual functioning, which involves inhibiting the strong first language in favor of the foreign language [11, 12]. See Figures 3 and 4 for test examples.

Finally, *task-switching*, the ability to shift between multiple tasks, operations, or mental sets [18], is hypothesized to reflect an aspect of cognitive control that is critical for efficient bilingual lexical selection and for advanced language tasks such as translation or switching between two languages [11, 19, 20]. The color-shape task shown in Figure 5 measures the mental cost of switching from deciding between shapes to distinguishing colors.

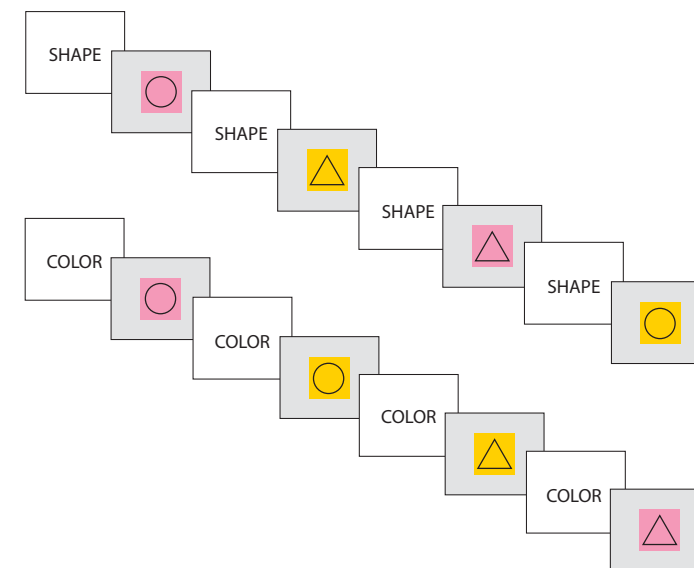
By incorporating measures of these sub-components of memory, along with the other constructs measured, Hi-LAB will provide the government with a more sensitive test of foreign language aptitude and one that is geared toward identifying language expertise. While Hi-LAB measures cognitive and perceptual measures of aptitude to predict high-level success, another CASL research and development effort, DLAB 2, incorporates additional, non-cognitive measures to make other predictions about foreign language learning.

### DLAB 2—revised Defense Language Aptitude Battery

Learning a foreign language is neither easy nor fast. To succeed, a learner must stick with the program for an extended period of time, push on when progress seems to have stalled, overcome setbacks, and work on maintenance once a goal has been reached. Gaining expertise in a foreign language requires that even the most talented learners commit years to study and practice. As a result, predicting which learners will be successful may require consideration not only of cognitive and perceptual abilities, but also non-cognitive factors such as motivation, interests, beliefs about learning, and differences in personality. CASL is currently examining these issues for the Defense Language Institute Foreign Language Center (DLIFLC) as part of the DLAB 2 project.

The goal of this project is to predict success in learning a foreign language within the intensive,

classroom-based, military environment at DLIFLC. Students who succeed in this setting may differ from those who would do well in a language immersion situation or in classes at a university. Success in the DLIFLC environment may depend on a number of attributes other than purely cognitive or perceptual aptitude. Recall the three prerequisites for success discussed initially—*talent*, *hard work*, and *good coaching*. While the Hi-LAB project focuses precisely on identifying the talent of individual learners, the DLAB 2 project aims to quantify potential for hard work. Thus, in addition to updating the existing cognitive and perceptual measures, this project incorporates a set of non-cognitive measures that may predict which learners will be most likely to persevere in an intensive program and achieve language learning success at DLIFLC. Although these are non-cognitive attributes, they may be included in a new version of the Defense Language



**Task-switching** task-switching refers to the process of shifting between multiple tasks, operations, or mental sets. There are significant individual differences in task-switching ability. Some individuals are markedly better at doing two or three things at once than others, controlling for the extent of prior practice and background knowledge relevant to the task situation. In this test, individuals see stimuli consisting of triangles or squares in yellow or pink blocks followed by a prompt. Depending on the prompt that is displayed, they must pay attention to shape and ignore color or the reverse, with no prior knowledge about when a switch in prompt will occur. Mean reaction times and accuracy are recorded for the single-task test blocks (shape and color) as well as for the mixed condition that contain switch and non-switch items in random order. Items that involve the same task as the previous item ("non-switch") can then be compared to those that immediately follow a task-switch ("switch"). In addition, non-switch items are compared to the baseline accuracy and reaction time information for each task when performed in the non-mixed condition.

Figure 5: Task-switching measure of attentional control [18]



Aptitude Battery, because they help to address such a test’s main goal—to predict which learners are most likely to succeed.

Like the original DLAB, the goal of the DLAB 2 test is to predict which learners will be able to learn quickly and reach minimum levels of foreign language proficiency to meet DLIFLC graduation requirements, given a set amount of training time. This time can vary from 25 weeks of full-time study for a language such as Spanish or French, up to 63 weeks of full-time study for a more difficult language such as Arabic or Mandarin Chinese. An important aspect for success in such an intense program is the learner’s ability to remain motivated to work week after week and to overcome the inherent difficulties and stresses involved in long-term language study. Because of these difficulties, student attrition is a major problem, and measures that predict who is likely to “stick with” the program may provide important additional information to the government. CASL researchers have identified and piloted a set of possible new measures and tested them. Each of the measures belongs to one of five major categories:

- Cognitive Abilities—such as general intelligence, memory, attention, and auditory perception
- Learning Orientation—such as preferred learning styles or activities and learning goals
- Personality—such as levels of conscientiousness, anxiety, and openness to experience
- Self Efficacy—such as beliefs in the ability to succeed and the ability to cope with setbacks
- Motivation—such as motivation to achieve, to learn, and to master difficult material

The new cognitive ability measures in DLAB 2 are a small subset of those measures in Hi-LAB and will update this aspect of the test. The addition of non-cognitive measures is expected to directly address course attrition. These measures recognize the fact that adults bring to language study an established set of preferences attitudes, learning strategies, and motivations. If those selected for language training at DLIFLC have what it takes to persevere, plus a talent for language learning, they can be expected to work hard and graduate with at least basic language proficiency [21,22].

### The future of aptitude research

There are many additional questions that can be asked about foreign language aptitude and a variety of new techniques that can be employed for aptitude research. One of the most promising areas for enhancing the measurement of aptitude is the use of techniques from *cognitive neuroscience*—a field of study that combines psychology with neuroscience through the use of modern technologies such as brain imaging. (See the article “Thinking Out of the Box” in this issue.) These techniques allow for visualizing individual differences in brain structure or measuring changes in brain state while a person completes test tasks. How might these techniques be used to enhance our understanding of foreign language aptitude? One example might be by using these technologies to provide a more sensitive measure of an individual’s ability to discriminate foreign language sounds.

We know from research on speech perception that infants are able to distinguish sound contrasts important for any of the world’s languages [21]. However, by the time an infant is one year old, his brain has “tuned in” to those contrasts necessary for the native language and learned to average

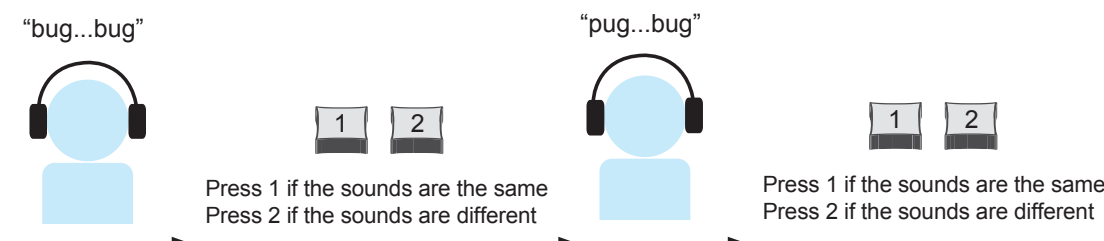
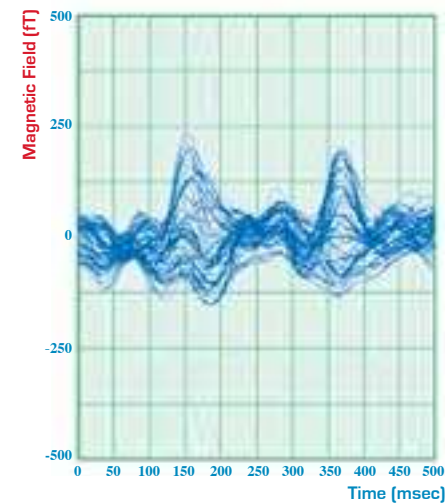


Figure 6: Phonemic discrimination task



**Figure 7: MEG measurement of stimuli**

Changes in the magnetic field surrounding the skull measured from onset of a stimulus and continuing for 500 ms. Each line represents the signal (in femtoTesla) recorded from one detector. Signals from multiple detectors are recorded over various areas of the scalp. Peaks in the signal represent purported sub-processes taking place during the brain's recognition of the input stimulus.

over contrasts that are not important for the native language [22]. Thus, older children and adults are often no longer able to discriminate contrasts that are not used in their native language, but that might be very important for a foreign language. Both the Hi-LAB and DLAB 2 projects examine the ability to hear these non-native contrasts by asking test-takers whether two sounds that are difficult to discriminate are “the same” or “different,” as shown in Figure 6.

The hypothesis is that some adults may retain the ability to distinguish these non-native contrasts even though most people have lost that ability. Such learners may truly have “an ear” for foreign languages. Perhaps an even more sensitive way to test this ability is by measuring changes in the brain's activity when a listener is presented with two difficult-to-discriminate sounds. This measurement can be accomplished using either electroencephalography (EEG), which measures changes in electrical conductance, or magnetoencephalography (MEG), which measures changes in the brain's magnetic field—both measured non-invasively by sensors at the scalp (Figure 7). Discrimination of sounds can then be measured using signals originating from the auditory cortex area of the brain, eliminating confounds with a learner's use of particular response strategies and guessing procedures, and providing a more sensitive measure of discrimination.

An additional future goal of aptitude research might be to match aptitude profiles, personalities, or learning preferences to particular types of instructional methods, thus allowing for the customization of language courses to particular learners. This sort of matching process is not yet widely used, but with increased understanding of the relationship between the variables currently under investigation, it may become an important

way to increase efficiency of language training. Furthermore, technology may assist instructors to provide individualized learning plans and activities that assist learners with their weaknesses, build upon their strengths, and help maintain motivation and interest during long-term study by accommodating learners' personalities and preferred learning activities. This type of instructional design would allow the findings from aptitude research to inform the third of our three keys for success—*good coaching*.

Finally, future research might address the issue of *differential aptitude*, allowing for a better matching of student to foreign language. *Differential aptitude* refers to the hypothesis that learners' patterns of strengths and weaknesses, identified through aptitude testing, might make some students more likely to succeed with special challenges of particular languages. For example, some languages require mastery of a new writing system, and these systems may vary in their complexity. Other languages may have a particularly complex grammatical system with the need to learn and correctly use a large number of case endings or verb conjugations, or may require the learner to acquire a particularly difficult sound system, such as a series of tonal contrasts. Imagine the situation shown in Figure 8, where a learner has relatively high general foreign language aptitude and must now be assigned to study either Arabic or Mandarin Chinese. Both are very difficult languages for English speakers to learn, but they present very different challenges. Learner 1, with a high perceptual acuity, might be better suited to learn Mandarin Chinese, whereas Learner 2, with an even aptitude profile, could be assigned to learn Arabic.

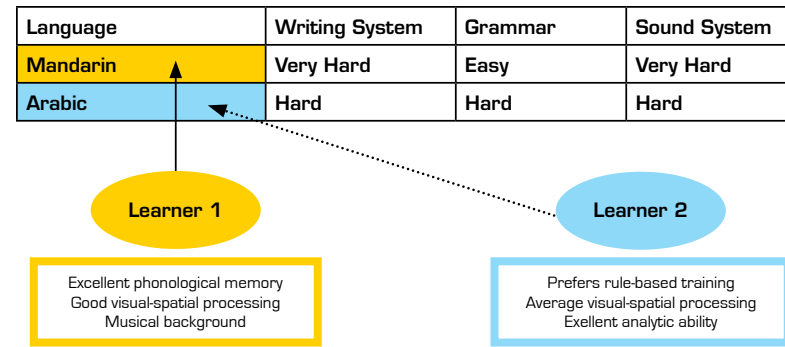


Figure 8:  
Optimizing language training

### Conclusion

Expertise in critical languages is in high demand. Despite the intense pressure, the demand is difficult to meet. By investigating the components of language talent and ways to uncover them through precise and sensitive measurement, CASL researchers are contributing to meeting the need. Eventually, through optimizing foreign language training in terms of rapid screening, precise selection of individuals likely to succeed, and placement into language programs suited to personal talents, it can be expected that our national language readiness will be greatly increased. 🇺🇸



## Notes

- [1] Carroll JB. Twenty-five years of research on foreign language aptitude. In: Diller KC, editor. *Individual differences and universals in language learning aptitude*. Rowley (MA): Newbury House; 1981.
- [2] For short descriptions of these tests, see: Bowles A, Bauman J, Winn M. An aptitude for tone? Music, memory and more. University of Maryland Center for Advanced Study of Language; 2007. 112-M.1.
- [3] For more detailed discussion, see: DeKeyser R, Larson-Hall J. What does the critical period really mean? In: Kroll JF, De Groot AMB, editors. *Handbook of Bilingualism: Psycholinguistic Approaches*. Oxford (UK): Oxford University Press; 2005.
- Johnson JS, Newport EL. Critical period effects in second language learning: the influence of maturational state on the acquisition of English as a second language. *Cognitive Psychology*. 1989;21(1):60-99.
- Long MH. Maturational constraints on language development. *Studies in Second Language Acquisition*. 1990;12:251-285.
- [4] DeKeyser R. The robustness of critical period effects in second language acquisition. *Studies in Second Language Acquisition*. 2000;22:499-533.
- [5] For descriptions of the measures, see: Doughty CJ, Bunting MF, Campbell S, Bowles AR. Internal consistency and test-retest reliability of the Hi-LAB measures. College Park: University of Maryland, Center for Advanced Study of Language; 2007.
- [6] Saffran JR, Aslin RN, Newport EL. Statistical learning by 8-month-old infants. *Science*. 1996;274:1926-1928.
- [7] Cowan N. *Attention and memory: An integrated framework*. New York: Oxford University Press; 1995. (Oxford Psychology Series, No. 26).

- [8] Baddeley AD, Hitch GJ. Working memory. In: Bower GA, editor. *The psychology of learning and motivation*. New York (NY): Academic Press; 1974. p. 47-89.
- [9] Baddeley A. Working memory and language: An overview. *Journal of Communication Disorders*. 2003;36:189-208.
- [10] Morris N, Jones DM. Habituation to irrelevant speech: Effects on a visual short-term memory task. *Perception & Psychophysics*. 1990;47:291-297.
- [11] Abutalebi J, Annoni J-M, Zimine I, Pegna AJ, Seghier ML, Lee-Jahnke H, et al. Language control and lexical competition in bilinguals: an event-related fMRI study. *Cerebral Cortex*. 2008;18:1496-1505.
- [12] Kroll JF, Bobb SC, Misra M, Guo T. Language selection in bilingual speech: Evidence for inhibitory processes. *Acta Psychologica*. 2008;128:416-430.
- [13] Stroop JR. Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*. 1935;12:643-662.
- [14] MacLeod CM. The Stroop task: The "gold standard" of attentional measures. *Journal of Experimental Psychology: General*. 1992;121:12-14.
- [15] Fan J, McCandliss BD, Sommer T, Raz A, Posner MI. Testing the efficiency and independence of attentional networks. *Journal of Cognitive Neuroscience*. 2002;14:340-347.
- [16] Posner MI. Orienting of attention. *Quarterly Journal of Experimental Psychology*. 1980;32:3-25.
- [17] Eriksen BA, Eriksen CW. Effects of noise letters upon the identification of target letter in a nonsearch task. *Perception & Psychophysics*. 1974;16:143-149.
- [18] Monsell S, Driver JS, editors. *Attention and Performance XVIII: Control of cognitive processes*. Cambridge (MA): MIT Press; 2000.
- [19] Hernandez AE, Martinez A, Kohnert K. In search of the language switch: An fMRI study of picture naming in Spanish-English bilinguals. *Brain and Language*. 2000;73:421-431.
- [20] Segalowitz N, Frenkiel-Fishman S. Attention control and ability level in a complex cognitive skill: Attention shifting and second-language proficiency. *Memory & Cognition*. 2005;33:644-653.
- [21] Best CC, McRoberts GW. Infant perception of non-native consonant contrasts that adults assimilate in different ways. *Language and Speech*. 2003;46:183-216.
- [22] Werker JF, Tees RC. Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behavior and Development*. 1984;7:49-63.



# What's the Bottom Line?

## Development of and potential uses for the Summary Translation Evaluation Tool—STET

Language analysts in the intelligence community (IC) confront huge amounts of foreign language material, some highly relevant to national security requirements and some of lesser value. Language analysts cannot provide full translations of everything they process, nor do other analysts have the time or need to read full translations. Summary translation enables language analysts to identify, distill, and present English translations of the important information contained in the original foreign language material, thereby efficiently communicating the most crucial information.

The term summary translation<sup>a</sup> can be used to cover a broad range of tasks that vary in their purposes and skill demands, from documenting essential elements of information contained in a single source item to writing a personality profile or situation assessment synthesizing information from many source items [1]. Despite this variability, a common attribute of summary translations produced within the IC context is that they are typically targeted summaries written in response to intelligence needs, i.e., customer-specified requirements or requests for information. In other words, writers may be looking for information about specific topics or answers to specific questions, rather than attempting to summarize all of the main points of the source item as they would in a generic summary.

The Summary Translation Evaluation Tool (STET) was created primarily for assessment of targeted summaries, which are uncommon in commercial and academic environments and therefore rarely studied and ill-understood. The STET was designed not only to help researchers develop a deeper understanding of targeted summary translation, but also to establish a standard for summary translations and to provide language analysts with a vital tool for both assessing and improving summary translation performance via standardized quality control (QC) and enhanced training.

## What is the STET?

The STET is a computerized form that offers a standardized framework for evaluating summary translation products<sup>b</sup>. The heart of the STET is a set of rating scales to assess summaries along six dimensions: Significance, Completeness, Accuracy, Omission of Irrelevant Information, Organization, and Writing. Each dimension is described in Figure 1. These dimensions were designed to cover all of the important elements of a summary's content, structure, and style, but the relative importance of each dimension may depend on the purpose for which the summary is written.

The STET also includes a description of the source item(s) on which the summary is based. Although users of the STET are instructed not to adjust their ratings based on the difficulty of the material, the Source Item Description identifies features of the material that may be especially challenging and helps provide context for the STET ratings.

As described in the STET user's manual, the Source Item Description and Summary Translation Assessment are "analogous to the difficulty and execution in an Olympic dive; one must describe

both the difficulty of the task and the skill with which it is executed in order to make a meaningful evaluation" [2].

Figure 1 shows the STET form along with descriptions of each major component. In the interactive version of the tool, pop-up windows provide more detailed information than is available on the one-page static form. For the Summary Assessment section on the right side of the form, each pop-up window describes the fundamental question that is addressed in the rating scale, often indicating what qualities a summary should possess to receive a high rating on that scale. Each pop-up window also provides labels for the end-points of the rating scale. The labels are tailored to each dimension; for example, the Organization scale ranges from "The summary is extremely poorly organized" to "The summary is extremely well organized." Finally, because the STET is intended to be consistent with the eight analytic standards issued by the Director of National Intelligence (Table 1), each pop-up window lists the particular standards that are addressed by that dimension.

## Potential uses of the STET

### Quality control

Language analysts and supervisors throughout the IC recognize the importance of quality control (QC), but there is currently little standardization of QC procedures. In addition to its critical function of ensuring that products are accurate, QC serves as a mechanism for training and providing feedback to language analysts. QC may also contribute to record-keeping and help supervisors determine work assignments. The STET is designed to facilitate all of these aspects of QC.

### Help for the QC provider—a more efficient and effective means of conducting QC

One of the most critical functions of QC is ensuring that intelligence products are of high quality. QC is important not just for a final report, but also for the translations, summaries, and other possible steps that may be completed along the way; if the original source material is translated and/or summarized inaccurately, there is a high risk that the final report will contain incorrect information.

QC is seen as especially crucial for junior analysts who may have limited

**Table 1: ODNI Standards of Analytic Tradecraft [3]**

1. Properly describes quality and reliability of underlying sources
2. Properly caveats and expresses uncertainties or confidence in analytic judgments
3. Properly distinguishes between underlying intelligence and analysts' assumptions and judgments
4. Incorporates alternative analysis where appropriate
5. Demonstrates relevance to US national security
6. Uses logical argumentation
7. Exhibits consistency of analysis over time, or highlights changes and explains rationale
8. Makes accurate judgments and assessments

<sup>a</sup> The term *gisting* is sometimes used synonymously with *summary translation*, but we will not use *gisting* here, because it often refers to a process that would not typically be evaluated with the Summary Translation Evaluation Tool (STET). For example, in some operational environments analysts use *gisting* to refer to the process of making brief notes about the content of an item for triage purposes, and *summary translation* to refer to a more formal summarization process.

<sup>b</sup> The STET has also been adapted for evaluation of other translation products. For example, one operational organization has created a spin-off called the Language Product Evaluation Tool (LPET), which can be used to evaluate summary translations, verbatim translations, or hybrids (in which some material is summarized and some is translated verbatim).



Figure 1: STET form

## Summary Translation Evaluation Tool

Item  Language Analyst

**A. Source Item Description**  voice  graphic  both

**A.1. Language Level** (Select overall level of source; mark all characteristics of significance in source.)   1  2  3  4

<input type="checkbox"/> cultural information	<input type="checkbox"/> lack of continuity
<input type="checkbox"/> diagrams, charts, graphs	<input type="checkbox"/> meaning beyond the literal
<input type="checkbox"/> greater than average length	<input type="checkbox"/> multiple objects or concepts
<input type="checkbox"/> high density of information	<input type="checkbox"/> rhetorical devices
<input type="checkbox"/> highly technical subject matter	<input type="checkbox"/> shared knowledge
<input type="checkbox"/> inference based on overt info	<input type="checkbox"/> spatial relationships
<input type="checkbox"/> intentional deception	<input type="checkbox"/> telling out of sequence

**A.2. Complicating Mode Factors** (Mark all of significance in source.)

<input type="checkbox"/> communicants speaking over one another	<input type="checkbox"/> non-standard colloquialisms or slang
<input type="checkbox"/> corrupt source	<input type="checkbox"/> omissions
<input type="checkbox"/> dialect	<input type="checkbox"/> one-sided conversation
<input type="checkbox"/> distortion	<input type="checkbox"/> poor grammar
<input type="checkbox"/> elliptical or telegraphic style	<input type="checkbox"/> poor handwriting
<input type="checkbox"/> heavy accent	<input type="checkbox"/> poor spelling
<input type="checkbox"/> more than one language or dialect or alphabet	<input type="checkbox"/> rapid speech
<input type="checkbox"/> non-standard abbreviations or specialized terminology	<input type="checkbox"/> sudden changes in subject
	<input type="checkbox"/> typographical errors
	<input type="checkbox"/> urgency (need for time-sensitive processing)

**A.3. Impact of Complicating Mode Factors** (Mark one.)  
 none  inconsequential  moderate  considerable  extensive

**B.8. QCer 2 Comments** (Use appropriate handling and classification markings, if needed.)

**Written Comments**—Users are encouraged to provide written comments. Ratings on a single dimension (compared to an overall rating) can reflect specific problems that might benefit from different types of interventions. For example, a Writing score could be poor because the summary is full of typographical errors or because the writer is a non-native speaker of English who does not have an adequate grasp of English grammar.

**Language Level**—Users determine the ILR level of the source item and indicate which characteristic(s) of the item contributed to that level. In the electronic version of the STET, the language levels and characteristics include pop-up windows with descriptions and examples.

**Complicating Mode Factors**—The presence of these factors can make an item more difficult to understand, translate, or summarize. Some factors are specific to one modality; other factors, such as colloquialisms or poor grammar, can be found in both voice and graphic items.

**Impact of Complicating Mode Factors**—Users assess the degree to which the complicating factors impact the ability to understand, translate, or summarize the item. For example, typographical errors may be relatively inconsequential in one text but may render another virtually incomprehensible.

**Significance**—The summary should clearly demonstrate why the source item is relevant to national security and should indicate how the relevant information relates to what is already known about a particular requirement. Users may indicate that the dimension is not applicable if, for example, a generic summary is required rather than a targeted summary.

QCer 1  QCer 2

**B. Summary Assessment** (Mark a number for each factor; add comments.)

**B.1. Significance:** How well does the summary relate the “so what” of the source item to requirements?  NA  1  2  3  4  5

**B.2. Completeness:** How much of the essential information is covered in the summary:  
who? what? when?  
where? why? how?  
relevant background? analytic comment?

**B.3. Accuracy:** How much of the information in the Summary is accurate?  1  2  3  4  5

**B.4. Omission of Irrelevant Information:** How well does the summary omit irrelevant information?  1  2  3  4  5

**B.5. Organization:** How well organized is the summary: “bottom line” up front? logical organization? well-structured paragraphs?  1  2  3  4  5

**B.6. Writing:** How well does the summary follow conventions for:  
grammar? spelling?  
punctuation? word usage?  
date, time, transliteration, etc.?  
 1  2  3  4  5

**B.7. QCer 1 Comments on Summary Translation** (Use appropriate handling and classification markings, if needed.)

**Completeness**—Users address the degree to which the summary contains all of the relevant information. In addition to presenting the facts that are explicitly stated within the source item, a good summary may need to include explanatory facts available elsewhere and analytic comments.

**Accuracy**—Many consider accuracy to be the most fundamental component of the STET. If a summary receives a low rating for accuracy, the summary will be of very limited value even if it receives high ratings for all of the other dimensions.

**Omission of Irrelevant Information**—A critical feature of summarizing is efficiently communicating the most important information in the source material. Because of the targeted nature of summary translation within the IC, the source material may often contain a great deal of information that is not relevant to national security.

**Writing**—Poor grammar, spelling, and punctuation can obscure the message in an otherwise good summary. This dimension takes into account conventions for reporting dates, times, and transliterations of foreign names—particularly important because foreign names can be spelled many different ways in English.

**Organization**—A good summary must be organized in such a way that the message is communicated clearly. One of the most highly valued organizing principles within the IC is “bottom line up front.” Sometimes language analysts will need to alter the original organization of the source material to convey the information in a way that most directly addresses the information need.

experience with the target and/or language. However, most professionals realize that even highly seasoned experts can benefit from having their work reviewed by colleagues. During structured interviews conducted by our summary translation research team, nearly all language analysts noted that “nothing goes out the door without being seen by at least two pairs of eyes.”

Because of the vast amount of material that is processed every day, QC places a heavy burden on the most experienced analysts. One goal of the STET is to facilitate the process of conducting QC. Although an initial time investment may be required for QCers to learn about the STET and become accustomed to using it, the STET can ultimately make QC faster and more effective by providing a standardized framework for evaluating summary translations.

#### **Help for the QC recipient—more detailed and useful feedback**

The “checking” aspect of QC emphasizes forward movement in the sense that each piece of work is checked and passed forward to the next person in the chain. For example, a language QCer may check a translation and pass the corrected version forward to a reporter.

Ideally, QC also involves a “backward” step in which feedback is provided to the language analyst who wrote the initial translation or summary. Such feedback is vital for improving the junior analysts’ language skills and helping them to avoid similar mistakes in the future. Unfortunately, the fast operational tempo sometimes makes it difficult for QCers to provide feedback in a timely manner.

The structure of the STET will allow QCers to generate feedback at the same time that they are checking the work, thus helping both the QC provider and recipient, and the multidimensional nature of the STET will ensure that the feedback is detailed enough to help language analysts pinpoint specific areas for improvement.

#### **Help for the supervisor—a mechanism for tracking progress**

Over time, the STET will help supervisors and managers keep track of strengths and weaknesses in order to determine work assignments and note opportunities for targeted training for individuals or groups. For example, a supervisor who is using the STET to track an analyst’s progress may note that the analyst performs very well with Level 2 material but less well with Level 3 material; this type of pattern can be useful in making appropriate work assignments. In addition, the detailed nature of the Source Item Description may reveal that the analyst excels in the face of certain challenges but struggles with others, allowing for identification of individually tailored professional development activities. Aggregated STET data may also help managers to determine whether an entire shop’s performance is affected by factors such as new software tools, new mentoring programs, or changes to the physical workspace.

#### **Training**

Not only will on-the-job training benefit from the STET via improved QC and feedback, but classroom training will also be able to capitalize on the STET. Perhaps most critically, the STET will provide instructors with a coherent framework for teaching students about the components of a good summary. Instructors can also use the STET to provide standardized feedback on student assignments, which will make the grading process more efficient for instructors and more useful for students. Using the STET in the classroom will also help students become accustomed to the way they will be evaluated on the job.

The Source Item Description section of the STET may also be helpful for instructors in guiding the selection of texts that are at an appropriate level for the class and that present students with particular challenges that are relevant to the lesson.

#### **Research**

At the University of Maryland Center for Advanced Study of Language (CASL),

one of the goals of research on translation is to understand the cognitive and procedural processes involved in summary translation, and to apply that understanding to improve the performance and training of language analysts. With respect to both aspects of this aim—understanding and application—evaluation is an essential component. The STET will provide researchers with a valuable mechanism for evaluating the summary translations that are produced in experiments. The STET is a powerful tool for experimentation because it allows the researchers to examine summarization performance along a variety of dimensions.

In one CASL experiment using the STET, we are examining the ways in which varying amounts of time pressure impact the quality of summary translations and the strategies that language analysts use to create them. In this experiment, we ask language analysts to summarize a different foreign language text in each of three time conditions: 2 hours, 1 hour, and ½ hour. (The order of the time conditions and the assignment of text to condition are counterbalanced across participants.)

With a holistic rubric, we would only be able to determine whether summary translations were “better” in one condition than in another. With the STET, however, we can look at the effects of time pressure on different aspects of the summarization process. For example, we might see that Significance is relatively unaffected by time pressure if language analysts prioritize the need to identify the critical intelligence value of the source item; Completeness, on the other hand, might suffer under extreme time pressure if the language analyst does not have sufficient time to include all important details. Similarly, Organization might be relatively stable, but Writing might be vulnerable to time pressure when language analysts do not have time to proofread or check their work.

This detailed level of analysis enabled by the STET will help researchers better understand the various components of the summary translation task and guide the development of interventions to help



analysts maintain key components of summary quality under trying conditions.

### Development of the STET

As part of CASL's first experimental study of summary translation, our research team developed a "holistic" rubric, which assigned a single qualitative rating to each summary. The scale was developed using a modified empirically based binary-boundary approach [4], meaning that we relied on collaboration and consensus of qualified professionals using an iterative process of categorizing and characterizing salient features to arrive at descriptors for each level of proficiency [5]. This type of scale is probably similar to informal evaluations used in the (IC) and is typically fairly quick to use. However, a holistic evaluation provides only a single rating of the summary, and two summaries could receive the same rating for very different reasons (e.g., one due to poor comprehension of the source item and one due to poor English writing skills).

We ultimately decided that an "analytic" rubric would be more useful for both experimental and applied purposes, as described above, i.e., summary translation evaluation via an analytic rubric returns much more informative feedback about performance, allowing for more powerful experimentation as well as more individualized on-the-job evaluation and training. The current version of the STET was developed as a collaborative effort between CASL researchers and our United States Government (USG) colleagues, capitalizing on scholarly literature, scientific methods, and the operational expertise of many language analysts.

### Characteristics of an analytic rubric

Our development of an analytic rubric had the goal of making transparent the component processes involved in summary translation while also ensuring that the STET would be easily understood and used. Preliminary effort sought to derive a set of unidimensional evaluatory elements, and was informed both by the

existing scientific literature and by an analysis of the original holistic rubric. Following standards in educational and psychological measurement [6], we strove to produce an analytic rubric characterized by the following qualities:

- All elements worthy of evaluation are included.
- Each element is unidimensional in that it cannot be further separated or partitioned. (Given this quality, we refer to each element as a dimension.)
- Ratings are distinct, comprehensive, and descriptive in that they cover the range of expected performance.
- Each element of the rubric communicates clearly to the user.
- The rating score on each element covers the range of performance, perhaps in the range of 3-7 levels.

### Characteristics of a good summary

To develop the appropriate dimensions for the STET, we needed to identify the most important components of summary translation based on existing science and operational needs. From a scientific perspective, we began by examining literature in fields such as language processing, memory, translation, reading comprehension, and spoken-discourse comprehension. To determine what constitutes a good summary for operational purposes, we conducted structured interviews with language analysts, intelligence analysts, QCers, and instructors. We asked interviewees to describe the ways in which they write or use summary translations on the job and what they look for in a good summary translation. In addition to these individual structured interviews, focus groups were convened to determine the qualities that users would look for in an evaluation tool and to solicit comprehensive feedback on preliminary drafts of the STET.

One of the most important characteristics of a good summary is that it accurately reflects the meaning of

the source text. According to Jonassen, Beissner, and Yacci [7], the meaning of a source text comes from its structure—the way its propositions are related to each other and organized. In other words, the reader must derive structural knowledge of the text and be able to integrate each successive part of the text with his or her prior representation of it. Related to this point, learning and understanding of the text take place by assimilating new knowledge with prior background knowledge.

For a language analyst to produce a quality summary, he or she must have (1) sufficient language proficiency to achieve a discourse-level understanding of the text and not just a word-by-word glossing and (2) adequate background knowledge to assimilate the text with prior structural knowledge. Preliminary CASL research on summary translation demonstrated that an insufficient discourse-level understanding of the text often led to gross misunderstandings and eventually to inadequate summaries [8]. A solid discourse-level understanding of the text is assessed most directly by the Accuracy dimension of the STET. In addition, a thorough understanding of the text and its relation to the relevant background knowledge are also necessary for the reader to determine and explain how and why the text relates to intelligence needs, as assessed by the Significance dimension of the STET.

Endres-Niggemeyer [9] emphasizes the importance of representing and understanding discourse via schemata (structured groups of concepts used to organize knowledge). This emphasis implies that summarizers must be able to identify the schematic elements of the text and the relationships and actions between them. Some of these elements will be crucial to the summary, and the success of their identification should be evaluated by the rubric. For example, does the summarizer correctly identify the relevant participants, their roles in the text, and the actions taking place? Identification of the key pieces of information in the text is



assessed by the Completeness dimension of the STET.

Another fundamental skill required for summarization is information reduction. Ultimately, in a quality summary only the relevant information should remain, and the irrelevant information should be discarded, as measured by the Omission of Irrelevant Information dimension of the STET.

Finally, the summary translation must be appropriately communicated, which might be evaluated as having macro-structural and microstructural dimensions (Organization and Writing in the STET). The former is clearly reflected in the organization of the summary and the latter in its grammar. Important macrostructure components include clear, coherent overall organization and a structure that makes evident how the summary responds to the relevant information need(s). Important microstructure components include proper grammar, spelling, and word choice, as well as a writing style that is clear at both the sentence and clause level.

#### Properties of the rating system

Once the dimensions were established, we had to decide on other properties of the rating system, including the appropriate number of points on the rating scale and the descriptors of the different levels.

We ultimately decided to use a 5-point rating scale. Although some of the potential users we consulted felt that a smaller number of points would make the scale faster and easier to use, research has demonstrated that reliability and validity are better in 5- to 7-point scales than in scales with fewer points [10-13]. Also, respondents tend to avoid using the endpoints of a scale [14], so having a 4-point scale could potentially concentrate most of the responses on only two points, which may not be sensitive enough to detect subtle differences in summary quality.

Another important decision was how to label the five points of the scale. We initially attempted to write a detailed

description of each rating of each dimension of the STET. We discovered, however, that most of the important information appeared in the descriptor for the highest rating, which listed the characteristics required for a high quality summary. By putting all of the desired characteristics up front after each question, the endpoints of the scale emerged naturally without the need for descriptions of the intermediate ratings. This decision was validated by research suggesting that the labels for intermediate scale intervals are not as critical as the choice of endpoint labels [15,16].

#### Refining the STET

Once the dimensions and rating system were established, we used several approaches (some of which are still in progress) to refine the STET.

##### Rigorous practical testing

Since a major aim of rubric development is the application of the rubric to the training and evaluation of language analysts working within government agencies, practical testing by representatives of those agencies is crucial. This “beta testing” will allow agencies to report experiences using the rubric and to provide feedback so we can make the rubric maximally user-friendly and useful for their needs. For example, it will likely be valuable to build into the rubric a degree of modularity, so that particular dimensions can be added or omitted as needs dictate. Similarly, the size of the rating scale might be collapsed or expanded according to practical needs. In coordination with practical testing, statistical testing will be used to confirm that adapted versions of the rubric are valid, sensitive, and reliable.

##### Rigorous statistical testing

Rigorous statistical testing is required to examine the validity, sensitivity, and reliability of the STET against gold standards, i.e., summaries pre-established by experts to represent certain levels of performance along the different dimensions. This testing will allow us to determine if variations in each aspect of summary quality are appropriately

reflected in ratings for the corresponding dimension (validity) and if the STET adequately assesses the full range of performance along each dimension (sensitivity). Statistical testing will also help to determine if the dimensions are treated independently or if, for example, grammatical errors affect Accuracy scores as well as Writing scores. Lastly, the STET will be tested thoroughly to establish its consistency across users and conditions (reliability). CASL researchers are currently conducting a set of experiments to accomplish these goals.

#### Conclusion

The STET is the result of a needs-based approach to research in which a multidisciplinary team of scientists collaborated with members of the operational workforce to develop a product that addresses both operational and scientific problems. This analytic rubric for evaluating summary translations was developed to benefit language analysts, QCers, and their managers, resulting in better reports and better language analysts. The STET will also allow CASL scientists to deepen our understanding of the summary translation process so we can continue to conduct rigorous research to enhance language performance in the IC. 📌

## References

- [1] Michael EB, Bailey B, Gannon-Kurowski S, Pinckney K. Description of summary translation task requirements for specific jobs and the uses of the summaries. College Park (MD): University of Maryland Center for Advanced Study of Language; 2007. Technical Report No.: M.19.
- [2] University of Maryland Center for Advanced Study of Language. The Summary Translation Evaluation Tool: What it is and how to use it. [User's manual]. College Park (MD); 2008. p. 3.
- [3] McConnell J. Intelligence community directive number 203: Analytic standards; 2007.
- [4] Upshur J, Turner C. Constructing rating scales for second language tests. *ELT Journal*. 1995;49:3-12.
- [5] Turner J. Report on working session: Rubric development and rater orientation. In: B. Bailey B, Turner J, Pinckney K, de Terra D, Michael E. Technical Report No. M.16: Compilation of milestones examining evaluation of translation summaries. College Park (MD): University of Maryland Center for Advanced Study of Language; 2007.
- [6] Crocker L, Algina J. Introduction to classical and modern test theory. Mason (OH): Thomson Wadsworth; 2006.
- [7] Jonassen DH, Beissner K, Yacci M. Structural knowledge: Techniques for representing, conveying, and acquiring structural knowledge. Hillsdale (NJ): Lawrence Erlbaum Associates; 1993.
- [8] Michael EB, Allison T, Danks J, de Terra D, Massaro D, Donavos D, Graham K, Klavans J. Does verbatim translation help summary translation? Paper presented at the 6th International Symposium on Bilingualism; Hamburg, Germany; May 2007.
- [9] Endres-Niggemeyer B. Summarizing information. Berlin (Germany): Springer-Verlag; 1998.
- [10] Bandalos DL, Enders CK. The effects of nonnormality and number of response categories on reliability. *Applied Measurement in Education*. 1996;9:151-160.
- [11] Dawes J. Do data characteristics change according to the number of scale points used? An experiment using 5-point, 7-point and 10-point scales. *International Journal of Market Research*. 2008;50:61-77.
- [12] Jenkins GD Jr, Taber TD. A Monte Carlo study of factors affecting three indices of composite scale reliability. *Journal of Applied Psychology*. 1977;62:392-398.
- [13] Lissitz RW, Green SB. Effect of the number of scale points on reliability: A Monte Carlo approach. *Journal of Applied Psychology*, 1975;60:10-13.
- [14] Beal DJ, Dawson JF. On the use of Likert-type scales in multilevel data: Influence on aggregate variables. *Original Research Methods*. 2007;10:657-672.
- [15] Gannon KM, Ostrom TM. How meaning is given to rating scales: The effects of response language on category activation. *Journal of Experimental Social Psychology*. 1996;32:337-360.
- [16] Klockars AJ, Yamishi M. The influence of labels and position in rating scales. *Journal of Educational Measurement*. 1988;25:85-96.

## Further Reading

Jonassen DH, Howland J, Moore J, Marra RM. Learning to solve problems with technology: A constructivist perspective. 2nd ed. Columbus (OH): Merrill/Prentice-Hall; 2003.

