An Integrated Approach to Language Capabilities in Humans and Technology

Colin Phillips, Director, Maryland Language Science Center, colin@umd.edu
Tess Wood, Assistant Director, Maryland Language Science Center, ewood1@umd.edu
Shevaun Lewis, Assistant Director, Maryland Language Science Center, shevaun@umd.edu

The world has been transformed by the Internet, mobile devices, globalization of markets, and security threats from small groups that can emerge anywhere, anytime. In recent years, awareness has grown in the US of the importance of training and maintaining capabilities in diverse languages - for diplomacy, business, security, and international development. However, the needs and challenges surrounding language capabilities are far broader than language training.

Language is the most direct and accessible means for identifying the intentions and mental states of leaders, groups and individuals. It is also the most straightforward medium for communicating with and influencing others. Directly observing interactions in a language that we understand natively remains unmatched as a source of information. This is especially true if we have information about the participants and if we can detect the nuances of what is and is not said, and how. But the detailed information conveyed in even a brief interaction is lost if we lack the detailed knowledge about the speakers and their language. There are over 6,000 languages in the world, but detailed expertise is available for only a small fraction of them, and even that expertise is often not in a form that is useful for end users. Meanwhile, language technology is needed to triage vast amounts of data for further analysis by humans, but even in the best understood languages this technology is rudimentary when it comes to detecting high-value information.

The language challenge is too broad be solved by brute force. The range of language expertise that could become critically necessary is so diverse that it is not possible to simply aggregate that expertise ahead of time. A more feasible approach is to be able to rapidly ramp up human and technological capabilities in response to needs as they arise. This requires establishing a multi-pronged approach that combines a powerful base of human and technological language expertise with techniques for quickly and flexibly leveraging that base for specific situations in specific languages or regions. The promise of this approach has been recognized for some time, especially since 2001. On the technology side it can be seen in a number of DARPA projects, such as the current LORELEI project for ‘low resource’ languages. On the human side it can be seen in the creation in the mid-2000s of the Center for Advanced Study of Language, a university-affiliated research center (UARC) based at the University of Maryland.

A limitation of much existing research in this area is that it has been distributed across loosely connected or disconnected fields. Scientific expertise in human language has traditionally been distributed across a range of fields, including linguistics, psychology, computer science, education, and health disciplines. Flexibly combining expertise from these areas is essential for addressing complex problems in language. “Language Science” is the integrated field that can address these challenges. Recent efforts to at integration across fields, along with advances in technology and methods, have created opportunities for rapid advances the science of language in many interdisciplinary areas. These include:

- Human Language Technology: Simultaneous machine translation, speech recognition, automatic language and dialect identification;
- Building and integrating resources on world languages for use by humans and technology, and crowdsourcing language expertise;
- Rapidly building language capabilities in the workforce, and maintaining flexible readiness;
- Measuring language performance in humans and machines.
- Addressing how listeners understand - or fail to understand - language in adverse conditions (noise, dialect differences, non-native proficiency, hearing impairment, cochlear implant, divided attention, preconceptions);
- Combined human and machine evaluation of attitudes, psychological states and persuasiveness of information.

Human language technology.
Advances in areas such as simultaneous machine translation and speech recognition benefit from research on how humans perform so well. For example, humans outstrip current technologies in their ability to accommodate variation in speech. What gives humans this edge, and is that within reach of near-term technologies? Meanwhile, simultaneous (word-by-word) machine translation (SMT) is made difficult by word order differences between languages, e.g., when translating a sentence from Japanese to English the verb much appear early in the English sentence, but it is the last word to appear in the Japanese source sentence. In such situations it is only possible for a SMT system to proceed smoothly if it predicts upcoming words in the source sentence. Such technologies could benefit from integration with research in psycholinguistics and cognitive neuroscience on how humans anticipate upcoming input during comprehension. More broadly, human language learners easily outperform current technologies, if the learning occurs in childhood, despite the fact that the humans learn from far less training data than is available to technologies for high-resource languages.

Integrating broad language resources
Global advances in language technology, education, health, commerce, and disaster response depend on the availability of interoperable resources in thousands of languages. But current knowledge is distributed, uneven, and rarely interoperable. To take a scenario that occurs frequently: if an earthquake, disease outbreak, or security threat occurs in a location where very little is known about the local languages, how can we create tools and information that can help people on the ground as quickly as possible?

Extensive, well organized digital resources are available in less than 1% of the world's 6000 languages, and those that exist are often designed for specific audiences or user needs, or of uncertain quality. For some languages excellent resources are available, but they are all written in Russian. For many languages some audio recordings are available, thanks to extensive missionary work, but the recordings may be inauthentic because they are recordings of non-native speakers. In some languages the resources are available only in a format that is useful to technologists, or language typologists. And in most cases there is simply very little information available digitally. In those cases, the expertise simply resides in the speakers themselves.

It is not feasible to turn all languages into high-resource languages like English or Chinese. So the greatest progress will come from learning how to do more with less, e.g., can we create language technologies that learn from 1 million words of text what is currently learned from 20 million words of text. And it will also come from learning how to get a head-start on under-resourced languages by leveraging what we already know about closely related well-resourced languages. This approach holds a lot of promise, but it presupposes a good understanding of relatedness among languages.

In order to address these challenges, we need integrated, interoperable resources, with the scope for integrating new material via expert contributions and via crowd-sourcing. The resources need to be accessible through multiple languages, though this could rely on lingua francas, e.g., a Swahili interface could be useful for engaging with speakers of many East African languages. Also, the resources need to be accessible to users with diverse interests and levels of expertise. Also, in order for a distributed global network of experts to be motivated to contribute, it is also essential that the resources or standards have a long-term trajectory.

Langscape (langscape.umd.edu) is a project based at the University of Maryland that aims to address this need. Langscape is an entirely open resource that aims to aggregate information on the world's languages through an easy to use GIS interface. The focus is on gathering expertise from experts around the world, and making it available to users with different levels of expertise and goals. It also has a non-public mirror version that is usable on government computing systems and that can integrate closed information sources. The mirror is maintained by a different group. The existing release is a proof of concept for the GIS interface. Future development plans include hosting diverse map layers, developing data standards for interoperability via a WikiData initiative, and developing crowdsourcing capabilities.

Language learning methods, aptitude and heritage learners
It is not possible for the government to have sufficient expertise on hand for all languages that might become important. So it is instead valuable to have the capability to rapidly develop new expertise, by training and retraining professionals for new languages, and leveraging a worldwide network expertise in related languages. Achieving this requires research on adult language learning, with a focus on how to identify and train learners with the greatest aptitude. This is a departure from most adult language learning research, which is based on cross-sectional samples of college students who have varying motivation and receive low-to-medium intensity training. How can we identify aptitude for language learning? How can we leverage skills already acquired in one language to facilitate learning another language? What can be achieved through more intensive bouts of training.

A particularly interesting target for research is so-called "heritage language learners". These are bilinguals who grew up in immigrant communities in the US, but may not be very strong in their home language and stronger in English. However, such bilinguals nevertheless have a strong head start over people who want to learn that language from scratch (L2 learners): heritage learners require less instruction and can achieve a higher level of competence. Understanding what they know and using this understanding to develop practical implications for language pedagogy is important socially, culturally and economically, because these speakers can meet a

number of needs for critical languages, and there are tens of millions of them in the US. Montrul (2015). *The Acquisition of Heritage Languages;* white paper from the National Heritage Language Research Center

Measuring performance in humans and machines
Research is needed on how to measure the performance of human language learners or new language technologies. On the human side, government incentives are tied to performance benchmarks, and so this creates a bias towards learning languages where tests exist. In order to grow diverse expertise, it is necessary to identify ways of reliably assessing proficiency in new languages without the time, money, and expertise that is available for major languages. On the machine side, there is a need to better assess how technologies fare in identifying high value information, rather than high frequency information, i.e., things that are said often. Current methods are limited, but they can benefit from collaboration with cognitive scientists and neuroscientists who are skilled in measuring the language capabilities of highly proficient humans.

Language understanding in adverse conditions
Although automatic speech recognition systems have become more reliable in quiet environments with a single talker, they are still rather poor at dealing with noisy, multi-talker environments. Humans excel in these challenging situations: when listening to our native language, we easily filter out substantial background noise and fill in missing or obscured information. This remarkable efficiency is made possible by cognitive and neural mechanisms that process multiple levels of acoustic and linguistic information simultaneously and allocate attention to relevant input. Multidisciplinary basic research on these mechanisms--ranging from animal models of auditory processing to psycholinguistic research on adults' real-time comprehension in a first or second language--may help improve automatic speech recognition.

Evaluating opinions and mental states
Computational linguists are developing methods to identify aspects of language that are connected with underlying mental state and mental status. This includes analyzing language for signals related to mental health, e.g. depression, schizophrenia, or suicidality. It also includes computational methods for identifying and characterizing differences in how individuals and groups frame events and issues, with broader applications involving not only mental health but computational political science. Friedenberg et al. (2016), Hong et al. 2016, Resnik et al. 2015, Iyyer et al. 2014 These technologies have complementary capabilities to human analysts. The technologies may be able to perform triage, and they can also identify data patterns that might be missed by humans. But the technologies are not so good at identifying the pragmatic nuances of a conversation, which often reveals a huge amount of information not directly conveyed by the words spoken. This is an area where basic research on cross-language and cross-cultural variability in pragmatics is needed.