# Establishing a Human Language Technology Center of Excellence

*G. Strong*, J. Eisner*†, and C. Piatko*‡*

**JHU Human Language Technology Center of Excellence, Baltimore, MD;*
*†JHU Whiting School of Engineering, Baltimore, MD;*
*and ‡JHU Applied Physics Laboratory, Laurel, MD*

*In January 2007, JHU was awarded a long-term multi-million dollar contract to establish and operate a Human Language Technology Center of Excellence (HLTCOE) adjacent to the JHU Homewood campus. The HLTCOE's research focuses on advanced* technology for automatically analyzing a wide range of speech, text, and document image data in multiple languages. Other key members in jointly establishing this HLTCOE included the Johns Hopkins Center for Language and Speech Processing, the University of Maryland College Park, and BBN Technologies. The focus of the technical program is on automatic population of knowledge bases from text, proof-of-concept experiments for robust speech technology, and stream characterization from content. These projects address key issues in extracting information from massive sources of text and speech.

Many important applications will become possible when systems can automatically produce language-independent structured representations of knowledge derived from unstructured text, speech, and document image data in a wide variety of languages and genres. The derived knowledge can be aggregated into a cumulative knowledge base, but it can also serve as input to a range of analytic and inference technologies (Fig. 1). Although humans could potentially extract the kinds of information needed (such as various classes of entities, relations, events, opinions, scenarios, and so forth), the large volumes of data, the complexity, and the required level of detail make such tasks impractical. Reasonably accurate, fully automatic methods are therefore essential.

In recognizing what a speaker is saying, a human listener uses an enormous amount of language knowledge and world knowledge that is difficult even to represent on a computer, much less to learn automatically. Therefore, automatic speech recognition has been a challenging and exciting area of research for several decades. However, even after decades of development that has been very successful in some areas, automatic systems still fall far short of human listeners in the ability to tolerate moderate changes in the speech that deviate from the current model. The challenge for the HLTCOE is to develop new methodologies that not only improve the overall performance of speech recognition and related tasks but also maintain this performance across a wide range of changes in the speech or language characteristics.

In many applications, in particular in both spoken and written language applications, there often is a large amount of data. Many experiments in language technology require that large quantities of these data be manually labeled so that automatic learning algorithms can build sophisticated models of the data. However, manual annotation of a large quantity of data is both expensive and time-consuming. A common challenge in both speech recognition and text-based language analysis is to turn the large quantity of data into a resource rather than a burden (Fig. 2). Meeting this challenge requires research at the cutting edge of automatic learning techniques. The challenge is to develop high-speed algorithms that are reasonably accurate and can deal with data in motion (rather than archival data stores).

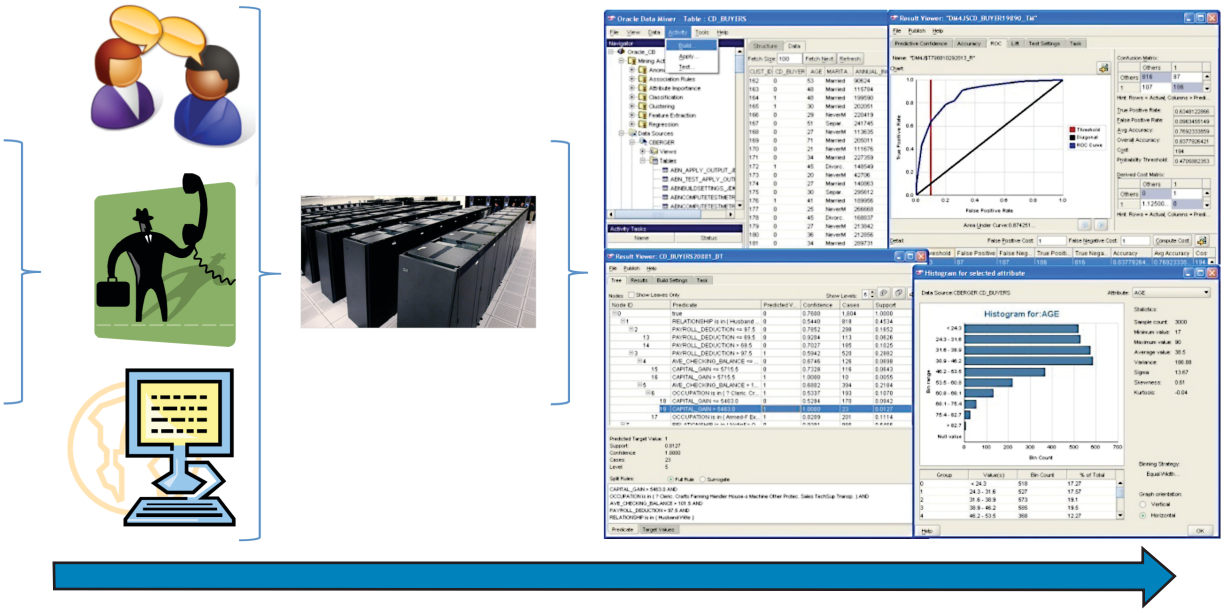


**Figure 1.** The cumulative knowledge base created from the text, speech, and document image data can serve as input to a variety of analytic technologies.

Some of the challenges faced by the HLTCOE are as follows:

- **Producing structured knowledge from unstructured language data**. The HLTCOE must create the ability to extract knowledge from various languages, genres, and media. Entities, events, and relations must be disambiguated and normalized. The knowledge base must represent changes over time, and the system must accommodate heterogeneous data and error-full inputs.
- **Creating robust technology for speech.** It is a necessity to create capabilities—including speech-to-text, speaker identification, and language identification—that work well on many languages and conditions. These capabilities should be easily ported to new languages, and the system should be no more susceptible to input signal differences than humans are.
- **Avoiding a data annotation bottleneck.** Efficient, effective ways must be devised to train algorithms. These ways must not require large quantities of manually annotated, domain-and-condition-specific data; may exploit vast quantities of unannotated data; and should work with resource-rich as well as resource-poor languages.
- **Characterizing the data stream automatically.** Some applications involve volumes of streaming data whose content must be analyzed to identify data similar to other data previously judged to be of interest, data that differ from previous norms or are otherwise anomalous, and overall stream characteristics (e.g., distribution of spoken languages). HLTCOE's challenge is to develop high-speed algorithms that are reasonably accurate, can analyze speech or text (either would be useful), and can deal with data in motion (rather than archival data stores).

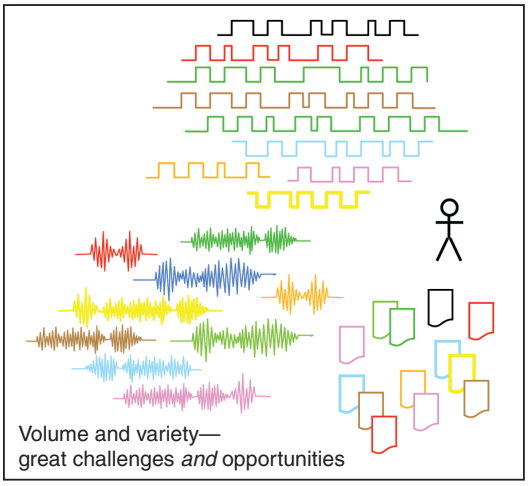


**Figure 2.** The great volume and variety of data to be analyzed (speech, text, and documents) from multiple languages is the HLTCOE's biggest challenge—and opportunity.

**For further information on the work reported here, see the references below or contact christine.piatko@jhuapl.edu.**

[1]JHU Human Language Technology Center of Excellence website, *Welcome!*, www.hltcoe.org (accessed 21 Dec 2009).

[2]JHU Human Language Technology Center of Excellence publications page, *Publications*, www.hltcoe.org/Publications.html (accessed 21 Dec 2009).