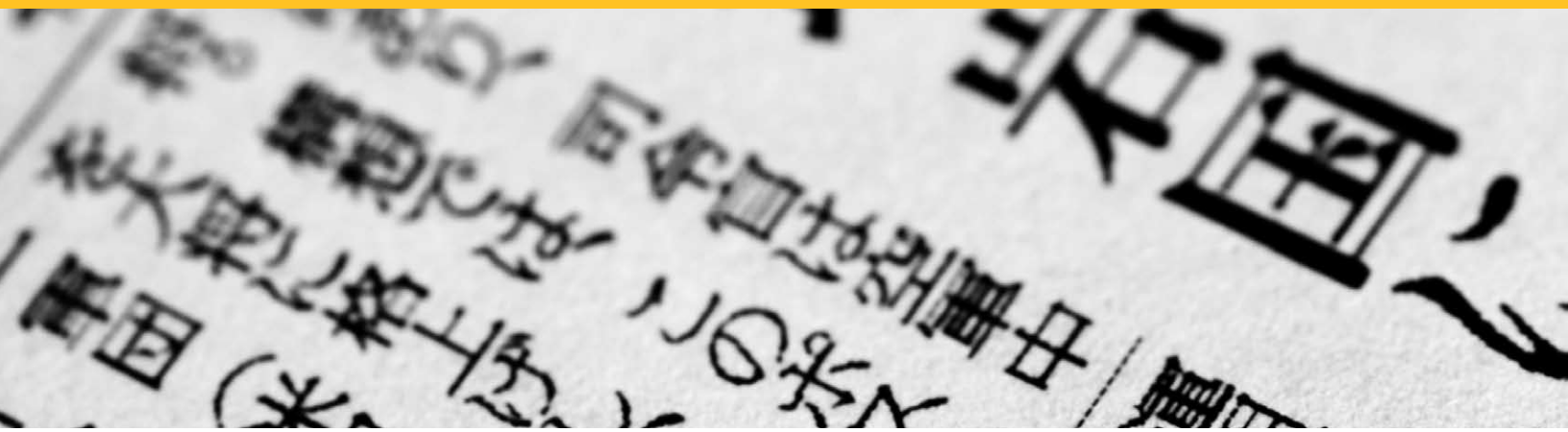
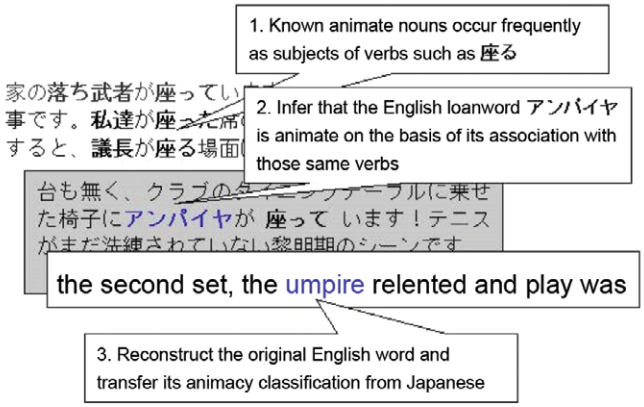


How do we communicate?

Project lead:
Christopher Brew, Ph.D., The Ohio State University
Research title:
Hybrid methods for acquisition & tuning of lexical information
Funding source:
National Science Foundation Career Grant



above: This diagram illustrates one of the ideas developed by linguists at The Ohio State University. The technique uses information in Japanese to get at information about animacy that is missing from English.

As a computational linguist, Ohio State University Professor Chris Brew merges computer science with the scientific study of language and communication. Computational linguists create and test computational models of linguistic theories, as well as develop and apply tools for computers to complete real world tasks, such as extracting information from text, translating languages, and synthesizing and recognizing speech.

At the center of Dr. Brew's work is natural language processing, the development of ways for computers to use and understand everyday human language. In particular, Dr. Brew is working on the problem of building and maintaining dictionaries (lexicography) and grammars that reflect not only the insights of linguists and lexicographers, but also the evidence from large bodies of actual text.

"Broad-coverage dictionaries and ontologies (terms and descriptions) for natural language processing are difficult and costly to create and maintain," Dr. Brew said. "Because of the quantity of text available, computer support is essential. But while all modern lexicography is computer-based, the information used by today's systems is fairly shallow, simply because of the scale of the enterprise.

"We're working on ways of enriching the information available to lexicographers and other professionals. The challenge is to do this without losing the benefits of scale," said Dr. Brew, whose team includes graduate students Kirk Baker and Jianguo Li. "We found this is possible, and that a representation known as "dependency triples," i.e. groups of words that relate directly to each other, gives a significant boost over just counting words."

Using the Ohio Supercomputer Center systems, Baker led the effort to create these dependency triples from the Gigaword corpus, which contains a billion words of newswire text. With a supercomputer, they were able to complete in 72 hours what otherwise would have taken all summer.

"While our immediate goal is to gain a better understanding of lexical tuning and acquisition," Dr. Brew said, "the resulting dictionaries, ontologies and mapping techniques have the potential to help information professionals such as librarians navigate through data, understand their significance and incorporate insights into their working practice." ■