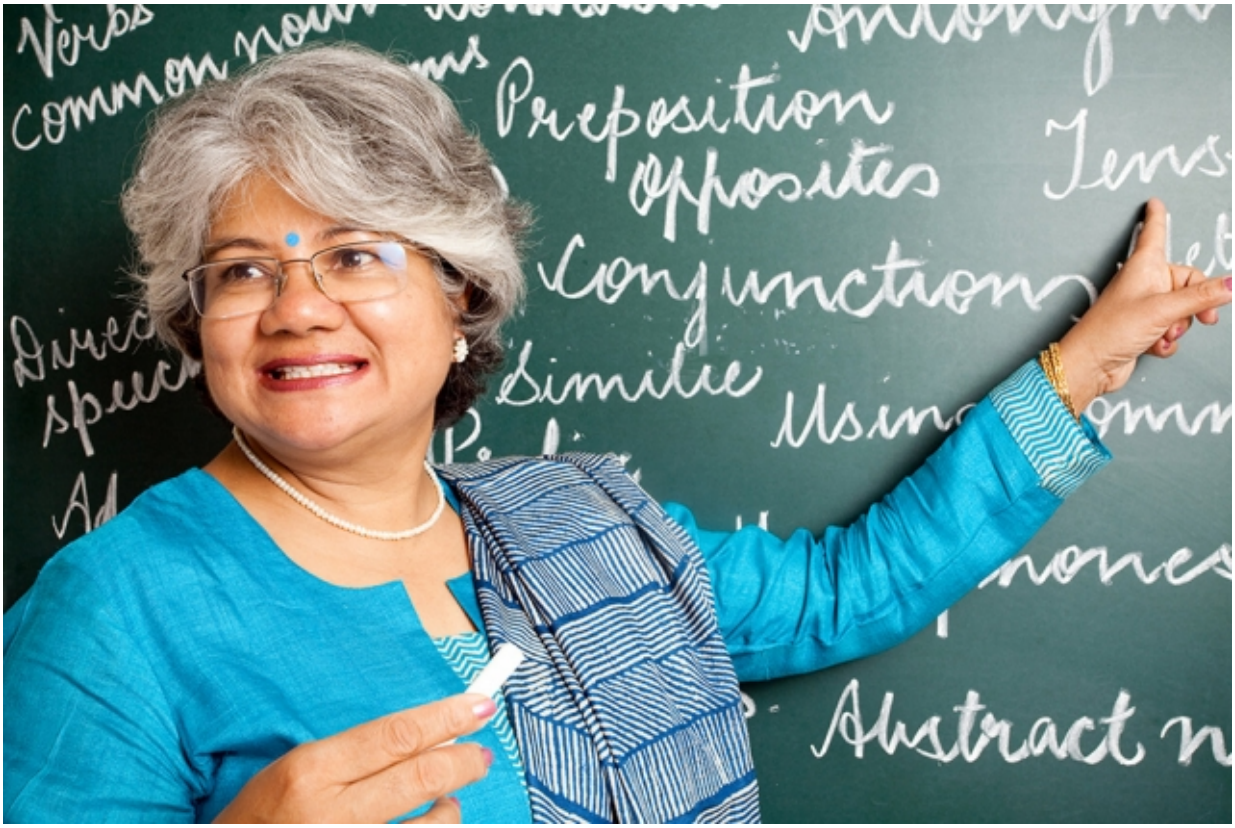
First major database of non-native English

July 29 2016, by Larry Hardesty



English is the most used language on the Internet, and globally most of the people who speak or write in English are non-native speakers. A new database of annotated English sentences written by non-native speakers could help improve how computers handle the spoken or written language of non-native English speakers. Credit: Massachusetts Institute of Technology

After thousands of hours of work, MIT researchers have released the

first major database of fully annotated English sentences written by non-native speakers.

The researchers who led the project had already shown that the grammatical quirks of non-native speakers writing in English could be a source of linguistic insight. But they hope that their dataset could also lead to applications that would improve computers' handling of spoken or written language of non-native English speakers.

"English is the most used language on the Internet, with over 1 billion speakers," says Yevgeni Berzak, a graduate student in electrical engineering and computer science, who led the new project. "Most of the people who speak English in the world or produce English text are non-native speakers. This characteristic is often overlooked when we study English scientifically or when we do natural-language processing for English."

Most natural-language-processing systems, which enable smartphone and other computer applications to process requests phrased in ordinary language, are based on machine learning, in which computer systems look for patterns in huge sets of training data. "If you want to handle noncanonical learner language, in terms of the training material that's available to you, you can only train on standard English," Berzak explains.

Systems trained on nonstandard English, on the other hand, could be better able to handle the idiosyncrasies of non-native English speakers, such as tendencies to drop or add prepositions, to substitute particular tenses for others, or to misuse particular auxiliary verbs. Indeed, the researchers hope that their work could lead to grammar-correction software targeted to native speakers of other languages.

Diagramming sentences

The researchers' dataset consists of 5,124 sentences culled from exam essays written by students of English as a second language (ESL). The sentences were drawn, in approximately equal distribution, from native speakers of 10 languages that are the primary tongues of roughly 40 percent of the world's population.

Every sentence in the dataset includes at least one grammatical error. The original source of the sentences was a collection made public by Cambridge University, which included annotation of the errors, but no other grammatical or syntactic information.

To provide that additional information, Berzak recruited a group of MIT undergraduate and graduate students from the departments of Electrical Engineering and Computer Science (EECS), Linguistics, and Mechanical Engineering, led by Carolyn Spadine, a graduate student in both EECS and linguistics.

After eight weeks of training in how to annotate both grammatically correct and error-ridden sentences, the students began working directly on the data. There are three levels of annotation. The first involves basic parts of speech—whether a word is a noun, a verb, a preposition, and so on. The next is a more detailed description of parts of speech—plural versus singular nouns, verb tenses, comparative and superlative adjectives, and the like.

Next, the annotators charted the syntactic relationships between the words of the sentences, using a relatively new annotation scheme called the Universal Dependency formalism. Syntactic relationships include things like which nouns are the objects of which verbs, which verbs are auxiliaries of other verbs, which adjectives modify which nouns, and so on.

The annotators created syntactic charts for both the corrected and

uncorrected versions of each sentence. That required some prior conceptual work, since grammatical errors can make words' syntactic roles difficult to interpret.

Berzak and Spadine wrote a 20-page guide to their annotation scheme, much of which dealt with the handling of error-ridden sentences. Consistency in the treatment of such sentences is essential to any envisioned application of the dataset: A machine-learning system can't learn to recognize an error if the error is described differently in different training examples.

Repeatable results

The researchers' methodology, however, provides good evidence that annotators can chart ungrammatical sentences consistently. For every sentence, one evaluator annotated it completely; another reviewed the annotations and flagged any areas of disagreement; and a third ruled on the disagreements.

There was some disagreement on how to handle ungrammatical sentences—but there was some disagreement on how to handle grammatical sentences, too. In general, levels of agreement were comparable for both types of sentences.

The researchers report these and other results in a paper being presented at the Association for Computational Linguistics annual conference in August. Joining Berzak and Spadine on the paper are Boris Katz, who is Berzak's advisor and a principal research scientist at MIT's Computer Science and Artificial Intelligence Laboratory; and the undergraduate annotators:Jessica Kenney, Jing Xian Wang, Lucia Lam, Keiko Sophie Mori, and Sebastian Garza.

The researchers' dataset is now one of the 59 datasets available from the

organization that oversees the Universal Dependency (UD) standard. Berzak also created an online interface for the dataset, so that researchers can look for particular kinds of errors, in sentences produced by native speakers of particular languages, and the like.

"What I find most interesting about the ESL [dataset] is that the use of UD opens up a lot of possibilities for systematically comparing the ESL data not only to native English but also to other languages that have corpora annotated using UD," says Joakim Nivre, a professor of computational linguistics at Uppsala University in Sweden and one of the developers of the UD standard. "Hopefully, other ESL researchers will follow their example, which will enable further comparisons along several dimensions, ESL to ESL, ESL to native, et cetera."

"The decision to annotate both incorrect and corrected sentences makes the material very valuable," Nivre adds. "I can see, for example, how this could be cast as a machine translation task, where the system learns to translate from ESL to English. The current corpus would essentially provide the parallel data necessary to train such a system, and the availability of syntactic annotation for both sides opens up more diverse technical approaches."

More information: Universal Dependencies for Learner English. arxiv.org/pdf/1605.04278v1.pdf

Treebank of Learner English: esltreebank.org/

Provided by Massachusetts Institute of Technology