Using an automatic parser as a language learner model

University of Zurich[UZH]

Gerold Schneider, Gintarė Grigonytė
gschneid@ifi.uzh.ch, gintare@ling.su.se

Introduction

NATIVE SPEAKERS perceive and produce semi-preconstructed phrases (Sinclair 1991, Stefanowitsch and Gries 2003). Lexical expectations (Hoey 2005) guide our interpretation, creative and analytic language use is restricted. Native speakers employ argument structures, alternations (Levin 1993), choice of synonyms as subtle operations (Pawley and Syder 1983). Although grammatical variation seems abundant (e.g. Rohdenburg & Mondorf 2003) but is severely restricted by complex, and interacting factors up to being nearly deterministic (Bresnan et al. 2007). Sentences are rendered in the way that they are due to many complex and interacting factors. Failures increase both the human and the automatic processing load up to creating ambiguity.

(1a) Original:	Usually , I go to the library , and I rent these books.
(1b) Corrected:	Usually , I go to the library , and I borrow these books.
(2a) Original:	I am going to the present for my family.
(2b) Corrected:	I am going to buy presents for my family.
(3a) Original:	Kindly and gently computer game I bought for them.
(3b) Corrected:	I bought a harmless computer game for them.

Method

AN AUTOMATIC robust probabilistic parser (Schneider 2008) is used as psycholinguistic model of syntactic and idiomatic expectation. A broad-coverage parser can be a psycholinguistic language model because it:

- predicts attachment decisions from grammar rules & lexical preferences
- has a statistical model that can be extended by any observed factors
- learns from real-word data (e.g. Penn Treebank).
- assigns higher scores to entrenched structures, as they are more expected.

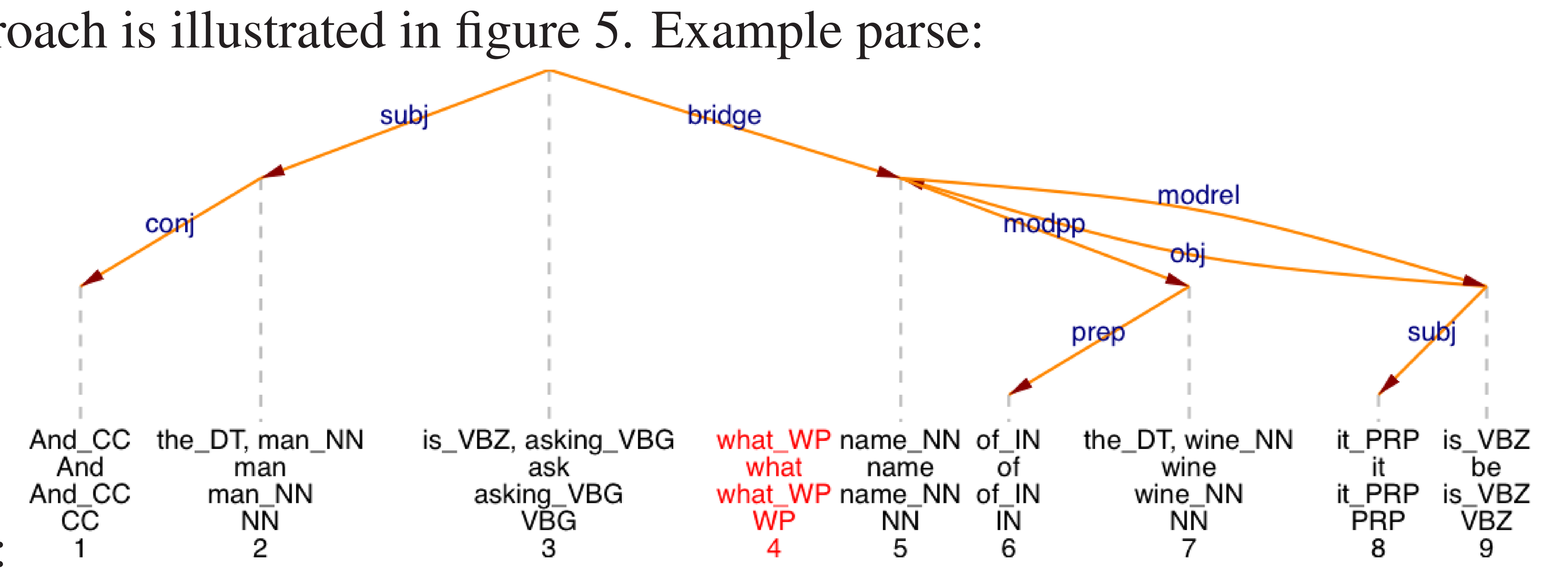
Keller (2010) suggests the use of broad-coverage robust parsers as cognitively plausible models.

Our hypothesis is: L2 utterances do not fit the model very well – equally the human listener and the computational parser model – and thus lead to

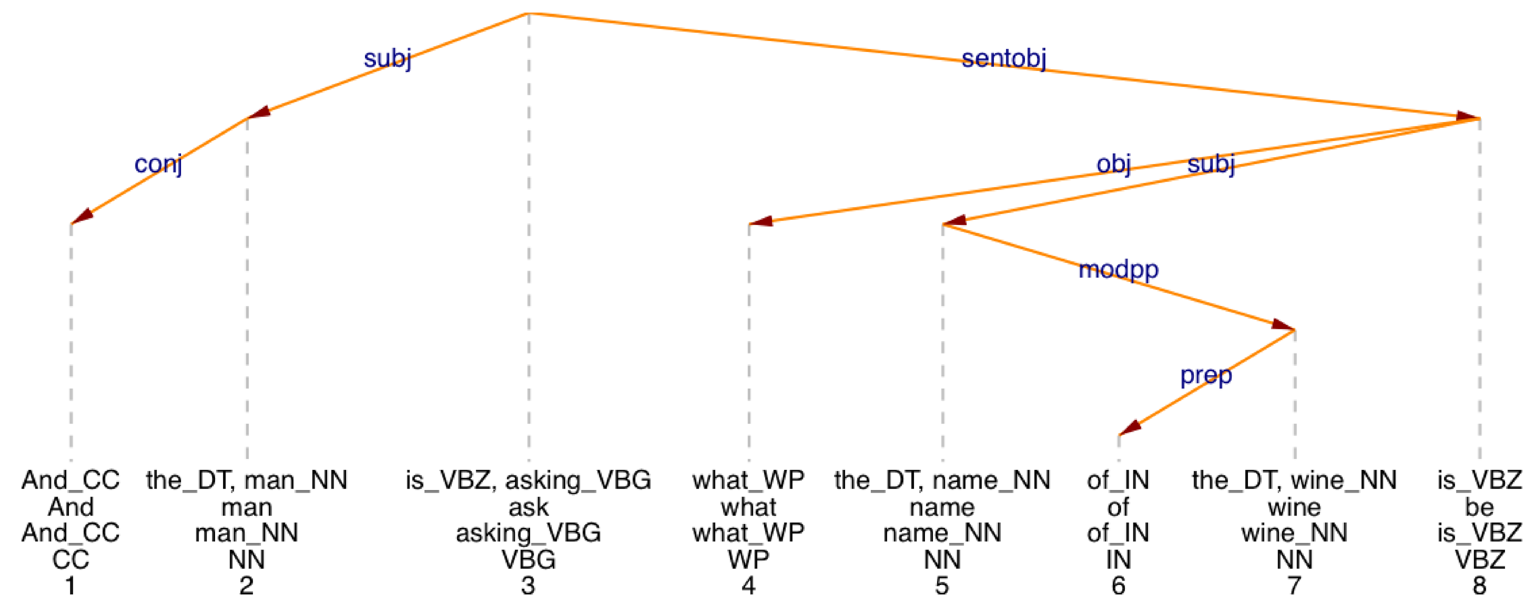
- more parsing errors and
- lower parser scores, in correlation to increased processing times for human listeners.

Our approach is illustrated in figure 5. Example parse:

Original:



Corrected:



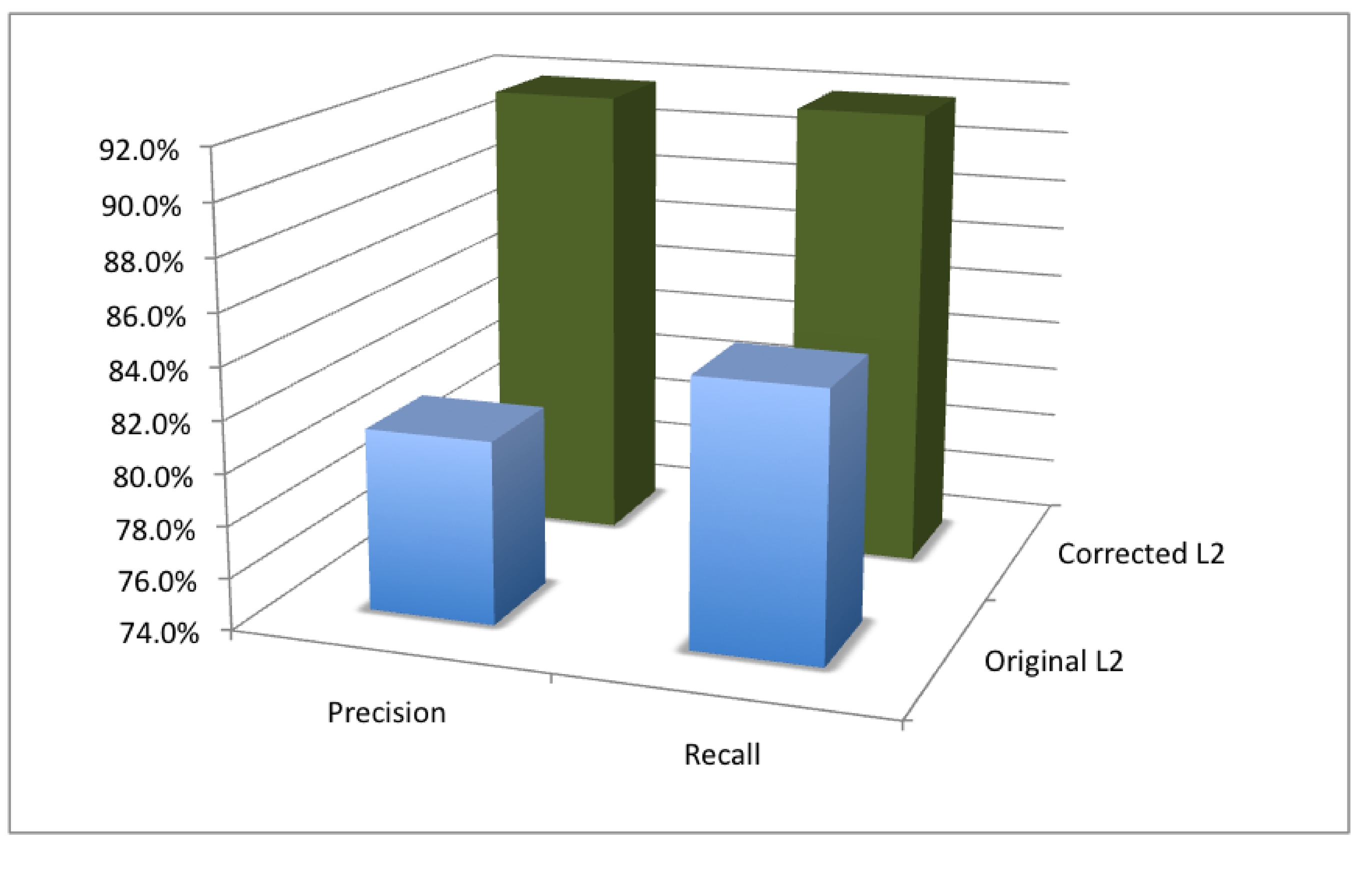


Figure 1: Parser error rate decreases on the corrected text.

Results

WE APPLY the parser to Learner English. We have manually annotated 100 sentence pairs from the NICT Japanese Learner English (JLE) Corpus [http://alaginrc.nict.go.jp/nict_jle/index_E.html]. It contains 120,000 sentence pairs of consisting of an original language learner sentence and a corrected sentence (see (1)-(3)). We show that:

- parser performance is significantly lower for the original Learner data than for the corrected (see Figure 1);
- parser scores are significantly lower for the original Learner data than for the corrected (see Figure 2);
- parse fragmentation is considerably higher for the original Learner data than for the corrected (see Figure 3).

We also tested the uncorrected essays from the CEEAUS (Corpus of English Essays Written by Asian University Students, (Ishikawa, 2009). [http://language.sakura.ne.jp/s/ceeause.html]

- There is a correlation between learner level and parser scores (Figure 4).

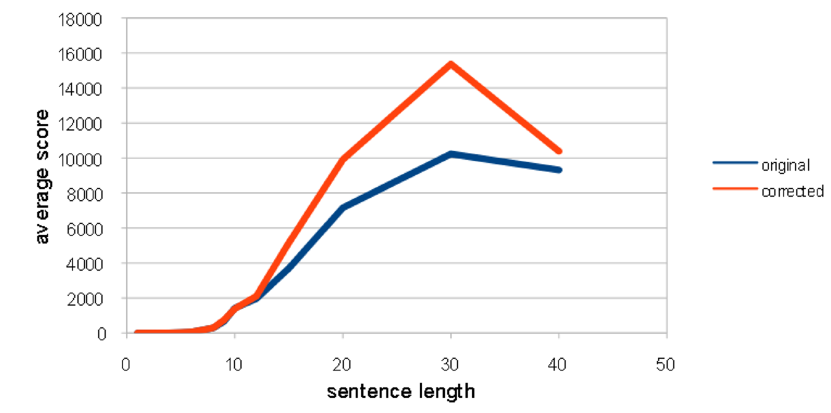


Figure 2: Parser scores, by sentence length.

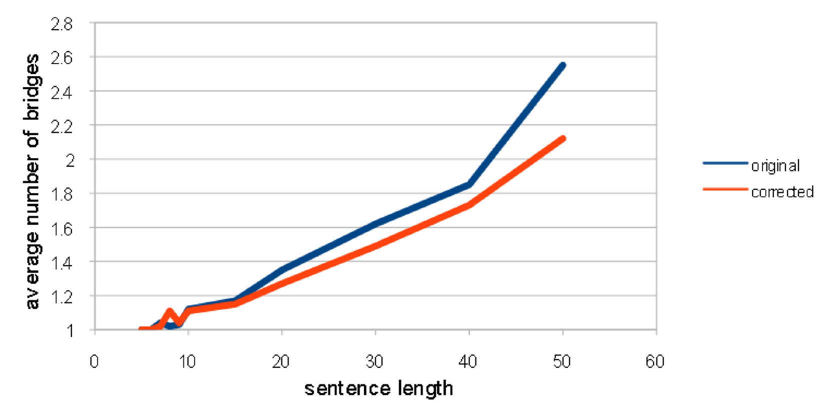


Figure 3: Parse Fragmentation.

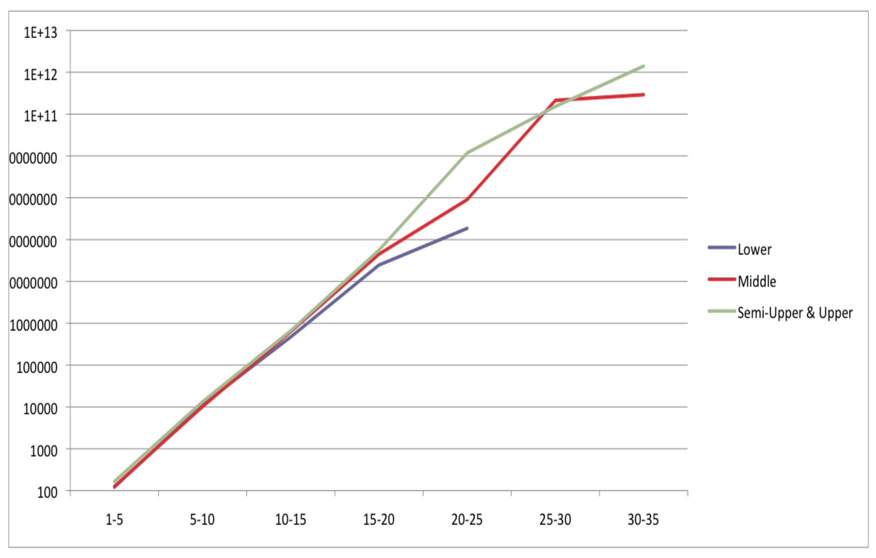


Figure 4: Parser scores, by sentence length, according to learner level in CEEAUS corpus.

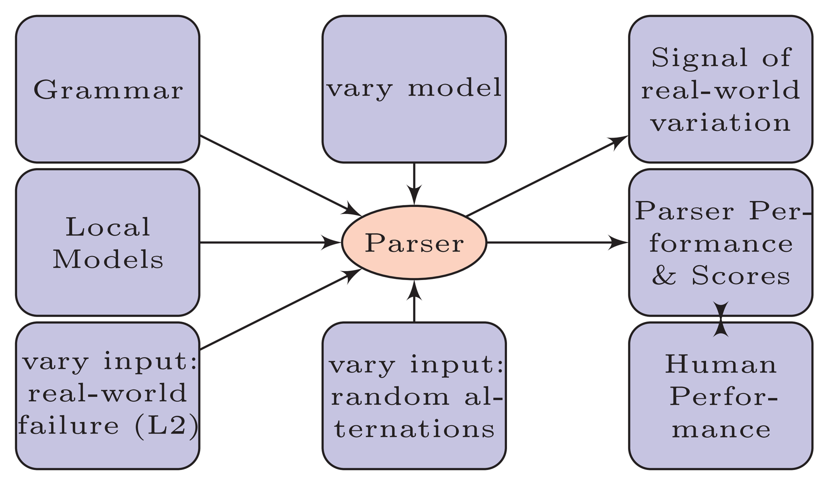


Figure 5: Overview of our approach.

For the investigation of highly gradient, complex and interacting factors a global language model is not just a nice add-on, but an essential base. We plan to use it as a psycholinguistic model in future research, for example to detect learner errors, similar to Gamon (2011) but using more features.

References

Bresnan, Joan, Anna Cueni, Tatiana Nikitina, and Harald Baayen (2007). Predicting the dative alternation. In G. Boume, I. Kraemer, and J. Zwarts, editors, Cognitive Foundations of Interpretation. Royal Netherlands Academy of Science, Amsterdam, pages 69-94.

Gamon, Michael (2011). High-order sequence modeling for language learner error detection. In Proceedings of the Sixth Workshop on Innovative Use of NLP for Building Educational Applications, pages 180-189, Portland, Oregon, June. Association for Computational Linguistics.

Hoey, Michael (2005). Lexical priming: A New Theory of Words and Language. Routledge.

Ishikawa, Shin (2009). Vocabulary in interlanguage: A study on corpus of English essays written by Asian university students (CEEAUS). In K. Yagi and T. Kanzaki, (eds): Phraseology, corpus linguistics and lexicography: Papers from Phraseology 2009 in Japan, pages 87-100, Nishinomiya, Japan. Kwansei Gakuin University Press.

Keller, Frank.(2010). Cognitively Plausible Models of Human Language Processing. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics: Short Papers, 60-67. Uppsala.

Levin, Beth (1993). English Verb Classes and Alternations: A Preliminary Investigation. Chicago: University of Chicago Press.

Pawley, Andrew and Syder, Frances Hodgetts (1983). Two Puzzles for Linguistic Theory: Native-like selection and native-like fluency. In Richards, J. C. & Schmidt, R. W. (Eds.), Language and Communication. London: Longman. 191-226.

Rohdenburg, Guenter & Mondorf, Britta, eds. (2003). Determinants of Grammatical Variation in English, Mouton de Gruyter, Topics in English Linguistics 43.

Schneider, Gerold (2008). Hybrid Long-Distance Functional Dependency Parsing. Doctoral Thesis, Institute of Computational Linguistics, University of Zurich.

Sinclair, John (1991). Corpus, concordance, collocation: Describing English language. Oxford: OUP.

Stefanowitsch, Anatol & Gries, Stefan Th. (2003). Collostructions: investigating the interaction between words and constructions. International Journal of Corpus Linguistics, 209-43.

Acknowledgement

THIS PROJECT is partially supported by the Zurich Center for Linguistics http://www.linguistik.uzh.ch