A systematic study of the rate of child over-irregularization errors

Vanna Willerton, Graham Adachi-Kriege (McGill University), Shijie Wu (Johns Hopkins University), & Ryan Cotterell (University of Cambridge)
savanna.willerton@mail.mcgill.ca

Overregularization errors, where a regular inflection is applied to an irregular verb, as in *teach/teached, fly/flied,* or *go/goed*, is a well-documented and intensely studied phenomenon in English language acquisition [1-3]. Conversely, *over-irregularization errors*, the overapplication of an irregular inflection to a regular verb like *trick/truck*, or the application of an incorrect irregular inflection to a non-target irregular verb, like *shake/shade,* or *bring/brang*, have been studied less thoroughly. This may be due to the difficulty of detecting over-irregularization errors in databases of child speech. Regularization can be detected simply by adding *-ed* (or +/d/) to irregular English verb stems, while searching for irregulars requires more complex methodology.

The current study is a pilot project testing the feasibility of a novel methodology for estimating the true rate of such over-irregularization errors in child English. An endeavour of this kind has not been attempted in over 20 years (since Xu and Pinker [4]), and there has since been a large increase in available data, new technologies for querying, as well as novel statistical methods for estimating the rate of rare events.

In order to find instances of over-irregularization errors, we searched the North American English databases of the CHILDES corpus [5], a large dataset of child speech, which includes nearly 47,000 instances of verbs used by children. From a list of existing irregular verbs for English past tense we generated a comprehensive set of context dependent rewrite rules and applied these to all monosyllabic English verbs using Pynini [6], a library for compiling a grammar of strings, regular expressions, and context-dependent rewrite rules into weighted finite-state transducers. The resulting list contained possible irregularizations of English verb forms. The rewrite rules required phonetic spelling of the words. The phonetic forms resulting from the application of rewrite rules were transcribed into regular orthography using a sequence-to-sequence phoneme-to-grapheme model for character level transduction [7]. In order to avoid under generating forms due to spellings, we also manually transcribed the forms to get up to five next-best spellings. The resulting list was used to search through CHILDES for matches.

Based on our pilot estimates, there are between 146 and 387 cases of irregularization errors in the North American English section of CHILDES. These findings give rates which are on the same order of magnitude as the similar 1995 Xu and Pinker study. The lower bound indicates clear cases of misuse of irregular inflection to regular or non-target irregular verbs. The upper bound includes mostly no-change (by analogy with *bet/bet,* or *put/put*) irregular matches for which it is unclear from the utterance context whether the verb is an irregularization error or simply an unmarked verb form. A common type of error which we did not search for in this pass is incorrect *-en/-n* suffixations for past participles, for example, erroneously using *shooten* as the past participle of *shot*. Additionally, future iterations of this work will look at per-child verb usage to ensure that over-irregularization errors are not being counted for children who may be in a stage of development where they fail to mark tense at all. If we remove all matches where no other verb in the utterance was used with tense marking then our lower bound is reduced to 102 cases of over-irregularization errors. This lower bound estimate is likely too conservative as many of these utterances contained only a single verb. Further within-participant investigation is necessary to increase the precision of our estimation.

References

[1] Marcus, G., Pinker, S., Ullman, M. T., Hollander, M., Rosen, J., and Xu, F. (1992). *Overregularization in language acquisition*. Monographs of the Society for Research in Child Development. University of Chicago Press, Chicago.

[2] Pinker, S. (1995). Why the child holded the baby rabbit: A case study in language acquisition. In Gleitman, L. R. and Liberman, M., editors, *An invitation to cognitive science, Vol. 1: Language*, pages 107–133. MIT Press, Cambridge, MA.

[3] Maslen, R., Theakston, A. L., Lieven, E. V., and Tomasello, M. (2004). A dense corpus study of past tense and plural overregularization in English. *Journal of Speech, Language and Hearing Research*, 47(6):1319–1333.

[4] Xu, F., & Pinker, S. (1995). Weird past tense forms. *Journal of Child Language, 22*(3), 531-556.

[5] MacWhinney, B. (2000). *The CHILDES Project: Tools for analyzing talk. Third Edition.* Mahwah, NJ: Lawrence Erlbaum Associates

[6] Gorman, Kyle. 2016. Pynini: A Python library for weighted finite-state grammar compilation. In *Proceedings of the ACL Workshop on Statistical NLP and Weighted Automata*, 75-80.

[7] Shijie Wu, Pamela Shapiro and Ryan Cotterell. Hard Non-Monotonic Attention for Character-Level Transduction. EMNLP. 2018