Comparison of a Morphological Learner to Child Acquisition of the Turkish Aorist

Kaan Bayar[β]

Linguistics Department, Bogazici University, Istanbul, Turkey

[β]kaan.bayar13@gmail.com

This research investigates a supervised morphological learning model to find out whether the model is behaving similarly to child acquisition data. Using the supervised morphological rule learner model by Albright & Hayes (2002), the model's ability to generalize Turkish morphology despite the complexities of vowel harmony and irregular forms is tested. Additionally, by manipulating the parameters included in the model, the research aims to pinpoint crucial parameters influencing rule learning. The output is compared to child acquisition data on the aorist case (Nakipoğlu & Ketrez, 2006). The findings are expected to contribute to a better understanding of how rule learning mechanisms interact with complex phonological and morphological constraints and explore the model's similarities to child language acquisition data.

Vowel harmony is a regular process in Turkish. It affects two contrastive features: {±round, ±back}. In principle, these features are underspecified for vowels of the affixes, and the values for these features spread rightwards from the stem-final syllable to all underspecified vowels. However, non high morphemes, i.e., A-type affixes, only accept [±back] harmony.

Vowel harmony has an effect on the morphology of Turkish, namely, on the realization of the morphemes as concrete pronounceable forms. The morphemes in Turkish can be split into two types: I-type and A-type. For example, the plural marker in Turkish is A-type (-lar, -ler), and the accusative case is I-type (-ı, -i, -u, -ü)[1]. While affixes in Turkish are categorized into either A-type or I-type, Turkish aorist has both affix types (-ar, -er, -ır, -ir, -ur, -ür). This marker has irregular forms and lexically conditioned allomorphic forms. This makes it challenging for children to acquire this form. This study uses Nakipoglu and Ketrez (2006) as a child language acquisition dataset to compare the model's performance. Nakipoglu and Ketrez (2006) claim that overregularization happens when a regular rule is extended to an irregular form, and irregularization happens when a regular form is treated as an exception and used with the irregular rule. Nakipoglu and Ketrez (2006) claim that children mostly make errors when a monosyllabic verb ends with a sonorant segment. As they start to encounter more verbs ending with -Ir, their error rates on -Ar ending verbs decrease, which can be explained by their irregularization patterns switching from preferring the -Ar form to the -Ir form. The results suggest that the model approximates the early language acquisition stage, where form selection is heavily influenced by lexical frequency rather than phonological or morphophonological rules. The baseline model was only able to correctly generate 4 out of 43 roots. After getting the baseline results, syllable information is introduced to the model. Because the aorist marker has two types of morphemes, the marker is fully rule-governed in both multisyllabic and vowel-ending roots[2] (Nakipoglu & Michon, 2020). This addition has removed the errors on rule-governed forms. The model with the syllable information correctly generated 25 out of 43 roots. More importantly, including the syllable information enabled the model to learn all rule-governed forms, while the lexically conditioned allomorphic forms could not be captured.

Moreover, the baseline model does not behave similarly to humans as the model always chooses the rule with the highest confidence. However, the implementation of syllable information made the model behave somewhat similarly to early language acquisition with respect to the regularization pattern. This means that the model makes regularization errors based on the dominant form in the

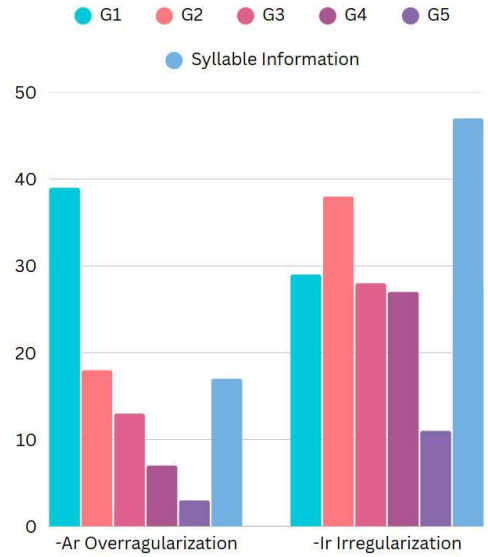
[1] [ɪ] is a high back unrounded vowel (ɯ), and [ü] is a high front rounded vowel (y).

[2] Light verb constructions with 'et', such as aff+et 'forgive', uses -Ar, but the rest of the multisyllabic verbs are used with -Ir (Nakipoglu and Michon, 2020).

lexicon. Figure 1 below shows the irregularization and overregularization rates of the baseline model, the model with the syllable information, and child language acquisition data (Nakipoglu & Ketrez, 2006).

Figure 1

Comparison of the regularization rate of the model to the groups in Nakipoglu & Ketrez (2006)



Note. Regularization error rates of groups (Nakipoglu & Ketrez, 2006, p. 406). The x-axis shows the regularization type, and the y-axis shows the error rate (%).

As shown in Figure 1, the baseline model does not behave similarly to humans as the model sticks to one morpheme; however, syllable information implementation made the model behave somewhat similarly to G1 with respect to the regularization pattern. Nakipoglu and Ketrez (2006) argue that children hear mostly monosyllabic forms that use -Ar, which explains the high overregularization rate. The dataset used to train the model is -Ir heavy with 130 -Ir, 61 -Ar, and 43 -r occurrences. Considering the dataset, the model makes overregularization errors based on the dominant form in the lexicon, i.e., dataset.

Future research will explore implementing corrective feedback. Since corrective feedback has a long-term effect on language acquisition (Hiller & Fernández, 2016), implementing this might show similarities towards later stages of acquisition. Additionally, adjusting the dataset's distribution to align with naturalistic language acquisition patterns, as described in Nakipoğlu & Ketrez (2006), is needed to justify the idea that the model outputs similar results to those of children.

The model captured general morphophonological rules for Turkish, albeit with challenges in handling exceptions, particularly with monosyllabic verbs ending in sonorants. The findings indicate that the model's error patterns, characterized by overregularization and irregularization, parallel those observed in child language acquisition data, underscoring the influence of lexical frequency on morphological generalizations.

Keywords: Computational Linguistics, Morphological Learning, Language Acquistion

References

Albright, A., & Hayes, B. (2002). Modeling English Past Tense Intuitions with Minimal Generalization. *Proceedings of the ACL-02 Workshop on Morphological and Phonological Learning*, 58–69. https://doi.org/10.3115/1118647.1118654

Hiller, S., & Fernández, R. (2016). A Data-driven Investigation of Corrective Feedback on Subject Omission Errors in First Language Acquisition. In S. Riezler & Y. Goldberg (Eds.),

Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning (pp. 105–114). Association for Computational Linguistics. https://doi.org/10.18653/v1/K16-1011

Nakipoglu, M., & Ketrez, F. (2006). *Children's overregularizations and irregularizations of the Turkish Aorist*.

Nakipoglu, M., & Michon, E. (2020). Abstraction vs. Analogy in the Turkish aorist. In A. Gürer, D. Uygun-Gökmen, & B. Öztürk (Eds.), *Studies in Language Companion Series* (Vol. 215, pp. 14–38). John Benjamins Publishing Company. https://doi.org/10.1075/slcs.215.01nak