Modeling Morphological Learning: Tolerance Principle on Turkish past tense -DI

Rüveyda Şahyar • Boğaziçi University • ruveyda.sahyar@std.bogazici.edu.tr

Universität Konstanz

Background

Turkish past tense suffix -DI has 8 allomorphs conditioned by voice assimilation and vowel harmony with regards to frontness and roundness [8]. Vowel harmony and voice assimilation rules apply consistently for all verbal stems.

(1)	gel	-di	(3)	ısır	-dı	(5)	oku	-du	(7)	gör	-dü
	come	-DI		bite	-DI		read	-DI		see	-DI
	They(sg.) came.			They(sg.) bit.			They(sg.) read.			They(sg.) saw.	
(2)	git	-ti	(4)	yap	-tı	(6)	somurt	-tu	(8)	düş	-tü
	go	-DI		do	-DI		frown	-DI		fall	-DI
	They(sg.) went.			They(sg.) did.			They(sg.) frowned.			They(sg.) fell	

-DI is used productively by Turkish-acquiring children as early as 1;5 of age with very little error [1][2]. Even the least frequent form of the morpheme to appear in the combined corpus for this study is reportedly used by a Turkish-acquiring child as young as 1;3 with less than seven verbs in their speech [2].

Tolerance Principle

Tolerance Principle

Let R be a rule applicable to N items, of which e are exceptions. R is productive if and only if

$$e \leq \theta N \text{ where } \theta N := N / \ln N \qquad [16]$$

Abduction of Tolerable Productivity: A greedy search algorithm that recursively generates a decision tree based on Tolerance Principle. [5]

The Present Study: tests ATP model on the rule-based allomorphy of the Turkish morpheme -DI.

- The suffix has many allomorphs that are completely rule-driven and even the least frequent form is acquired very early.
- TP formulates rules that minimize the number of exceptions. A rule defined over a small set is more 'tolerant', more learnable.

Hypothesis: The model should be successful with these phonologically conditioned allomorphs that exhibit no irregularity.

Claim: The challenges of the data are such that they evaluate ATP's ability to learn complex yet regular rules with limited occurrence.

References

[1] Aksu-Koç, A. & Slobin, D.I. (1985). The acquisition of Turkish. In The Crosslinguistic Study of Language Acquisition, Vol. 1, D.I. Slobin (ed.), 839–878. Hillsdale NJ: Lawrence Erlbaum Associates.

[2] Aksu-Koç, A. & Ketrez, F.N. (2003). Early verbal morphology in Turkish: Emergence of inflections. In Mini-paradigms and the Emergence of Verb Morphology, D. Bittner, W.U. Dressler & M. Kilani-Schoch (eds), 27–52. Berlin: Mouton de Gruyter.

[3] Albright, A., & Hayes, B. (2003). Rules vs. analogy in English past tenses: A computational/experimental study. Cognition, 90(2), 119–161. DOI: 10.1016/s0010-0277(03)00146-x

[4] Altinkamis, F. (2003). CHILDES Database Turkish Altinkamis Corpus. DOI: 10.21415/T5H89W

[5] Belth, C., Payne, S., Beser, D., Kodner, J., & Yang, C. (2021). The Greedy and Recursive Search for Morphological Productivity. Proceedings of the Annual Meeting of the Cognitive Science Society, 43(43).

[6] Berko, J. (1958). The Child's Learning of English Morphology, WORD, 14(2-3), 150-177, DOI: 10.1080/00437956.1958.11659661.

[7] Bird, S., Klein, E., & Loper, E. (2009). Natural language processing with Python. O'Reilly. www.nltk.org/book/

[8] Göksel, A., & Kerslake, C. (2005). Turkish: A comprehensive grammar. London: Routledge.

[9] Kuzgun, A., Cesur, N., Arıcan, B. N., Özçelik, M., Marşan, B., Kara, N., Aslan, D. B., & Yıldız, O.T. (2020). On building the largest and cross-linguistic Turkish dependency corpus. In 2020 Innovations in Intelligent Systems and Applications Conference (ASYU), 1-6. IEEE.

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

[11] Nakipoğlu, M., Uzundağ, B. A., & Ketrez, F. N. (2023). Analogy is indispensable but rule is a must: Insights from Turkish. Journal of Child Language, 50(2), 437–463. doi:10.1017/S0305000921000921

[12] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. In Journal of Machine Learning Research 12(85), 2825–2830.

[13] Slobin, D. (1982). Universal and particular in the acquisition of language. In E. Wanner & L. Gleitman (Eds.), Language acquisition: The state of the art, 128-172. New York: Cambridge University Press. DOI: 10.21415/T5SP6C

[14] Štěpánek, J. & Pajas, P. (2010). Querying Diverse Treebanks in a Uniform Way, in Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10), European Language Resources Association (ELRA), 1828-1835. Valletta: Malta.

[15] Straka, M. (2018). UDPipe 2.0 Prototype at CoNLL 2018 UD Shared Task. In Proceedings of CoNLL 2018: The SIGNLL Conference on Computational Natural Language Learning, 197-207, Association for Computational Linguistics, Stroudsburg, PA: USA.

[16] Yang, C. (2016). The price of linguistic productivity: How children learn to break the rules of language. MIT Press.

Methods

Seven experiments were conducted where ATP model [5] was trained on a combined corpus of 751 Turkish verbs inflected with -DI. Each experiment isolated phonologically conditioned allomorphs of the morpheme.

The train/test split was done using sklearn [12] with 563 verbs for training and 188 for testing.

Data

- 328 verbs from child-produced and child-directed speech from CHILDES Turkish corpora [4] [13] extracted using UDPipe 2.0 [15]
- 900 most frequent verbs in Universal Dependencies Turkish Penn 2.10 Treebank [9] queried through PML Tree Query [14]
- combined by removing overlapping instances, then inflected using a Context-Free Grammar with NLTK [7]

Form	Features	Occurrence
-dı	[+VOICE] [+BACK] [-ROUND]	270
-di	[+VOICE] [-BACK] [-ROUND]	180
-tı	[-VOICE] [+BACK] [-ROUND]	89
-ti	[-VOICE] [-BACK] [-ROUND]	75
-du	[+VOICE] [+BACK] [+ROUND]	55
-dü	[+VOICE] [-BACK] [+ROUND]	35
-tu	[-VOICE] [+BACK] [+ROUND]	25
-tü	[-VOICE] [-BACK] [+ROUND]	22
Total		751

Evaluation

- Precision, recall and F1 calculations on the test data
- wug-test [6] of 8 nonce words
- Decision trees provided by the model for explicit analysis of formulated rules.

Results

	Features	Precision	Recall	F1
Experiment 1	[+/-VOICE]	1.0	1.0	1.0
Experiment 2	[+/- BACK]	0.955539	0.934803	0.943099
Experiment 3	[+/- ROUND]	0.734524	0.650497	0.675638
Experiment 4	[+/-VOICE] [+/- BACK]	0.951042	0.942859	0.946500
Experiment 5	[+/-VOICE] [+/- ROUND]	0.867888	0.777437	0.805699
Experiment 6	[+/- BACK] [+/- ROUND]	0.906071	0.891674	0.893532
Experiment 7 (Turkish forms)	[+/-VOICE] [+/- BACK] [+/- ROUND]	0.883886	0.888727	0.880219

Metrics for experiments

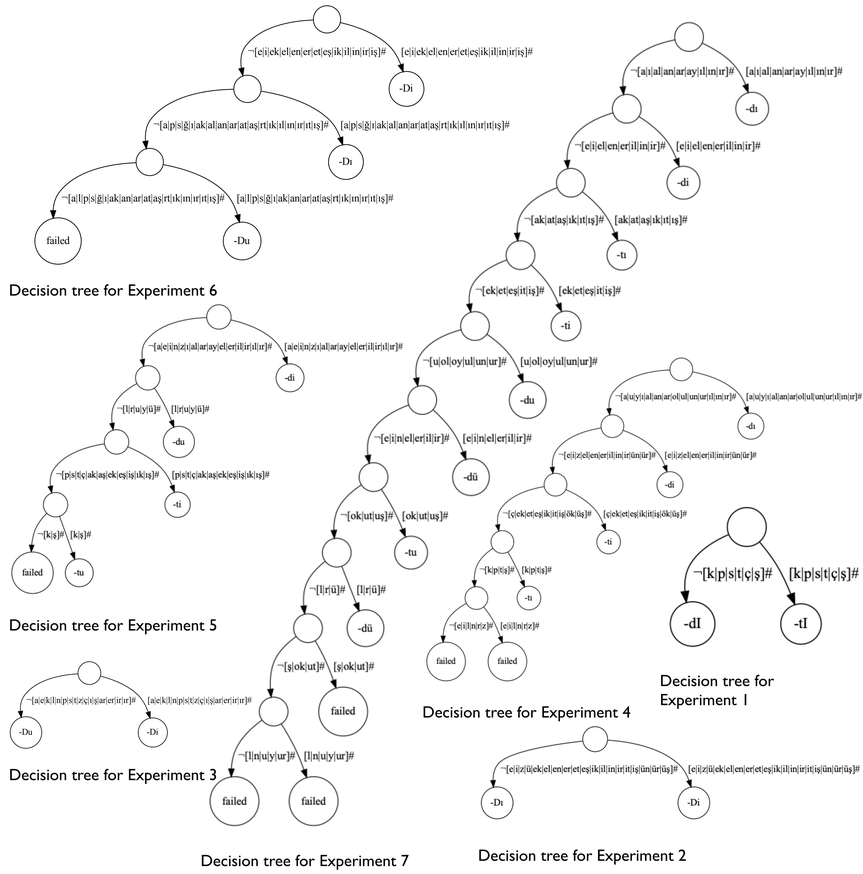
		Output							
		dı	di	tı	ti	du	dü	tu	tü
Expected	dı	57	0	0	0	0	0	0	0
	di	0	43	0	0	0	0	0	0
	tı	0	0	31	3	0	0	0	0
	ti	0	1	1	17	0	0	0	3
	du	0	0	0	0	9	3	0	0
	dü	1	0	0	0	0	9	0	1
	tu	0	0	0	0	0	0	2	0
	tü	0	0	0	0	0	1	0	6

Confusion matrix for Experiment 7

	nonce words							
	apıt	lemir	şatır	kutul	şöpür	deriş	kuruf	pülüs
Experiment 1	apıttI	lemirdI	şatırdI	kutuldI	şöpürdI	derişti	kurufdI	pülüstI
Experiment 2	apıtDı	lemirDi	şatırDı	kutulDı	şöpürDi	derişDi	kurufDı	pülüsDı
Experiment 3	apıtDi	lemirDi	şatırDi	kutulDi	şöpürDu	derişDi	kurufDu	pülüsDi
Experiment 4	apıttı	lemirdi	şatırdı	kutuldı	şöpürdi	derişti	kuruftı	pülüsti
Experiment 5	apıtti	lemirdi	şatırdi	kutuldu	şöpürdu	derişti	kurufdu	pülüsti
Experiment 6	apıtDı	lemirDi	şatırDı	kutulDu	şöpürDü	derişDi	kurufDu	pülüsDı
Experiment 7	apıttı	lemirdi	şatırdı	kutuldu	şöpürdü	derişti	kuruftu	pülüstü

wug-tests for each nonce word

Results



Discussion

- Decision trees given by ATP fail to capture the allomorphy rules of -DI, especially in experiments where roundness is tested. However, given that this is not observed in the metrics, one could argue that it captures children's productive use of the morpheme.
- ATP is a rule-based model. Previous acquisition literature on Turkish verbal morphology suggest analogy effects based on token frequency in children's production errors. Token frequency of sequences with roundness harmony (incl. those outside the verbal domain) were found to be higher than sequences without it (uru > ura) in a developmental corpus [11]. ATP's shortcomings, especially with regards to roundness harmony, might be because an analogy-based process is involved in the acquisition of this morpheme.
- ATP tests the final segment of a lemma for a given suffix against the Tolerance Principle, then the final two segments in case it is not productive under TP, and so forth. The allomorph of -DI for a given verb can depend on as much as three final segments.
- At worst, ATP would have to consider, 21 consonants in Turkish orthography + 21 x 21 + 8 vowels in Turkish orthography x 21 x 21 = 3990 possible rules for each allomorph.
- Considering this, the acquisition of these different forms might simply require abstraction. Under feature theory, a Turkish-acquiring child would be able to generalize over features and natural classes, which ATP is unable to do.