

Learning inflection

Commentary on [Michael Ramscar \(2021\)](#). “A discriminative account of the learning, representation and processing of inflection systems.” In: *Language, Cognition and Neuroscience*

Olivier Bonami

`olivier.bonami@u-paris.fr`

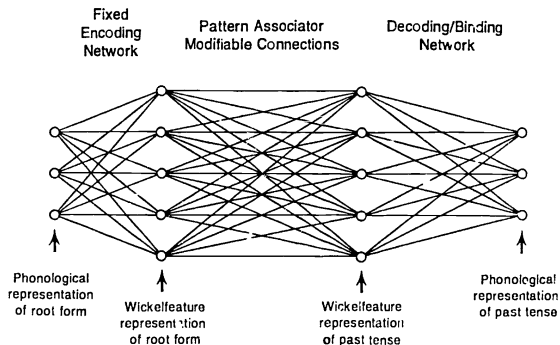
Université de Paris — Lexical Matters — 16/02/2022

Words and rules

- ▶ Traditional vision of learning inflection: general rules + lists of exceptions
- ▶ Early psycholinguistic research (e.g. Miller 1967) took the productivity of regular morphology as an indication of innate knowledge of linguistic structure.
 - ▶ If all speakers do is reproduce patterns of cooccurrence they have already encountered, how can they inflect novel lexemes?
 - ▶ Suggests speakers are attempting to acquire abstract rules.
- ▶ In addition, two early intriguing results:
 - ▶ In wug tests (Berko, 1958), speakers readily extend irregular patterns if wugs are similar enough to existing irregulars (Bybee & Slobin 1982).
 - ▶ U-shaped learning of inflection: as vocabulary grows, drop in accuracy followed by new gain (Brown 1996): first *mice*, then *mouses*, then *mice*.

Rumelhart & McClelland

- ▶ Rumelhart and McClelland (1986) argue that all three observations (productivity, subregularity, u-shape) can be accounted for without positing rules.
- ▶ Basic idea: speaker behavior simply follows from the statistical structure of the data they are exposed to. Regular inflection is more frequently used because it is more frequent in the data.
- ▶ Implementation based on a simple, early 1-layer neural network.
 - ▶ Arguably the most spectacular early success of using neural networks in cognitive science.



The dual route alternative

- ▶ This led to a fury of criticism and new research, initially mostly led by Steven Pinker and colleagues (Pinker and Prince, 1988; Marcus et al., 1992; Prasada and Pinker, 1993; Marcus et al., 1995; Pinker, 1999).
- ▶ Main criticisms:
 - ▶ Rumelhart and McClelland's U-shaped learning results from a specific and unrealistic timing of training data.
 - ▶ Identification of various phenomena not predicted by Rumelhart and McClelland's approach but widely attested.
 - ▶ Regular inflection is used "by default" for borrowings, compounds, etc., even on strings that are undistinguishable from known irregulars: *low-lifes*, not **low-lives*, etc.
 - ▶ Irregular inflected forms are lexical unites, not regular ones: *mice-eater* vs. *rats-eater*, etc.
- ▶ Main innovation: dual-route learning and processing
 - ▶ Regulars are derived by rule.
 - ▶ Irregulars are stored in an associative memory akin to Rumelhart and McClelland's network.

The role of context I

- ▶ Pinker and colleagues argued on the basis of homophones (e.g. *brake* vs. *break*) that phonology was not enough to predict correct past tenses (Pinker & Prince 1988), but independently that semantics information played no role, and would actually get in the way, as irregulars sharing a pattern typically do not share semantic features (*sing*, *sting*, *swing*, etc.)
- ▶ In early work, Ramscar (2002) showed that speaker's willingness to use an irregular pattern is affected by semantic similarity:
 1. In a traditional spring rite at Moscow University Hospital, the terminally ill patients all *frink* in the onset of good weather, consuming vast quantities of vodka and pickled fish. In 1996, his favorite vodka glass in hand, cancer patient Ivan Borovich ----- around 35 vodka shots and 50 pickled sprats; it is not recorded whether this helped in his treatment.
 2. In a classical symptom of Howson's syndrome, patients all *frink* in their right eye if they are left handed or left eye if right handed, their eyelids opening and closing rapidly and uncontrollably. In 1996, in extreme discomfort due to his bad eye, Howson's patient Ivan Borovich ----- around 35 times per minute for two days, causing severe damage to the muscles in his left eyelid.

The role of context II

- ▶ In addition, this still holds if the verbs are clearly presented as denominal, a context where Pinker and colleagues predict that regularity should prevail.
 1. A **frink** is the Muscovite equivalent of the Spanish tapas; it is served in bars, and usually comprises chilled vodka and some salted or pickled fish. In a traditional spring rite at Moscow University Hospital, the terminally ill patients all **frink** in the onset of good weather, consuming vast quantities of vodka and pickled fish. In 1996, his favorite vodka glass in hand, cancer patient Ivan Borovich ----- around 35 vodka shots and 50 pickled sprats; it is not recorded whether this helped in his treatment.
 2. The **frink** is the common name for the motor muscle that controls the opening of the eyelid. It is especially prone to neurological interference. In a classical symptom of Howson's syndrome, patients all **frink** in their right eye if they are left handed or left eye if right handed, their eyelids opening and closing rapidly and uncontrollably. In 1996, in extreme discomfort due to his bad eye, Howson's patient Ivan Borovich ----- around 35 times per minute for two days, causing severe damage to the muscles in his left eyelid.
- ▶ Relatedly, Ramscar and Dye (2011) provide evidence that the acceptability of **rats-eater* type examples is affected by manipulating the semantic context.

What is to be modelled?

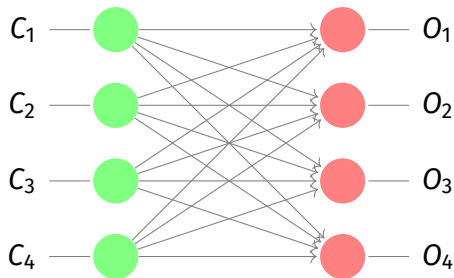
- ▶ Ramscar (2021)'s central argument:
 - ▶ Rumelhart and McClelland were doing one thing right: modelling learning using discriminative networks.
 - ▶ Both Rumelhart and McClelland and Pinker and colleagues are doing one thing wrong: focusing on an unrealistic learning task.
- ▶ Children are not exposed to pairings of stems and inflected forms, they are exposed to inflected forms in context.
- ▶ Ramscar suggests that Speaker's behavior in wug tests (at any age) is a **consequence** of what they have learned, but it is not the learning objective.
 - ▶ The learning objective is understanding and using individual words correctly.

Discriminative learning

- ▶ Three different but related senses of *discriminate*:
 1. In animal learning: **discrimination learning** is learning to associate different responses with different stimuli (Rescorla & Wagner, 1972)
 2. In machine learning: a **discrimination model** is a machine learning model attempting to maximize $P(\text{target} \mid \text{predictor})$ (Ng & Jordan, 2002)
 3. In human learning: **discriminative learning** names the family of error-driven learning models, including neural networks and cognitive models grounded in Rescorla & Wagner's (1972) equations (Ramscar et al. 2010)
- ▶ For 15 years Ramscar has advocated that the Rescorla-Wagner model of behavioral conditioning:
 - ▶ Should be seen conceived as **discriminative** rather than **associative**, as it relies crucially on learning from errors and not just remembering associations.
 - ▶ Grounds human learning of language.
 - ▶ Explains various properties of human language.

The Rescorla-Wagner model I

- ▶ We are interested in finding out how agents produce **outcomes** when exposed with **cues**.
- ▶ This is modelled by a two-layer fully connected network, where weights on the edges indicate how strongly a given cue signals an outcome.



- ▶ Learning the network is learning to adjust the weight matrix so that the right cues predict the right outcomes.

The Rescorla-Wagner model II

- ▶ Simplified formulation of the Rescorla-Wagner equations Chuang and Baayen (2021):
Given a network relating cues C_i to outcomes O_i with weights w_{ij} , a learning event at time t leads to the following update:

$$w_{ij}^{t+1} = w_{ij}^t + \Delta w_{ij}^t, \text{ where}$$

$$\Delta w_{ij}^t = \begin{cases} 0 & \text{if ABSENT}(C_i, t) \\ \alpha \left(1 - \sum_{\text{PRESENT}(C_k, t)} w_{kj} \right) & \text{if PRESENT}(C_i, t) \text{ and PRESENT}(O_j, t) \\ \alpha \left(0 - \sum_{\text{PRESENT}(C_k, t)} w_{kj} \right) & \text{if PRESENT}(C_i, t) \text{ and ABSENT}(O_j, t) \end{cases}$$

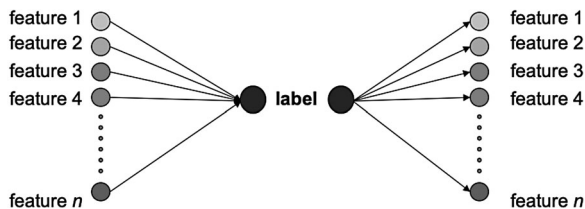
with the learning rate α small.

The Rescorla-Wagner model III

- ▶ In other words, when the learner witnesses a set of cues together with a set of outcomes:
 - ▶ The link of these cues with these outcomes is upgraded
 - ▶ The link of these cues with these outcomes is downgraded
 - ▶ The extent of this upgrading/downgrading is a function of the total previous weights of all presently seen cues.
- ▶ Note the (non-coincidental) similarity with backpropagation in neural networks.

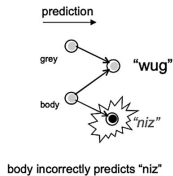
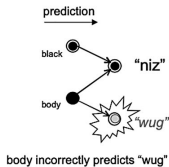
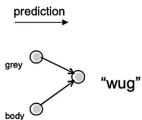
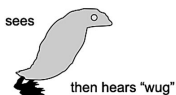
Learning a word I

- ▶ Ramscar et al. (2010) apply the RW model to learning a word, i.e. how a unique discrete label relates to a large set of features in the context.



- ▶ They argue that it is crucial for learning to go from features to label (i.e. from meaning to form).

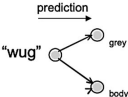
Learning a word II



Cue	Outcome	S0	S1	S2	S3
grey	wug	0	0.1	0.1	0.181
grey	niz	0	0	0	-0.01
black	wug	0	0	-0.01	-0.01
black	niz	0	0	0.1	0.1
bird	wug	0	0.1	0.09	0.171
bird	niz	0	0	0.1	0.09

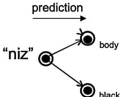
Learning a word III

hears "wug"
then sees



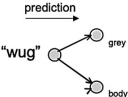
"wug" correctly predicts *red* and *body*

hears "niz"
then sees



"niz" correctly predicts *blue* and *body*

hears "wug"
then sees



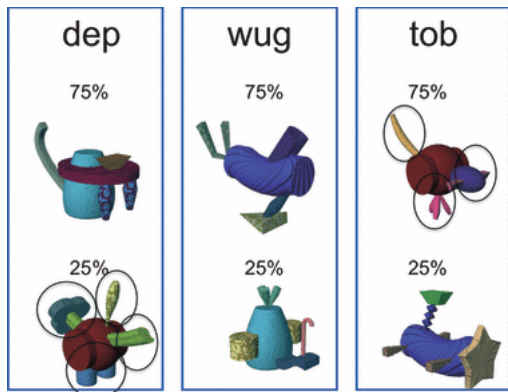
"wug" correctly predicts *red* and *body*

Cue	Outcome	S0	S1	S2	S3
wug	grey	0	0.1	0.1	0.19
niz	grey	0	0	0	0
wug	black	0	0	0	0
niz	black	0	0	0.1	0.1
wug	bird	0	0.1	0.1	0.19
niz	bird	0	0	0.1	0.1

Cue	Outcome	S0	S1	S2	S3
grey	wug	0	0.1	0.1	0.181
grey	niz	0	0	0	-0.01
black	wug	0	0	-0.01	-0.01
black	niz	0	0	0.1	0.1
bird	wug	0	0.1	0.09	0.171
bird	niz	0	0	0.1	0.09

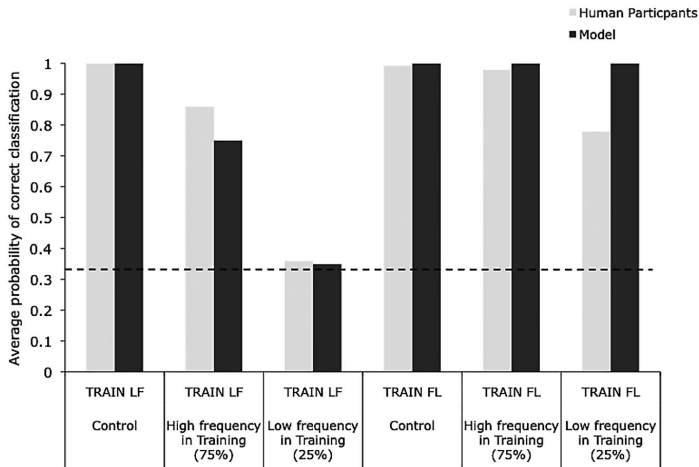
Learning a word IV

- ▶ Experimental confirmation: learning ambiguous words such that
 - ▶ There is a prominent nondiscriminative feature (main shape)
 - ▶ Other features are fully discriminative, but have an unbalanced distribution



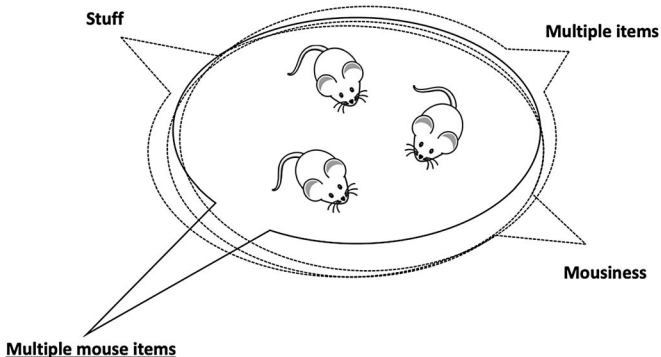
Learning a word V

Human learning matches the predictions of the model: low frequency associations are learned only in the FL presentation order.



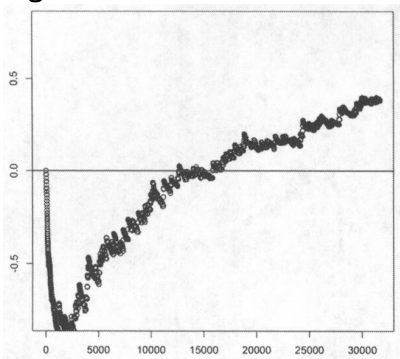
Learning an inflection system I

- ▶ The results above suggest that agents do not learn “a meaning” for words (or morphemes), they learn to carve regions of the feature space that correspond to a word.
- ▶ Application to inflection: when hearing an inflected word in context, the hearer must decide which features are coded.



Learning an inflection system II

- ▶ Simulation: using the RW equations, learn associations between collections of 4 features and 1 or 2 morphs (2 for regular plurals, 1 for singulars and irregular plurals)
- ▶ Likelihood of producing *mice* over time:



- ▶ U-shaped learning follows from the distribution of forms in the input data plus discriminative learning of words from contexts.

Why do languages have irregular forms?

- ▶ Conventional view:
 - ▶ Irregular morphology is a bug, not a feature: languages would be better off without it.
 - ▶ Irregular morphology is the consequence of unsystematic language change.
- ▶ Ramscar et al. (2018):
 - ▶ Regulars and irregulars contrast in discriminative value: suppletion entails maximal discriminability in the form dimension.
 - ▶ Suppletion is the extremum of a gradient of discriminability, not a discretely different case.
 - ▶ The value of regular morphology lies in its poor discriminative power: as the cues for a morphosyntactic distinction are unspecific, they can be redeployed productively to make sense of unseen words.

Why then is language uniquely human?

- ▶ The paper ends with speculations on why, if the learning mechanisms crucial to language are shared with all animals, only humans have language.
- ▶ Suggestive evidence from neuroscience: compared to other primates, humans are characterized by a slower elimination of synaptic connections in the brain that is also uneven in its pace across brain areas.
- ▶ Leads to the following conjecture: an inability to filter attention to the input is crucial to acquiring structured conventional knowledge.
 - ▶ Children need to not jump to conclusions to learn that the plural of *mouse* is *mice*.









Key insights

- ▶ Learning inflection is learning to pair meanings with words.
- ▶ Training data + a general, well-understood learning algorithm are enough to make sense of basic insights on the learning and processing of morphology, once one focuses on the right learning task.
 - ▶ Compare Malouf (2017): training data + general machine learning methods (LSTMs) are enough to produce an inflectional synthesizer with superhuman performance.
- ▶ A series of conjectures on the nature of meaning, the design properties of language, etc.

Roadmap

- ▶ This was the first of a series on discriminative learning and morphology:
 - ▶ 25/2: Naive discriminative learning, based on [Fabian Tomaschek et al. \(2021\)](#). “Phonetic effects of morphology and context: Modeling the duration of word-final S in English with naïve discriminative learning.” In: *Journal of Linguistics* 57.1, pp. 123–161
 - ▶ 04/03: (relevant) intermission: Distributional semantics and morphological relatedness
 - ▶ 11/03: Morphology reading group
 - ▶ 18/03: Linear discriminative learning, (probably) based on [R. Harald Baayen et al. \(2019\)](#). “The Discriminative Lexicon: A Unified Computational Model for the Lexicon and Lexical Processing in Comprehension and Production Grounded Not in (De)Composition but in Linear Discriminative Learning.” In: *Complexity* 2019, p. 4895891

References I

-  Baayen, R. Harald et al. (2019). “The Discriminative Lexicon: A Unified Computational Model for the Lexicon and Lexical Processing in Comprehension and Production Grounded Not in (De)Composition but in Linear Discriminative Learning.” In: *Complexity* 2019, p. 4895891 (cit. on p. 22).
-  Chuang, Yu-Ying and R. Harald Baayen (2021). “Discriminative learning and the lexicon: NDL and LDL.” In: *Oxford Research Encyclopedia of Linguistics*. Ed. by Mark Aronoff. Oxford University Press (cit. on p. 10).
-  Malouf, Robert (2017). “Abstractive morphological learning with a recurrent neural network.” In: *Morphology* 27.4, pp. 431–458 (cit. on p. 21).
-  Marcus, Gary F. et al. (1992). *Overregularization in language acquisition*. Wiley (cit. on p. 4).
-  Marcus, Gary F. et al. (1995). “German Inflection: The Exception That Proves the Rule.” In: *Cognitive Psychology* 29, pp. 189–256 (cit. on p. 4).
-  Pinker, Steven (1999). *Words and Rules*. New York: Basic Books (cit. on p. 4).
-  Pinker, Steven and Alan Prince (1988). “On language and connectionism: analysis of a parallel distributed processing model of language acquisition..” In: *Cognition* 28, pp. 73–193 (cit. on p. 4).
-  Prasada, Sandeep and Steven Pinker (1993). “Similarity-based and rule-based generalizations in inflectional morphology.” In: *Language and Cognitive Processes* 8, pp. 1–56 (cit. on p. 4).

References II



Ramscar, Michael (2021). “A discriminative account of the learning, representation and processing of inflection systems.” In: *Language, Cognition and Neuroscience* (cit. on pp. 1, 7).



Rumelhart, D. E. and J. L. McClelland (1986). “On Learning Past Tenses of English Verbs.” In: *Parallel distributed processing*. Ed. by J. L. McClelland, D.E. Rumelhart, and the PDP Research Group. Vol. 2. Cambridge: MIT Press, pp. 216–271 (cit. on pp. 3, 4, 7).



Tomaschek, Fabian et al. (2021). “Phonetic effects of morphology and context: Modeling the duration of word-final S in English with naïve discriminative learning.” In: *Journal of Linguistics* 57.1, pp. 123–161 (cit. on p. 22).