International Conference on Error-Driven Learning in Language (EDLL 2021)

Wednesday, March 10 — Friday, March 12, 2021

University of Tübingen

Book of Abstracts

Edited by Jessie S. Nixon, Elnaz Shafaei-Bajestan and Harald Baayen

Foreword

This volume includes the proceedings abstracts of the International Conference on Error-Driven Learning in Language, EDLL 2021, held online, Tübingen, Germany, 10-12 March, 2021. This is the first conference of its kind.

We believe error-driven learning models have made and will likely continue to make an important contribution to our understanding of language. Therefore, it is with great pleasure that we bring together researchers working in this exciting area for the first conference dedicated to error-driven learning in language. We are happy to have such a great line up of Keynote speakers: Randall O'Reilly, Adele Goldberg and Petar Milin. We have papers on various levels of linguistic processing, from speech sounds to morphology, syntax and semantics, on first and second language acquisition, perception, comprehension and production, neural correlates of error, as well as investigations into specific details of the learning algorithms. We have participants joining us from around the globe. We would like to thank all the participants and presenters for their contributions. We couldn't have done it without you. Thanks also to Adnane Ez-Zizi for help with the programme and to the ERC for financial support (grant number 742545). Finally, we look forward to further research and discussion into the future!

Jessie Nixon
Elnaz Shafaei-Bajestan
Harald Baayen

Tübingen, March 2020

March 10, 2021

9:00 - 9:15	**COFFEE/TEA**	
9:15 - 9:30	**WELCOME**	
KEYNOTE 1		
9:30 - 10:30	**Randall O'Reilly**	Predictive Error-driven Learning in the Brain.➥
10:30 - 10:40	**BREAK**	
10:40 - 11:10	Special event for junior researchers: Q&A with Randall O'Reilly	
11:10 - 11:15	**BREAK**	
SESSION 1		
11:15 - 11:45	**Yung Han Khoe, Chara Tsoukala, Gerrit Jan Kootstra and Stefan Frank**	Error-driven learning as a mechanism for cross-language structural priming ➥
11:45 - 12:15	**Chiara Gambi and Katherine Messenger**	The role of prediction errors in 4-year-olds' learning of English direct object datives ➥
12:15 - 12:45	**Jessica Nieder, Fabian Tomaschek and Ruben van de Vijver**	Modeling Maltese broken and sound plurals with Naive Discriminative Learning ➥
12:45 - 13:00	**BREAK**	
SESSION 2		
13:00 - 13:30	**Ben Ambridge and Kristen Liu**	Balancing information-structure and semantic constraints on construction choice: A discriminative learning model of passive and passive-like constructions in Mandarin Chinese (and Balinese and Hebrew). ➥
13:30 - 14:00	**Dušica Filipović Đurđević**	Uncertainty of polysemous word senses in the light of discrimination learning ➥
14:00 - 14:30	**Ksenija Mišić, Dušica Filipović Đurđević**	Interaction of semantic and syntactic ambiguity in the light of discrimination learning ➥
14:30 - 14:45	**BREAK**	
14:45 - 15:45	**POSTER SESSION**	

March 11, 2021

SESSION 3		
14:30 - 15:00	**Ronny Bujok, Sybrine Bultena, James McQueen and Mirjam Broersma**	Accent Adaptation through Error-Based Learning ➥
15:00 - 15:30	**Kristin Lemhöfer**	Error-driven learning in L2 vocabulary and syntax: ERP correlates ➥
15:30 - 15:45	**BREAK**	
SESSION 4		
15:45 - 16:15	**Sanne Poelstra, Jessie S. Nixon and Jacolien van Rij**	Does learning occur in the absence of cues? ➥
16:15 - 16:45	**Adnane Ez-Zizi, Dagmar Divjak and Petar Milin**	Error-correction mechanisms in language learning: tracking individual differences ➥
16:45 - 17:15	**Vsevolod Kapatsinski**	When backward transitional probabilities can be learned using forward prediction ➥
17:15 - 17:30	**BREAK**	
KEYNOTE 2		
17:30 - 18:30	**Adele E. Goldberg**	Explain me this: Coverage encourages generalization and Statistical Preemption constrains it➥
18:30 - 18:40	**BREAK**	
18:40 - 19:10	Special event for junior researchers: Q&A with Adele E. Goldberg	

March 12, 2021

SESSION 5		
9:30 - 10:00	**Theres Grüter, Yanxin (Alice) Zhu and Carrie N. Jackson**	Can forcing second language learners to generate prediction errors increase learning?
10:00 - 10:30	**Yanxin (Alice) Zhu, Yang Zhao and Theres Grüter**	Second language learners' sensitivity to competing alternatives is modulated by proficiency: Evidence from L2 Mandarin
10:30 - 11:00	**Chi Zhang and Min Wang**	Effects of input type frequency on structural priming and statistical preemption in the acquisition of L2 dative construction
11:00 - 11:15	**BREAK**	
SESSION 6		
11:15 - 11:45	**Harish Tayyar Madabushi, Dagmar Divjak and Petar Milin**	Less is more? Language learning, between simple and deep embeddings
11:45 - 12:15	**Laurence Romain, Petar Milin and Dagmar Divjak**	Learnability and Tense Aspect combinations in English: unveiling a dual system grounded in experience
12:15 - 12:45	**Benjamin Tucker, Dagmar Divjak and Petar Milin**	A learning perspective on the emergence of abstractions
12:45 - 13:00	**BREAK**	
KEYNOTE 3		
13:00 - 14:00	**Petar Milin**	What can be used from learning?
14:00 - 14:15	**CLOSING REMARKS**	
14:15 - 14:20	**BREAK**	
14:20 - 14:50	Special event for junior researchers: Q&A with Petar Milin	
14:50 - ...	**SOCIAL EVENT**	

Poster session

Booth #1	**Jessie S. Nixon and Fabian Tomaschek**	Infant speech acquisition through error-driven learning of the acoustic speech signal.➥
Booth #2	**Marion Coumel, Ema Ushioda and Kate Messenger**	Can error-based models account for language processing via syntactic priming? Investigating the effects of task and learner characteristics ➥
Booth #3	**Kun Sun**	Semantic Vectors Based on Discriminative Learning as Predictive of Lexical Psychological Properties ➥
Booth #4	**Maja Linke and Michael Ramscar**	How Distributional Context Solves the Variance Problem in Speech Sampling ➥
Booth #5	**Michaela M. Vann, Giulia Bencini and Virginia Valian**	Learning trajectories in L2 and bilingual language development: a structural priming investigation ➥
Booth #6	**Xuefeng Luo, Yu-Ying Chuang and Harald Baayen**	Linear Discriminative Learning in Julia ➥
Booth #7	**Motoki Saito, Fabian Tomaschek and R. Harald Baayen**	Triphone meanings co-determine tongue shape during articulation: An ultrasound study ➥

Predictive Error-driven Learning in the Brain

Randall O'Reilly
University of California, Davis; oreilly@ucdavis.edu

I will present some recent computational models of brain circuits that can support predictive error-driven learning, along with a discussion of prior work on how the brain might support something like error backpropagation more generally. Error backpropagation is the engine of modern deep neural network models, and there has been a bit of a resurgence of interest in its possible biological basis recently. Top-down connections in the cortex can potentially provide a mechanism of error propagation, and there are various proposals that make distinct biological predictions, which will be reviewed. Predictive learning provides an attractive solution to a remaining challenge: where do all the error signals come from in the first place? Specific circuits between the thalamus and cortex appear ideally configured to support a form of predictive learning, which differs significantly from other machine-learning / Bayesian approaches. Our models show that this mechanism can learn abstract categorical representations from movies of rotating and translating 3D objects, and captures classic statistical learning phenomena in speech recognition. It is also consistent with how many current deep neural network models are trained.

Explain me this:
Coverage encourages generalization and Statistical Preemption constrains it

Adele E. Goldberg
Princeton University; adele@princeton.edu

How is that native English speakers find novel patterns such as *She tweeted them the story* unremarkable but stubbornly judge *She explained him the story* unacceptable? This apparent paradox is addressed by recognizing speakers' goal: to express their intended messages while obeying the conventions of their language community. Experimental evidence indicates that productivity is encouraged by Coverage (roughly the extent to which the required generalization has been previously attested), while productivity is constrained by statistical preemption: the existence of a more conventional, accessible alternative.

What can be used from learning?

Petar Milin
p.milin@bham.ac.uk
University of Birmingham UK

In my talk I will present work done with the Out Of Our Minds team [outofourminds.bham.ac.uk]. Through selected case studies I aim to show the range and reach that learning has in our research. Firstly, as a starting point, we see the role of learning along the lines of Poggio's revision of Marr's levels of understanding *any* complex system, language included: "understanding at the level of learning is […] perfectly adequate as an explanation all by itself" (2012, p. 1019). But the question about what learning is, more specifically, still needs an answer.

Many would accept that *learning is* (relatively permanent) *change*. Such a broad proposal allows diverse types of change to be considered as learning. In the first two studies we compare Memory-Based and Error-Correction learning (MBL vs. ECL). MBL is championed by (computationally inclined) usage-based linguists; in it, change happens when new exemplar gets added to memory. ECL, however, formalizes change as filtering (in machine learning) or discrimination (in cognitive science), to minimize error in predicting an outcome. Our results show that ECL is a worthy opponent: it fits the experimental data better and has better biological (or cognitive) credibility.

In the next two studies, we take further steps. Naïve Discrimination Learning (NDL), our main computational modelling framework, makes use of certain *ad hoc* abstractions for input cues and outcomes in error-correction (discriminative) learning. We ask what would happen if those abstractions were linguistically informed: Langacker considers *units* as being "abstracted from usage events [...] through the reinforcement of recurring commonalities" (2019, p. 346). This is, indeed, suspiciously similar to a *relatively permanent change from experience*, once we resist the temptation to misinterpret usage-based *units* as static and idealized (an assumption usage-based linguists don't hold). With this in mind, we tested whether bottom-up cues, from the original NDL setup, can be coupled with and benefit from top-down, theoretically motivated cues, in learning the same lexical outcomes. The results, based on data from self-paced reading, show that both types of cues do discriminate lexical outcomes and contribute significantly to predicting reading latencies, although they have somewhat different roles. The results also show that learning is not limited to the actual linguistic cues focused on in the experiment; instead, learning applies to any and all cues present in the situation. Using insights from Reinforcement Learning we frame our findings in terms of exploration and exploitation.

I conclude with a simple point that the results our work, jointly taken, provide empirical traction for the theoretical point of Spreat & Spreat that "much like the law of gravity, the laws of learning are always in effect" (1982, p. 593).

The role of prediction errors in 4-year-olds' learning of English direct object datives.

Chiara Gambi, Cardiff University, gambic@cardiff.ac.uk
Kate Messenger, University of Warwick, K.Messenger@warwick.ac.uk

Is children's acquisition of structural knowledge driven by prediction errors? According to error-driven models of language acquisition (e.g., [1],[2]), children generate linguistic expectations about upcoming words, compare them to the linguistic input, and when they detect a mismatch (i.e., prediction error signal) they update their long-term linguistic knowledge. But we only have limited empirical evidence for this learning mechanism. Prediction error (induced by violations of verb-specific structural preferences) modulates the magnitude of both short-term [3] and cumulative priming effects [4] in production tasks using the English dative alternation. However, these studies tested 5 and 6 year olds, who are already able to understand and use both prepositional object (PO) datives and the more difficult direct object (DO) datives.

In contrast, we do not know whether younger children's acquisition of DOs is driven by prediction error. We test whether 4-year-olds' understanding of DOs improves more when children are exposed to input that encourages the generation of prediction error signals. Specifically, we contrast training conditions where the input allows children to generate strong expectations about upcoming words (which later turn out to be incorrect) to training conditions where the input does not support expectation-generation: Strong expectations, when disconfirmed, should lead to larger improvements in understanding.

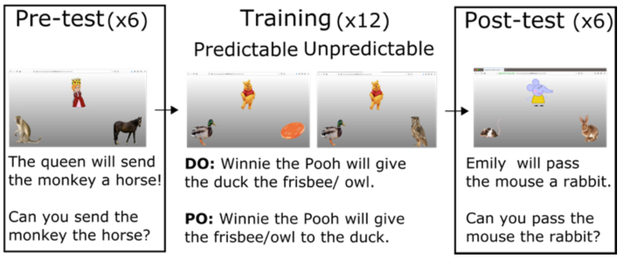
We developed a novel web-based touchscreen task, comprising of three phases. In the first phase, we assessed children's baseline comprehension of DOs (pre-test, see Fig 1). Children listened to pre-recorded DO sentences whilst viewing pictures of the theme and recipient on a touchscreen, and then acted out their interpretation: Correct answers required dragging the theme picture (e.g., horse) towards the recipient picture (e.g., monkey). Then, we exposed children to one of four different training conditions, and finally assessed their DO comprehension skills again (post-test), using a different set of sentences.

During training, children (N = 98) listened to 12 dative sentences - either all POs or DOs (between participants). PO training conditions control for structural priming effects (exposure to DOs may increase children's post-test performance regardless of prediction error). Critically, we contrasted (also between participants) training sentences with an inanimate theme (e.g., frisbee in Fig 1) *versus* an animate theme (e.g., owl); the recipient was always animate (e.g., duck). Since inanimate referents are more likely to be themes, children looking at a frisbee and a duck could predict the frisbee would be the theme even before they heard the sentence [5]; but children looking at an owl and a duck could not make this prediction. Thus, after hearing *Winnie the Pooh will give…*, children exposed to inanimate themes should generate a stronger expectation for the theme (Pred condition), compared to children exposed to animate themes (NonPred). Importantly, this predictability manipulation should only affect learning for children trained on DOs because only these children had their expectations disconfirmed in the Pred condition by hearing the recipient before the theme (e.g., *Winnie the Pooh will give...the duck...the frisbee.*).

We assessed post-test performance (while controlling for pre-test scores) separately for test items with inanimate themes (AI) and those with animate themes (AA); AI items are easier for children [6], a finding we replicated (54.59% vs. 39.97% accurate). Interestingly, there was no structural priming effect, nor an interaction between sentence type (PO vs. DO) and predictability for the easier AI test trials (all p's > .360; see left-hand panel of Fig. 2). Importantly, however, for the more difficult AA test trials, we found not only a priming effect (B=0.65, SE=0.32, z=2.06, p=.039), but also a larger average improvement in comprehension accuracy (from pre- to post-test) for children exposed to DOs in the Pred condition (24.24%, N=22), compared to those exposed to DOs in the NonPred condition (<1%; N=20); improvement following PO training was low overall (see right panel, Fig. 2).

These findings provide preliminary evidence that prediction error drives acquisition of difficult direct object sentences in 4 year olds. However, the interaction between sentence type and predictability for AA test trials was only marginally significant (B=1.24, SE=0.63, z = 1.95, p =.051). We plan to resume data collection when the COVID-situation allows.

Figure 1. Schematic summary of the three-phase study design. For each of the three phases (pre-test, training, post-test) we show one representative item, with visual input at the top and sample sentences at the bottom. Pictures are screenshots from the web-based task, which children completed on a touchscreen tablet.



The queen will send the monkey a horse!

Can you send the monkey the horse?

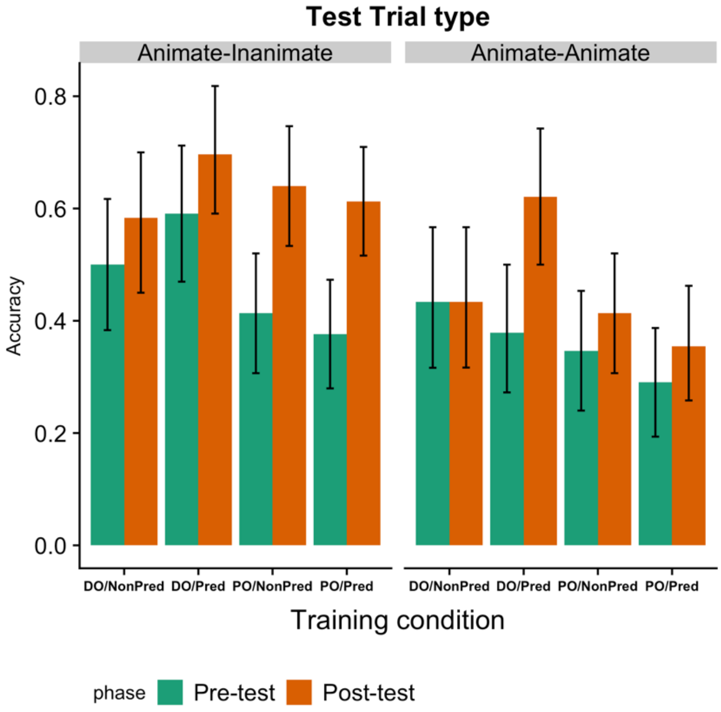
DO: Winnie the Pooh will give the duck the frisbee/ owl.

PO: Winnie the Pooh will give the frisbee/owl to the duck.

Emily will pass the mouse a rabbit.

Can you pass the mouse the rabbit?

Figure 2. Mean pre-test (green bars) and post-test (orange bars) comprehension accuracy in the four training conditions (PO = prepositional object, DO = direct object, NonPred = unpredictable condition, Pred = predictable condition); the left panel shows data for the easier Animate-Inanimate (AI) test trials, while the right panel shows data for the more difficult Animate-Animate (AA) trials. Error bars represent 95% bootstrap CIs.



References

[1] Elman, J. L. (1990). Finding structure in time. *Cognitive Science, 14(2),* 179-211.
[2] Chang, F., Dell, G. S., & Bock, K. (2006). Becoming syntactic. *Psychological Review, 113(2),* 234-272.
[3] Peter, M., Chang, F., Pine, J. M., Blything, R., & Rowland, C. F. (2015). When and how do children develop knowledge of verb argument structure? Evidence from verb bias effects in a structural priming task. *Journal of Memory and Language, 81,* 1-15.
[4] Fazekas, J., Jessop, A., Pine, J. M. & Rowland, C. F. (in press). Do children learn from their predictions mistakes? A registered report evaluating error-based theories of language acquisition. *Royal Society Open Science*.
[5] Thothathiri, M. & Snedeker, J. Syntactic priming during language comprehension in three-and four-year-old children. *Journal of Memory and Language, 58(2),* 188-213.
[6] Buckle, L., Lieven, E. & Theakston, A.L. (2017). The effects of animacy and syntax on priming: A developmental study*. Frontiers in Psychology, 8*, doi:10.3389/fpsyg.2017.02246

Balancing information-structure and semantic constraints on construction choice: A discriminative learning model of passive and passive-like constructions in Mandarin Chinese (and Balinese and Hebrew).

Ben Ambridge

University of Liverpool, UK

ESRC International Centre for Language and Communicative Development (LuCiD)

Li Liu

Guangdong University of Foreign Studies, China

The goal of this study was to build a discriminative-learning model of how Mandarin speakers choose between one of four truth-value-identical constructions when producing a two-argument utterance, given two (often competing) constraints: (a) An information structure constraint which specifies that “The denotation of the *by*-phrase NP in a passive clause must denote something at least as new in the discourse as the subject”. (Pullum 2014:64) and (b) A construction-semantic constraint such that the BEI Passive (1) and BA Active constructions (2), but not the Notional Passive (3) and SVO Active constructions (4), are associated with the meaning of affectedness of the PATIENT (i.e., of the OBJECT of the active forms).

	PATIENT Affected	PATIENT not (necessarily) affected
Topic = PATIENT	Mandarin O-BEI-SV passive (English OVS passive)	Mandarin OSV notional passive
Topic = AGENT	Mandarin S-BA-OV active	Mandarin SVO Active (English SVO Active)

(1) Lisi bei Zhangsan jiu le.
Lisi was saved by Zhangsan.

(2) Zhangsan ba Lisi jiu le.
Zhangsan saved Lisi.

(3) Zaofan Zhangsan chi le.
Zhangsan finished his breakfast.

(4) Zhangsan jiu le Lisi.
Zhangsan saved Lisi.

First, we conducted a grammaticality judgment study with 60 native speakers which confirmed that, across 57 verbs, semantic affectedness – as determined by a further 16 native speakers – determined each verb’s relative acceptability in the BEI Passive and BA Active constructions, but not the Notional Passive and SVO Active constructions.

Second, in order to simulate acquisition of these competing constraints, we built a discriminative learning model that learns to map from corpus-derived input (information structure + verb semantics + lexical verb identity) to an output representation corresponding to these four constructions. The model was able to predict judgments of the relative acceptability of the test verbs in the BA Active and BEI Passive constructions, obtained in Study 1, with model-human correlations in the region of *r*=0.48 and *r*=0.33, respectively.

Third, in ongoing work, we are extending the model to simulate equivalent already-collected data for passive(-like) constructions in Balinese and Hebrew.

Together, these findings contribute to a growing body of evidence showing that error-driven-learning models in general, and Resorla-Wagner/Widrow-Hoff style discriminative learning models in particular, hold considerable promise as mechanistic accounts of language acquisition, extending this evidence to a new domain (verb argument structure) and to new languages.

Figure 1. Grammaticality Judgment scores (y axis) as a function of human semantic affectedness ratings (x axis)

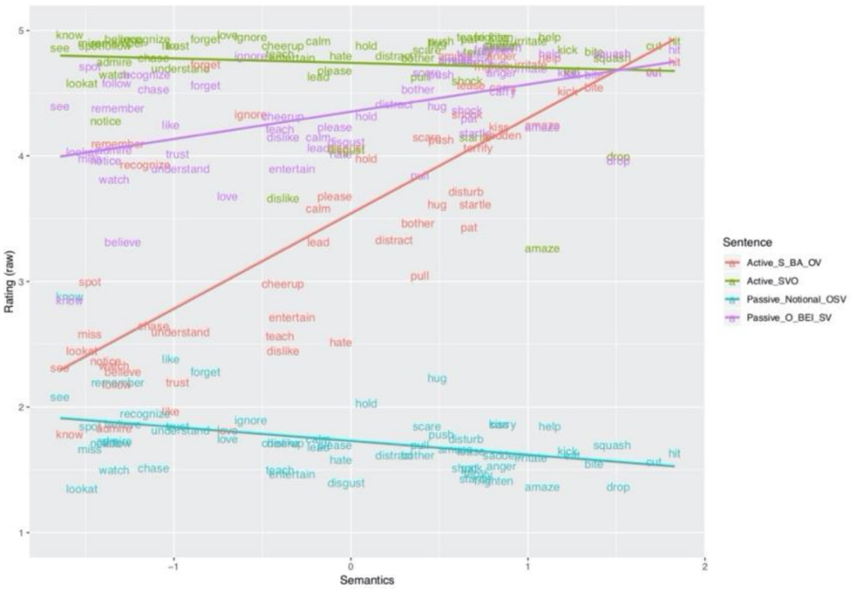


Figure 2. Architecture of the discriminative-learning model

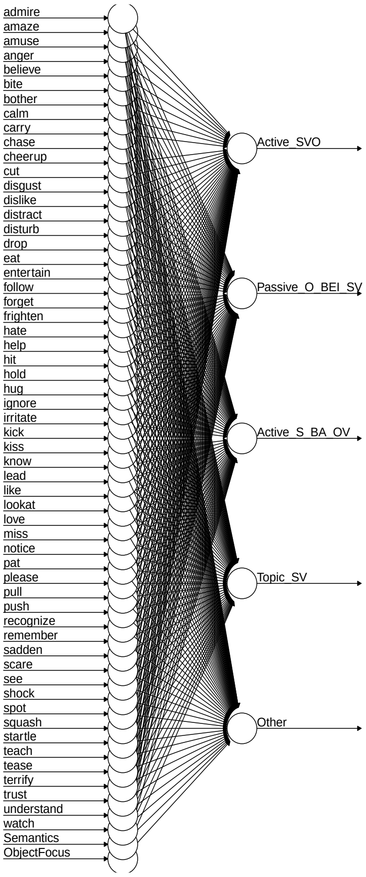
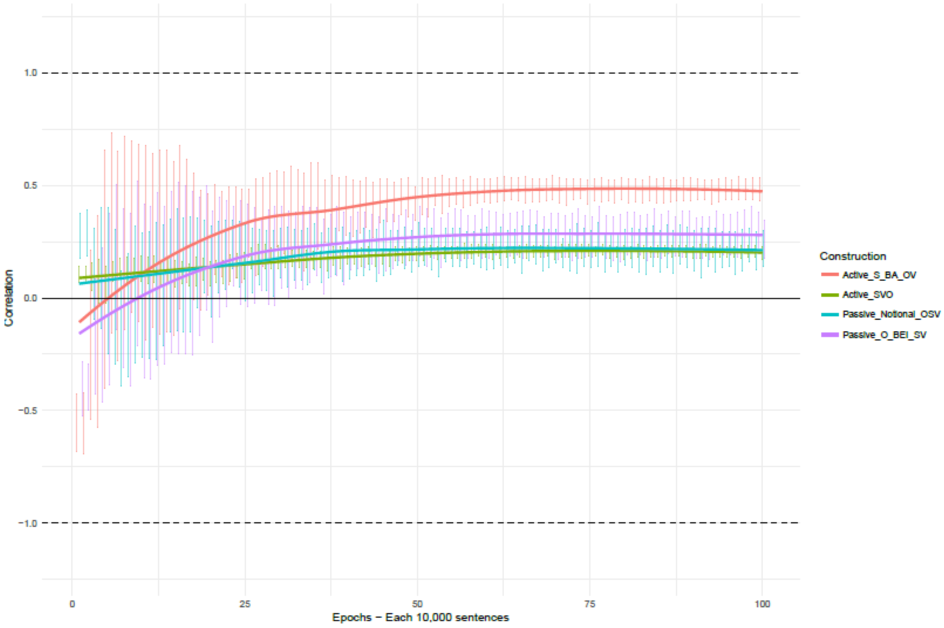


Figure 3. Model-human correlations



Can error-based models account for language processing via syntactic priming? Investigating the effects of task and learner characteristics

Marion Coumel[a], Ema Ushioda[b], Katherine Messenger[a]
[a]Psychology Department, University of Warwick, Coventry, UK
[b]Department of Applied Linguistics, University of Warwick, Coventry, UK

Recent psycholinguistic models propose that first (L1) and second language (L2) syntactic priming relies on an implicit, error-based language processing and learning mechanism[1,2]. Empirical support for this account mostly comes from studies obtaining L1 and L2 long-term priming[3,4] and inverse frequency effects[5,6] but some of the model's predictions remain largely unexplored. First, the model states that priming magnitude should vary with individuals' learning rate which should itself be determined by task characteristics. However, few studies have examined, for instance, whether task aspects such as the modality of prime sentences (i.e., auditory vs. visual) influences priming[7,8]. Second, few studies have tested the model's prediction that learner characteristics such as individual differences in attention and motivation should affect one's learning rate and thus, priming[9,10]. While most research testing these predictions targets L1 speakers, we expected L2 speakers' priming behaviour to be more sensitive to variation in task and learner characteristics given their overall reduced experience with the target language. Indeed, presenting prime sentences visually vs. auditorily might be particularly helpful for L2 speakers and attention and motivation are very relevant to second language learning[11,12]*.* Thus, the present study examined the effect of prime sentences' modality and individual differences in attention and motivation on L2 and L1 speakers' immediate and long-term syntactic priming.

Using an online written picture description task, we compared French L2 English speakers' and English L1 speakers' primed production of the active/passive syntactic alternation (Fig. 1). We manipulated between-subjects whether participants listened to (listening condition) or read the prime sentences (reading condition). We assessed attention (L2 and L1 speakers) and motivation (L2 speakers only) with questionnaires. We measured immediate priming (repeating a syntactic structure after a prime) and long-term priming (producing more target structures in immediate and delayed post-tests without primes relative to pre-tests).

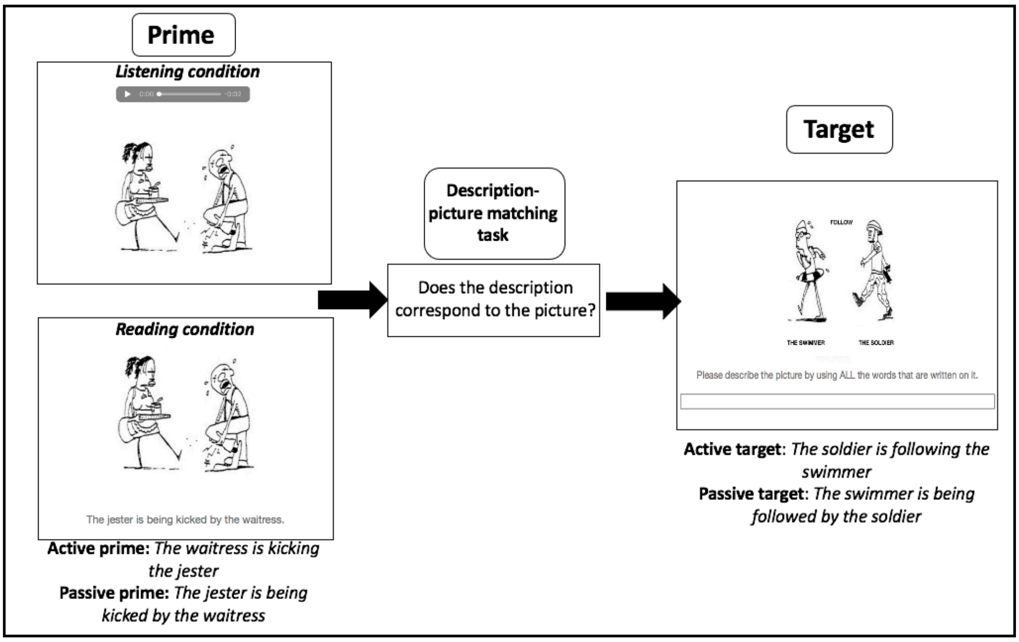
Overall, we predicted that both speaker groups would show immediate and long-term priming and that higher attention levels (both groups) and higher motivational levels (L2 speakers) would lead to larger priming effects. However, because reading the primes (vs. listening to them) was expected to facilitate L2 speakers' processing of the target structures, we expected L2 speakers to show more immediate and long-term priming in the reading than in the listening condition. For the same reason, we predicted that higher attention and motivation levels would be more helpful in the listening than in the reading condition and thus, boost L2 speakers' priming more in the former than in the latter condition. On the contrary, we expected that L1 speakers would exhibit the same priming strength and that attention would have the same effect irrespective of modality conditions.

As predicted, both speaker groups experienced immediate and long-term priming in the immediate post-test (Fig. 2 & 3). However, only L2 speakers exhibited long-term priming in the delayed post-test. Regarding the effect of modality, the results support our predictions regarding L1 but not L2 speakers. Prime modality did not affect priming in either group. Moreover, we did not find any interaction between attention or motivation, prime modality and any of the three priming types in L2 speakers. Only L1 speakers who were more attentive to the stimuli and the task were also more likely to experience long-term priming in the immediate post-test across modality conditions.

Overall, the long-term priming effects provide evidence that syntactic priming is a language learning mechanism. The between-group difference in the delayed post-test is in line with the model's prediction that less experienced speakers should experience more learning. Yet, the findings do not support the predictions regarding the effect of modality and provide limited support for the predictions regarding the effect of learner characteristics.

References:

[1]Chang, Dell, & Bock, (2006), *Psychological Review*, *113*, 234-272
[2]Jackson, (2018), *Second Language Research, 34*, 539-552
[3]Hartsuiker and Kolk, (1998), *Language and Speech, 41*, 143-184
[4]Jackson & Ruf, (2017), *Applied Psycholinguistics, 38,* 315-345
[5]Hartsuiker & Westenberg, (2000), *Cognition, 75*(2), B27-39
[6]Kaan & Chun, (2017), *Language and Cognition,* 1-15
[7]Cleland & Pickering, (2006), *Journal of Memory and Language, 54*(2), 185-198
[8]Hartsuiker et al., (2008), *Journal of Memory and Language, 58*(2), 214-238
[9]Bock et al., (1992), *Psychological Review, 99*(1), 150-171
[10]Ivanova et al., (2020), *Journal of Memory and Language, 110*
[11]Takahashi, (2005), *Applied Linguistics, 26*, 90-120
[12]Ushioda, (2016), *Language Teaching, 49*, 564-577



Prime

Listening condition

Reading condition

The jester is being kicked by the waitress.

Active prime: *The waitress is kicking the jester*
Passive prime: *The jester is being kicked by the waitress*

Description-picture matching task

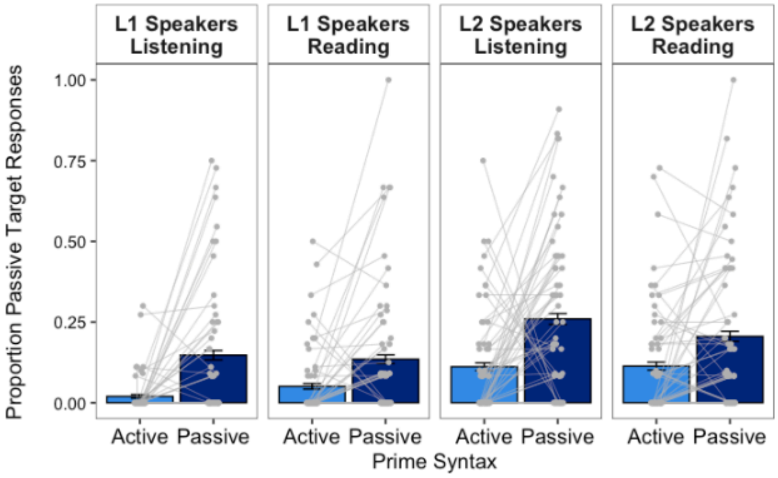
Does the description correspond to the picture?

Target

Please describe the picture by using ALL the words that are written on it.

Active target: *The soldier is following the swimmer*
Passive target: *The swimmer is being followed by the soldier*

Figure 1. Experimental trial and example of active/ passive alternation.



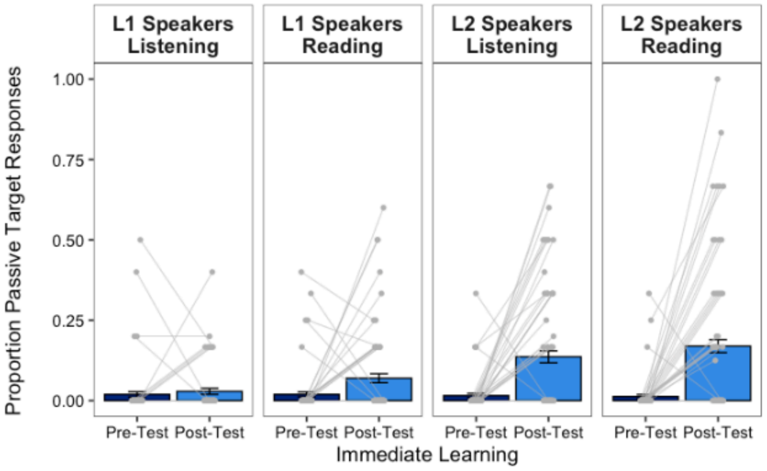


Figure 2. Passive responses in the immediate priming phase. Mean proportion of passive responses out of all transitive responses by prime syntax, prime modality and group condition in the immediate priming phase. Error bars indicate the standard error of the mean, grey dots indicate individual data points and grey lines individual priming effects.

Figure 3. Passive responses in the pre- and immediate post-test. Mean proportion of passive descriptions in the pre- and immediate post-test out of all transitive descriptions by section, prime modality and group condition. Error bars indicate the standard error of the mean, grey dots indicate individual data points and grey lines individual priming effects.

Does learning occur in the absence of cues?

Sanne Poelstra[a], Jessie S. Nixon[b], Jacolien van Rij[a]
[a]University of Groningen, [b]University of Tübingen
s.poelstra.1@student.rug.nl, jessie.nixon@uni-tuebingen.de, j.c.van.rij@rug.nl

Discriminative, Error Driven Learning (EDL) is a theory and set of equations that model bottom-up learning by minimising the uncertainty in the learner's expectations about upcoming events. Well-known formalisations of EDL include the Rescorla-Wagner model (1972) and the almost identical Delta Rule (Widrow & Hoff, 1960). Generally, we model learning using a fully connected, two-layer network (i.e. input layer: cues; output layer: outcomes; no hidden layers). The informativeness of cues is a key notion in EDL: only if cues are present are the connection weights between cues and outcomes updated. With each learning event the connections between present cues and outcomes are strengthened, while the connections between present cues and absent outcomes are weakened.

However in their frequently cited paper, Van Hamme and Wasserman (1994) have argued based on experimental data, that we can also learn from absent cues. They proposed an adjustment to the Rescorla-Wagner model: An absent cue should be encoded negatively, which leads to a weakened connection between an absent cue and present outcome and a strengthened connection between an absent cue and an absent outcome.

In the present study we aim to disentangle these two models of EDL. We implemented two computational simulations that model the experimental study reported by Van Hamme and Wasserman (1994). One simulation implements the Rescorla-Wagner model; the second implements the adaptation proposed by Van Hamme and Wasserman, which allows for learning from absent cues. In this experiment, participants had to indicate how likely it was that certain foods caused an allergic reaction. There were three types of food, of which two occurred on each trial together with an outcome (an allergic reaction or not). The participants then estimated the causal relation on a scale from 0 to 8 for *all three foods*.

Figure 1 shows the results of our computational simulations. To model the rating scale, we calculated weights to *Allergic reaction* minus weights to *No reaction*. The simulations show that with the Van Hamme & Wasserman experiment design - specifically, when the response measure (rating) includes both outcomes (allergy, no reaction) - there are no substantial differences in weight development between the Rescorla-Wagner and the Van Hamme-Wasserman models. Although the strength of activations is numerically different, we do not have a link function sufficient to evaluate which model best describes the data. Therefore, the two models make essentially the same predictions. These simulations demonstrate that Van Hamme & Wasserman's experiment design was not able to tease apart which model performs better: so, whether or not we learn from absent cues remains an open question.

However, our simulations also showed that the two models do make different predictions during the later phases of the experiment – *if* the individual outcomes are tested separately (see Figure 2). When weights to *Allergic* are separated from weights to *No reaction*, the Rescorla-Wagner model (left) predicts that, for example, 'bran' continues to predict the allergic reaction; in contrast, by the end of Block 3, the Van Hamme-Wasserman model (right) predicts that 'bran' is a negative predictor of the allergic reaction.

Based on our simulations, in ongoing work, we are running a series of experiments, all modifications of Van Hamme and Wasserman's experiment, to test the predictions of the two model variants. We will test outcomes separately at the end of Block 3. In addition, it is not clear whether Van Hamme and Wasserman's experiment reflects implicit learning, because they explicitly measured participants' ratings of present and absent cues. However, we argue that EDL is an implicit process, which may be hindered by explicit inference. Therefore, we will also employ a forced-choice paradigm and a speeded response manipulation to test the effects of explicit reasoning vs. implicit error-driven learning.

References

Rescorla, R., & Wagner, A. (1972). A theory of pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. *Classical conditioning ii*, *64*, 99.

Van Hamme, L. J., & Wasserman, E. A. (1994). Cue competition in causality judgments: The role of nonpresentation of compound stimulus elements. *Learning and motivation*, *25*(2), 127–151.

Widrow, B., & Hoff, M. E. (1960). *Adaptive switching circuits* (Tech. Rep.). Stanford Univ Ca Stanford Electronics Labs.

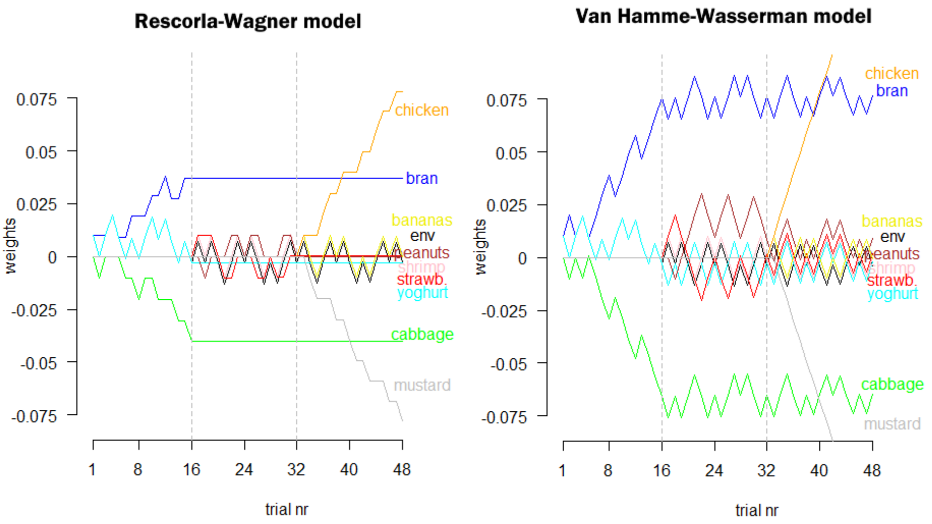


Figure 1: **Weights to *Allergic* minus the weights to *Not Allergic*** for each of the foods asked. Left: Rescorla-Wagner model. Right: Van Hamme-Wasserman model

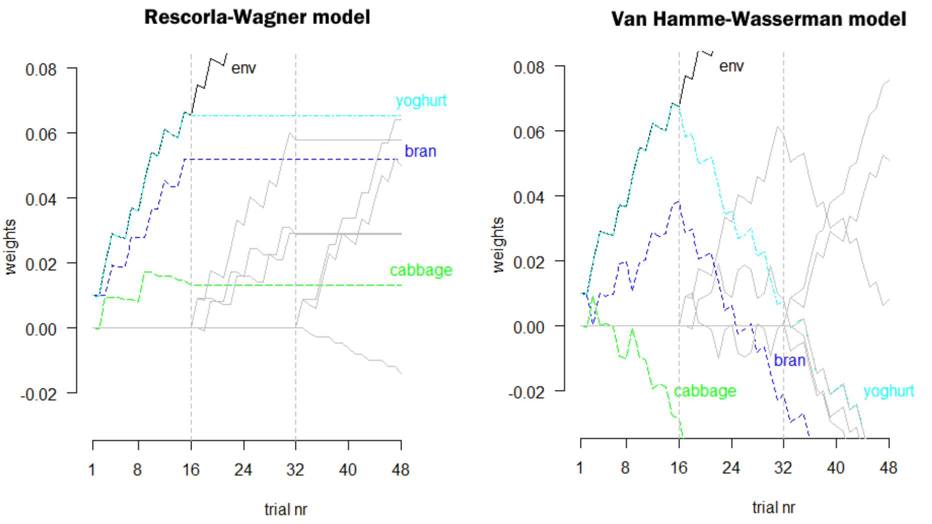


Figure 2: **Weights to *Allergic***. Left: Rescorla-Wagner model. Right: Van Hamme-Wasserman model

When backward transitional probabilities can be learned using forward prediction

Vsevolod Kapatsinski
University of Oregon

Backward transitional probability (BTP) is the probability of a word given the word that follows it, i.e., $p(\text{word}_i|\text{word}_{i+1})$ where *i* is position in an utterance. Perruchet and DeSaulty (2008) and Pellucchi et al. (2009) showed language learners to be capable of learning backward transitional probabilities from an input in which forward transitional probabilities, $p(\text{word}_i|\text{word}_{i-1})$, were controlled. These results have been argued to be problematic for predictive models of language learning in which learning results from predicting the future, and to provide decisive support for models that are capable of forming chunks based on either type of information (French et al., 2011; Perruchet, 2019). The present paper argues that this conclusion is premature: predictive models can become sensitive to either forward or backward transitional probabilities depending on a specific parameter setting.

While the discussion above has focused on recurrent networks, I trained a simpler two-layer network with an architecture previously proposed by Arnon and Ramscar (2012). The network was trained to predict the next word in the Switchboard Corpus (Godfrey et al., 1992) using its semantics and the identity of the preceding word, and used the Rescorla-Wagner learning rule (RW; Rescorla & Wagner, 1972). Semantic representations were either simple local codes or discretized Latent Semantic Analysis representations (Landauer & Dumais, 1997). In either case, they were more predictive of most of the words than the preceding word was. This greater predictiveness if crucial for obtaining the results below.

In its simplest form, RW updates cue→outcome associations based on the following two equations. For present outcomes and present cues, $\Delta w_{c\to o} = \alpha_c\beta_1(1-w)$; for absent outcomes and present cues, $\Delta w_{c\to o} = \alpha_c\beta_0(0-w)$. Crucially, the equations use different β parameters for present and absent outcomes. Figure 1 shows that if β_0 is much smaller than β_1(here it was set to zero and β_1 to .01 to show the extreme case),the item-to-item associations learned by the model reflect backward transitional probabilities and not forward ones. In contrast, if β_0 is set to the same value as β_1, the associations reflect forward transitional probabilities.

Since α and β are thought to reflect salience, a possible interpretation of this parameter manipulation is that the model's behavior depends on how much attention is allocated to absent forms. That is, sensitivity to backward transitional probability comes from paying little attention to absences.

This hypothesis generates a fresh perspective on differences between individuals and languages. It is often found that RW fits the data best if β_0 is smaller than β_1. Mckenzie and Mikkelsen (2007) have argued that this is because absences are less informative than presences, which suggests that the allocation of attention to absences may be adaptively adjusted by learners. A plausible mechanism for this adjustment is selection of attention allocation policies based on the prediction error that results from following a policy (Harmon et al., 2019). This explanation of sensitivity to BTP fits well with the finding that learners pick up on either BTP or FTP in an ambiguous artificial language based on which statistic is more informative in the participant's native language (Onnis & Thiessen, 2013).

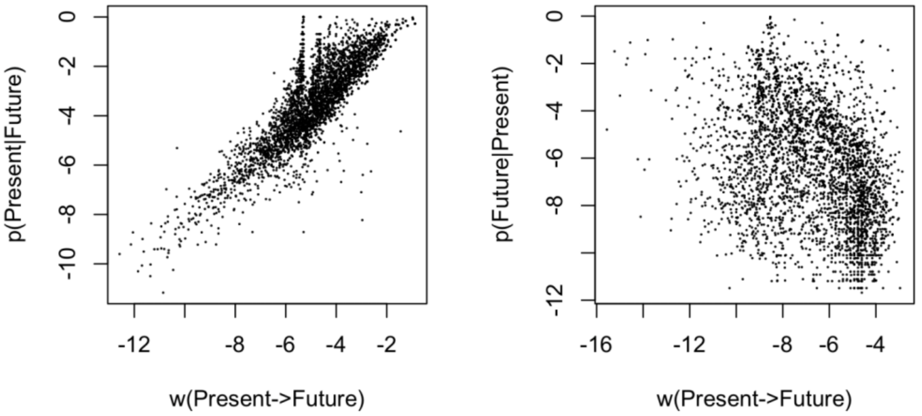


Figure 1. Forward associations (w) in a predictive model can track backward transitional probability (left panel, *r* = .76) rather than forward transitional probability (right panel, *r* = -.37) if there is cue competition between top-down and preceding-word cues to upcoming words, top-down cues are more predictive, and presences are much more salient than absences. Axes are log scaled. Semantic representations are localist. Points are individual tokens of words.

Arnon, I., & Ramscar, M. (2012). Granularity and the acquisition of grammatical gender: How order-of-acquisition affects what gets learned. *Cognition*, *122*(3), 292-305.

French, R. M., Addyman, C., & Mareschal, D. (2011). TRACX: A recognition-based connectionist framework for sequence segmentation and chunk extraction. *Psychological Review*, *118*(4), 614-636.

Godfrey, J. J., Holliman, E. C., & McDaniel, J. (1992). SWITCHBOARD: Telephone speech corpus for research and development. *IEEE International Conference on Acoustics, Speech, and Signal Processing* (Vol. 1, pp. 517-520).

Harmon, Z., Idemaru, K., & Kapatsinski, V. (2019). Learning mechanisms in cue reweighting. *Cognition*, *189*, 76-88.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*(2), 211-240.

McKenzie, C. R., & Mikkelsen, L. A. (2007). A Bayesian view of covariation assessment. *Cognitive Psychology*, *54*(1), 33-61.

Onnis, L., & Thiessen, E. (2013). Language experience changes subsequent learning. *Cognition*, *126*(2), 268-284.

Pelucchi, B., Hay, J. F., & Saffran, J. R. (2009). Learning in reverse: Eight-month-old infants track backward transitional probabilities. *Cognition*, *113*(2), 244-247.

Perruchet, P. (2019). What mechanisms underlie implicit statistical learning? Transitional probabilities versus chunks in language learning.*Topics in Cognitive Science*, *11*(3), 520-35.

Perruchet, P., & Desaulty, S. (2008). A role for backward transitional probabilities in word segmentation? *Memory & Cognition*, *36*(7), 1299-1305.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning II: Current research and theory* (pp.64-99). New York: Appleton-Century-Crofts.

Infant speech acquisition through error-driven learning of the acoustic speech signal

Jessie S. Nixon, Fabian Tomaschek
Quantitative Linguistics, University of Tübingen
jessie.nixon@uni-tuebingen.de, fabian.tomaschek@uni-tuebingen.de

Infants start out life with the ability to detect subtle changes in sensory information. Discrimination gradually changes depending on the predictive structure of the environment in a number of domains (e.g. Baker, Golinkoff and Petitto, 2006; Hannon and Trehub, 2005; Singh, Loh and Xiao, 2017), including the subject of the present study, the speech signal (e.g. Werker & Tees, 1984). What drives learning in speech perception is the subject of much ongoing debate.

In the present study, we investigate whether early infant acquisition of speech cues could occur through discriminative, error-driven learning (e.g. Ramscar & Yarlett, 2007; Ramscar et al., 2013) of the acoustic speech signal. We use a simple two-layer Rescorla-Wagner network (Rescorla & Wagner, 1972) trained on speech recordings from the CHILDES database. Because we were interested in learning in young infants of a few months of age, no lexical items were included in the model. Neither did we assume that infants have pre-existing representations of sound units, such as phonemes or phonetic features. Instead, both input cues and outcomes in the model were 25 ms by 0.47 mel components of spectral intensity extracted from the speech recording. The model was trained on a moving window with spectral components of three 25 ms temporal windows as cues predicting spectral components of one temporal window as outcomes. The model thus uses incoming acoustic cues to predict upcoming acoustic cues.

The model was evaluated against infant behaviour in the high-amplitude sucking (HAS) paradigm. Studies using this paradigm have shown that young infants can discriminate [i] vs [I] (Swoboda et al., 1976) and [s] vs [sh] (Eilers & Minifie, 1975). Swoboda et al. (1976) addtionally found that infant perception of [i] vs [I] is linear: within-category and between-category differences were discriminated equally well. We ran two tests to evaluate the model against these data. Firstly, summed activations from the spectral component cues of the target (e.g. [i]) to the target were compared to activation from the spectral components of the competitor ([I]) to the target. The model predicted significantly higher activation from target than competitor spectral components for both the vowels and the fricatives. This demonstrates that expectations developed through predicting upcoming speech signal from incoming speech signal enabled the model to discriminate the sound pairs, simulating infant behaviour in a common infant speech perception task.

Secondly, continua were created for the vowels and the fricatives. Summed activations from the cues were calculated for each step on the continuum to each endpoint of the continua for each spectral frequency band. With decreasing distrance from the competitor, the model predicted lower activation in the expected spectral frequency ranges for the vowels (F2 and F3) and lower activation of the competitor over a broad spectral frequency range for the fricatives (see Figure 1, first and second columns). Moreover, the activation of the vowels showed a linear decrease with decreasing acoustic distance from the competitor (Figure 1, top row, fifth column) as found by Swoboda et al. (1976). Activation for the fricatives was nonlinear over the continuum (Figure 1, bottom row, fifth column). This nonlinear perception has not yet been tested in infants, but is typical in adults (e.g. Mann & Repp, 1980).

In summary, using unstructured acoustic input cues to predict upcoming signal in running speech, the error-driven discriminative learning model learned to weight cues in such a way as to discriminate pairs of vowels and consonants – a standard measure of young infants' speech perception ability. The results suggest that error-driven learning of the acoustic signal may be a feasible model for infant acquisition of speech cues.

References

Baker, S.A., Golinkoff, R.M., Petitto, L.A. (2006). New insights into old puzzles from infants' categorical discrimination of soundless phonetic units. Language Learning and Development. 2, 147–162.

Eilers, R. E., Minifie, F. D., (1975). Fricative Discrimination in Early Infancy. Journal of Speech and Hearing Research 18, 158–167.

Hannon, E.E., Trehub, S.E. (2005). Metrical categories in infancy and adulthood. Psychological science 16, 48–55.

Mann, V. A., & Repp, B. H. (1980). Influence of vocalic context on perception of the [ʃ]-[s] distinction. Perception & Psychophysics, 28(3), 213-228.

Ramscar, M., Dye, M., & McCauley, S. M. (2013). Error and expectation in language learning: The curious absence of mouses in adult speech. Language, 89, 760–793.

Ramscar, M., Yarlett, D., (2007). Linguistic self-correction in the absence of feedback: A new approach to the logical problem of language acquisition. Cognitive Science 31, 927–960.

Rescorla, R. and Wagner, A., (1972). A theory of Pavlovian conditioning. Black, A. H., Prokasy, W. F. (Eds.), Classical conditioning II: Current research theory. Ap.-Cent-Crofts, New-York,. 64– 99.

Singh, L., Loh, D., Xiao, N.G., (2017). Bilingual infants demonstrate perceptual flexibilityin phoneme discrimination but perceptual constraint in face discrimination. Frontiers inPsychology 8, 1563.

Swoboda, P.J., Morse, P.A., Leavitt, L.A. (1976). Continuous vowel discrimination in normaland at risk infants. Child Development , 459–465

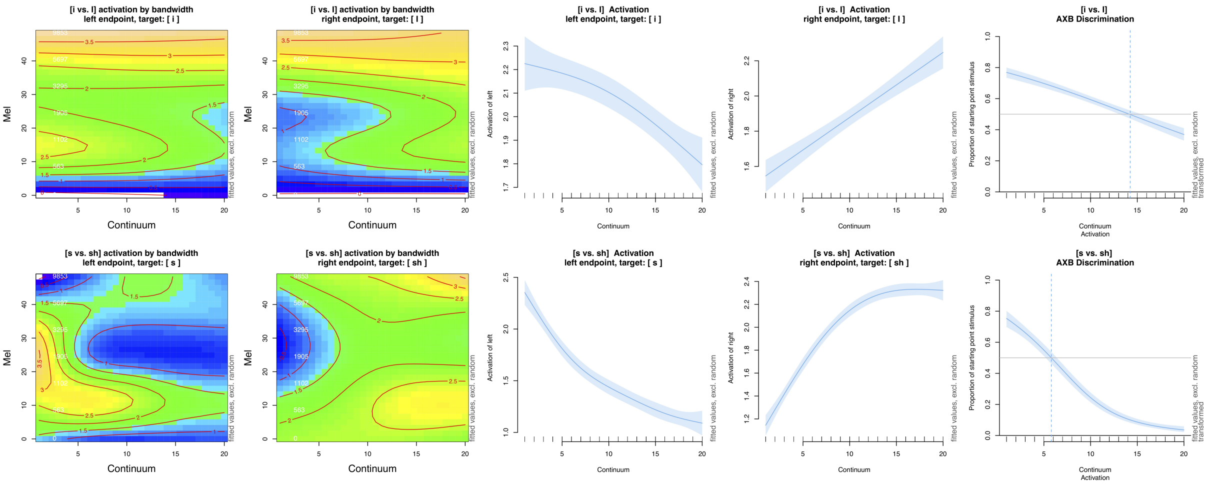


Figure 1: *Results for the vowels [*i - I*] (top row) and [*s - S*] (bottom row). The topographic plots show the estimated effect of the interaction between stimulus number and frequency on the activation of the left and right endpoint stimuli (first two columns). The x-axis represents the continuum step. The y-axis represents the spectral frequency (mel: outer axis label in black; Hz: inner axis label in white). Activation is represented by means of contour lines and color coding, where blue represents low activation; green, mid activation and yellow, high activation. Note that the z-limits differ between sound pairs. Third and fourth column: Average activation (y-axis) across the continuum (x-axis). Rightmost column: The smooth illustrates the probability (y-axis) of selecting the left endpoint stimulus in the AXB classification test along the continuum (x-axis). Y-axis values were back- transformed to probabilities.*

Accent adaptation through error-based learning

Ronny Bujok[1], Sybrine Bultena[2,3], James McQueen[3], Mirjam Broersma[2]

[1]Max Planck Institute for Psycholinguistics, Nijmegen
[2]Radboud University Nijmegen, Centre for Language Studies
[3]Radboud University Nijmegen, Donders Institute

Correspondence: Ronny.Bujok@mpi.nl

The ability of listeners to adapt to native accented speech (e.g., Maye et al., 2008), as well as foreign-accented speech (Bradlow & Bent, 2008), points to a high degree of flexibility in our speech perception. While the ability to adapt may be evident, the question of how listeners are able to adapt to accents so rapidly is still largely unanswered. It has been suggested that top-down knowledge (e.g., lexical knowledge) can guide accent adaptation (e.g. Norris et al., 2003). However, in the absence of sufficient context or linguistic information, for example in short and isolated utterances, other mechanisms must be at play. Language users monitor their errors internally to correct them and decrease their occurrence in the future. Because accented sounds can deviate starkly from the norm, how they are perceived is challenging and prone to errors. We thus suggest that one form of accent adaptation can be understood as being the development of specific internal error monitoring. We examined if accent adaptation can be explained in terms of feedback-driven error-based learning.

We created a novel accent which shifted various vowels downward, and applied it to a list of monosyllabic, highly frequent Dutch words (e.g.,'blik' /blɪk/ sounded like 'bluk' /blʏk/). Dutch native participants listened to the resulting accented words as a part of a 2AFC task, which asked them to decide which word on screen matched the accented auditory stimulus. Visually presented items always included a target ('blik') and distractor ('bleek') that formed a minimal pair. The task comprised two types of trials: accented words were either non-words (training), or sounded like actual Dutch words (test). Furthermore, in a proportion of test trials, the distractor word on screen was identical to the form of the auditory stimulus, resulting in error-prone items that allowed us to test how well participants had adapted to the accent. The task included 3 rounds, each consisting of 2 blocks (training block and mixed block, presenting only training items, and all items respectively), and participants received explicit feedback on their performance, such that they could learn from their mistakes. Using EEG, we measured participants' error detection as reflected by the error-related negativity (ERN). The ERN reflects internal error monitoring (Gehring et al., 2012).

Participants responded faster and their performance improved quickly in the course of the experiment (see Figure 1). Test items generally triggered more errors than training items. Test items with a distractor identical to the auditory stimulus led to more errors only in the first block. Moreover, the electrophysiological results (see Figure 1) show that initially the difference between response-locked negativities for correct and incorrect responses (i.e., an ERN effect) was small but significant, and this increased in later rounds. The effect did not differ between training and test items.

This study provides further evidence for the speed and flexibility of accent adaptation. The presence of the ERN effect in the first round demonstrates that internal monitoring develops very rapidly within just a few trials. It also appears robust as it extends to words that mismatch with stored lexical representations (i.e., test items). Moreover, it suggests that its development can be driven by explicit feedback. Taken together our findings support the idea that error-based learning is a mechanism of accent adaptation.

References:

Bradlow, A. R., & Bent, T. (2008). Perceptual adaptation to non-native speech. *Cognition*, *106*(2), 707–729. https://doi.org/10.1016/j.cognition.2007.04.005

Gehring, W. J., Liu, Y., Orr, J. M., & Carp, J. (2012). The error-related negativity (ERN/Ne). In *The Oxford handbook of event-related potential components* (pp. 231–291). Oxford University Press.

Maye, J., Aslin, R. N., & Tanenhaus, M. K. (2008). The Weckud Wetch of the Wast: Lexical Adaptation to a Novel Accent. *Cognitive Science*, *32*(3), 543–562. https://doi.org/10.1080/03640210802035357

Norris, D., McQueen, J. M., & Cutler, A. (2003). Perceptual learning in speech. *Cognitive Psychology*, *47*(2), 204–238. https://doi.org/10.1016/S0010-0285(03)00006-9

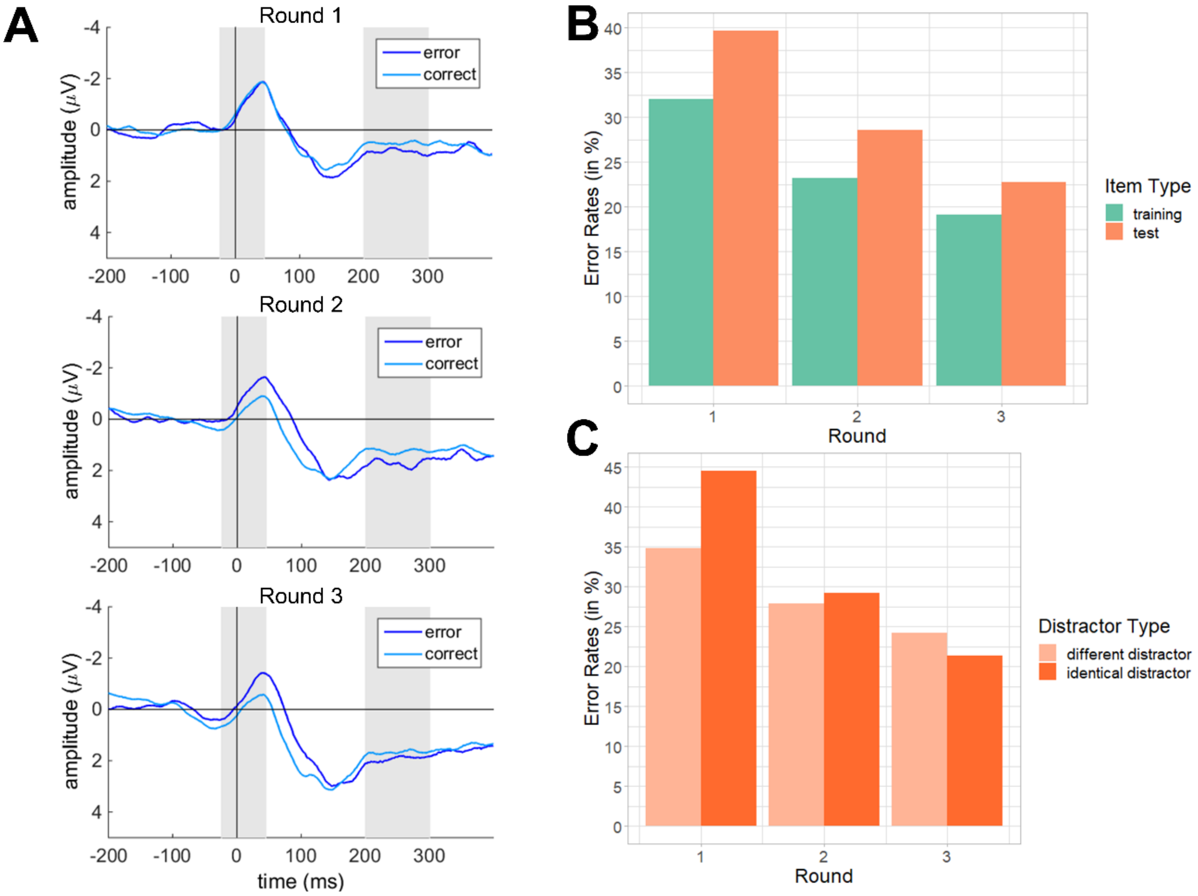


Figure 1. A) Response-locked data for error (ERN) and correct (CRN) responses across rounds at electrode location FCz. Grand averages for a subset of participants (n= 33). Shaded areas indicate the average latency of trough and peak (ERN) and the time window across which the PE was averaged. B) Comparison of error rates across rounds and item types (training vs. test). C) Error rates for test items are further split up to compare the two test item types (test items with distractors identical to the auditory stimulus vs. test items with a different distractor).

Semantic vectors based on discriminative learning as predictive of lexical psychological properties

Kun Sun
Department of Linguistics, University of Tübingen
email: kun.sun@uni-tuebignen.de

Psycholinguistic word information (i.e. lexical psycholinguistic properties, *LPP*, such as word concreteness, emotion, imageability, and familiarity) concerns the psycholinguistic properties of words that influence the processing and learning of words. LPP data has been collected through massive online rating. On the other hand, the meanings of words, well represented by semantic vectors, are essential to the properties of words. However, the LPPs has seldom been investigated to relate with the meanings of words. Several models concerning word meanings have been employed to predict word properties concerning cognitive and neural characteristics. The two most typically used are those that respectively use word co-occurrence and distributional semantics. However, these popular models do not give a plausible interpretation of the underlying cognitive mechanism. Additionally, the methods of using correlations or simple linear regression in past studies could not analyze the delicate interactions among the variables and detect whether fixed effects of predictors take place.

In order to overcome these limitations, the present study uses semantic vectors trained by the discriminative learning model to relate with LPP data, and further investigate whether semantic vectors can predict LPP data. To this end, generalized mixed-effects statistical models are used to compare what kind of semantic vectors (trained by discriminative model or word2vec) has stronger fixed effect on LPP data. Meanwhile, Bayesian multilevel statistical models are used to verify this. All results demonstrate that the semantic vectors based on the discriminative learning model is a good predictor of the LPP data. Additionally, the other LPPs or the interaction between semantic vectors and one LPP can also predict the other LPP that is taken as a response variable.

This study thus directly concerns the provision of effective methods for seeking a plausible cognitive mechanism for the connection between the LPPs and the word semantic information. It will further help in better understanding the nature of lexical semantics and word properties.

Key words: lexical psycholinguistic property, semantic information, L1norms, fixed effect, interaction

Learnability and Tense Aspect combinations in English: unveiling a dual system grounded in experience.

Laurence Romain, Petar Milin, Dagmar Divjak

In a usage-based approach to language, we assume that for linguistic categories to be plausible, they should be learnable from exposure to language. In this paper we enquire whether this is true of English Tense-Aspect (TA) categories. To do so, we trained an error-correction learning (ECL) model – Naïve Discriminative Learning (NDL; Baayen, Milin, Đurđević, Hendrix, & Marelli, 2011) that implements the Rescorla-Wagner (1972) rule. More specifically, we focused on the *learnability* of the various TA combinations in English to draw inferences about the type of cues and their informativity (i.e., ability to discriminate) for each TA category.

We trained our model on a sample of about 7 million sentences extracted from the British National Corpus. The dataset contained instances of all 12 possible TA combinations in English, but with varying frequencies. As the distribution of TA frequencies was, expectedly, Zipfian, we removed instances of low-frequency tenses, such as the future perfect progressive (<0.01%; e.g., *I will have been working for 8 hours by then.*). This left us with 11 possible outcomes. The remaining 11 TA categories had very different frequencies, with the present simple making up 46% of the data and the past simple 38%. For cues, we used the infinitive form of the verb whose TA combination we tried to predict and word n-grams within sentence boundaries (with n = [1,4]). We specifically focused on two questions: (1) how the skewed frequency distribution of TA outcomes co-affect NDL's prediction accuracy (viz. Boyd & Goldberg, 2009; Ellis, 2002) and furthermore (2) how *locality* vs. *contextuality* of cues (i.e., n = 1 vs. n > 1, or single vs. multiple word n-grams) affects informativity (discrimination) of TA outcomes.

Overall, our model achieved 68% prediction accuracy, which is well above the reference accuracy thresholds, either if predictions for each of the 11 categories are made randomly or if the most frequent TA (present simple) is always predicted. Nevertheless, we find that simpler forms, which are also the most frequent ones, are much easier for NDL to predict. We also find that the cues that are the most informative for these TA combinations are verb infinitives, whereas for the other TA combinations, higher-order n-grams make up the bulk of the most predictive cues. Thus, to answer our research question, we find that due to their high frequency, simpler forms rely mostly on local cues (n = 1; lexical). These two elements – high frequency and cue locality, conspire to make them easier to learn. The low frequency of complex TA combinations and their consequent reliance on more contextual (n-grams) cues makes them more difficult to learn. We believe this difficulty in learnability is closely related to the cognitive complexity (or lack thereof) of speakers' conceptualisations of temporal events.

Baayen, R. H., Milin, P., Đurđević, D. F., Hendrix, P., & Marelli, M. (2011). An amorphous model for morphological processing in visual comprehension based on naive discriminative learning. *Psychological Review, 118*(3), 438-481.

Boyd, J. K., & Goldberg, A. E. (2009). Input effects within a constructionist framework. *The Modern Language Journal, 93*(3), 418-429.

Ellis, N. C. (2002). Frequency effects in language processsing: A Review with Implications for Theories of Implicit and Explicit Language Acquisition. *Studies in Second Language Acquisition, 24*(2), 143-188. doi:10.1017/S0272263102002024

Rescola, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning ll: Curent research and theory* (pp. 64-99). New York: Appleton Century Crofts.

Error-driven learning in L2 vocabulary and syntax: ERP correlates

Kristin Lemhöfer

Models of error-driven learning assume that the brain makes predictions about the world, and in particular, about probabilistic relations (“if the sun is close to the horizon in the evening, it will soon be dark”; “if I turn the steering wheel to the right, the car will move to the right”). Learning is supposed to be a result of predictions that turn out to be incorrect or inaccurate, through the adaptation of underlying associations to reduce future prediction errors.

However, how this kind of learning works in case of adults learning a second language (L2) is not clear. Many L2 speakers are immersed in the L2 and thus receive ample correct input, yet often they do not improve beyond a certain proficiency level any more (so-called “L2 fossilization”). Error-driven learning models imply that L2 speakers make predictions about the incoming input and compare them to the actual input; learning would be driven by instances in which the two do not match. However, the premature learning asymptote seen in most adult L2 speakers questions whether this is really what they are, and keep, doing.

Combining ERP data from three previous projects, I raise the question whether learning-relevant input in an L2 is indeed compared to internally derived predictions, such that it can subsequently lead to learning, as well as what the electrophysiological correlates of this potential comparison process are. We did so both in the lexical and in the syntactic domain.

In Experiment 1*,* we looked at the moment of incidental *word learning* in L2. We recorded the EEG in native speakers of Dutch while they were incidentally exposed to previously unknown words in L2 English in a dialogue-like game with a ‘virtual’ (i.e. pre-recorded) partner (see de Vos et al., 2019, for a description of a similar paradigm). The ERPs show an enhanced late positive component (LPC) at the moment of hearing novel compared to already known words, an electrophysiological signature that is similar to what is generally observed for the recruitment of declarative memory resources. Furthermore, the LPC was even larger for those novel words that were subsequently successfully produced by the participant, compared to those that were not. This *subsequent memory effect* replicates other findings from (typically non-incidental) memory studies, but has so far not been observed for incidental L2 word learning.

In contrast, in Experiments 2 and 3, we looked at the moment of encountering corrective *syntactic* input. German learners of Dutch processed spoken or written sentences with a focus on comprehension. The sentences, which were all correct, contained both article-noun phrases (e.g., *het pistool*, the pistol) that the participants themselves had previously produced incorrectly (**de pistool*, driven by between-language gender incompatibility with German, *die Pistole),* as well as article-noun phrases for which their own production had been correct (e.g., *het huis*, which is gender-congruent with German *das Haus*). In Experiment 2, participants read the sentences for comprehension, while we again made use of a spoken dialogue game in Experiment 3. While there was clear behavioural evidence of learning (fewer errors after than before input), none of the two experiments showed any ERP effect of encountering a correct, but unexpected article-noun phrase (*het pistool)* as opposed to an also correct, but expected phrase (e.g., *het huis*). In particular, there was no evidence of an P600 that is standardly found for (outright) determiner violations. Thus, it seems that in contrast to the lexical findings of Exp. 1, the comparison process between a possible syntactic prediction and actual input either does not take place, or does not possess an ERP correlate. I would very much like to exchange thoughts about this unexpected finding on this extremely relevant and interesting workshop.

Reference

de Vos, J. F., Schriefers, H., ten Bosch, L., & Lemhöfer, K. (2019). Interactive L2 vocabulary acquisition in a lab-based immersion setting. *Language, Cognition and Neuroscience, 34*(7), 916-935. doi:10.1080/23273798.2019.1599127

A learning perspective on the emergence of abstractions

Benjamin V. Tucker (University of Alberta)
Dagmar Divjak (University of Birmingham)
Petar Milin (University of Birmingham)

Many theories of language presuppose the existence of abstractions with the aim to organize the extremely rich and varied experiences language users have. Evidence of how these abstractions would emerge from experience is, however, in many cases lacking: the amount of theoretical speculation about the existence of abstractions is inversely proportional to the amount of empirical work that has been devoted to providing evidence for or against the existence of such abstractions. This is specifically true for attempts to model computationally how such abstractions might emerge from exposure to input. We present a case study on the sounds of English in which we computationally model whether an abstract phone could emerge from exposure to speech sounds. In our study, each model was presented with over four hours of speech produced by one speaker, drawn from the MALD dataset (Tucker et al, 2019). We use two computationally very different approaches: Memory-Based and Error-Correction learning (MBL, ECL; for details see Milin, Divjak et al. 2016). For the latter, we use Widrow-Hoff (WH) and Temporal Difference learning (TD; a direct generalization of WH).

MBL, in essence, relies on a large pool of veridically stored experiences and a mechanism to match new ones with already stored exemplars, focusing on efficiency in storage and retrieval. The ECL models, on the other hand, learn to unlearn irrelevant dimensions of experience via the cue competition inherent in the experience. They focus on those dimensions that help deal with the environment on a ‘good enough' basis. What remains would then represent an abstraction: it is less detailed (or more schematic) than the input exemplars and more parsimonious because, as we learn, we discard or ignore what is non-predictive, i.e., not helpful for improving performance (by error-correction), and we retain only the relevant `core'. Importantly, however, that abstraction is not a given and static, but rather an ever-evolving information-rich residue of the experience (Nosofsky, 1986; Love, 2004; Ramscar, 2019).

We put both types of learning algorithms (MBL vs. ECL: WH and TD) through four tests that probe a variety of dimensions of the quality of their learning by assessing the models’ ability to predict data that is very similar to the data they were trained on (same speaker, different words) as well as data that is rather different from the data they were trained on (different speaker, same/different words). We also assessed the consistency or stability of what the models have learned and their ability to give rise to abstract categories. As expected, both types of models fare differently with regard to these tests. While MBL outperforms ECL in predicting new input from the same speaker, none of the models was able to generalize to input from a new speaker. The ECLs outperformed MBL in terms of consistency and performed well at the individual level rather than at the amalgamated, ‘population’ level. Finally, we found that ECL learning models can reliably identify at least part of the phone inventory. We conclude that abstractions of auditory experiences are, at least in part, learnable from input. The empirical exploration of whether abstractions are learnable should precede any decisions about whether to accept or reject them.

References

Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). Sustain: A Network Model of Category Learning. *Psychological Review*, 111 (2), 309-332.

Milin, P., Divjak, D., Dimitrijević, S., & Baayen, R. H. (2016). Towards cognitively plausible data science in language research. *Cognitive Linguistics*, *27*(4), 507-526.

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, 115 (1), 39-57.

Ramscar, M. (2019, March). Source codes in human communication (preprint). PsyArXiv. Retrieved 2020-05-27, from https://osf.io/e3hps

Tucker, B. V., Brenner, D., Danielson, D. K., Kelley, M. C., Nenadić, F., & Sims, M. (2019). The Massive Auditory Lexical Decision (MALD) Database. *Behavior Research Methods*, 51 (3), 1187–1204.

Can forcing second language learners to generate prediction errors increase learning?

Theres Grüter (U. of Hawai‘i), Yanxin (Alice) Zhu (U. of Hawai‘i), Carrie N. Jackson (Penn State U.)
theres@hawaii.edu

Effects of structural priming and adaptation have been argued to arise as a result of the computation of prediction error (ChangEtAl2006, Jaeger&Snider2013). Top-down factors such as explicit instructions to predict (BrothersEtAl2017) and social characteristics of the interlocutor (WeatherholtzEtAl2014) have been shown to modulate the size of prediction and priming effects. Within the context of second language (L2) acquisition, the view of priming as an (implicit) learning mechanism has led to the exploration of structural priming as a tool for L2 learning (McDonoughEtAl2015) and offered a potential theoretical framework for more unified study of L2 processing and L2 learning (Jackson&Hopp2020). Yet while of immediate relevance to applied and theoretical goals in L2 acquisition, the modulating roles of top-down factors such as explicit prediction and speaker characteristics on L2 priming and learning remain largely unexplored. We present evidence from two written production priming experiments with Korean L2 learners of English, focusing on double-object datives, to address the following questions:
RQ1: Do explicit task instructions to predict a partner's utterance increase effects of (i) immediate priming, and (ii) learning as measured by change from baseline to posttest?
RQ2: Do the partner's social and linguistic status as a native or non-native speaker affect the size of (i) immediate priming, and (ii) learning?

Method. In both experiments, participants in the 'guessing-game' (GG) group (Exp1: *n*=18, Exp2: *n*=27) had to predict how a virtual partner would describe a picture prior to seeing the actual prime sentence, which they then evaluated as the same or different from their initial guess (Fig1). This manipulation was intended to explicitly induce prediction and computation of prediction errors. Participants in the control group (CC; Exp1: *n*=17, Exp2: *n*=26) only re-typed the prime sentence in a standard repetition priming procedure (Fig2). The virtual partner consistently used double-object datives (DOs: *The girl fed the squirrel a nut*) with ditransitives, thus priming and adaptation should manifest in terms of increased use of DOs compared to prepositional datives (POs: *The girl fed a nut to the squirrel*), the strongly preferred construction for Korean learners (Kaan&Chun2018). The partner was presented as a native speaker of English ('Jessica') in Exp1 and as a Korean learner of English ('Soo-Min') in Exp2. In a baseline-priming-posttest design (Table1), participants alternated between repeating(CC)/guessing(GG) the partner's picture descriptions (primes) and describing pictures themselves (targets).

Results. Mixed logit models showed increases in DO production from baseline to priming phase in both experiments (*b*s>2, *p*s<.001; Fig3). While effects appeared numerically larger in GG vs CC groups, interactions with group were not robust (Exp1: *b*=1.32, *p*=.06; Exp2: *b*=.52, *p*=.3). Yet group significantly modulated change from baseline to posttest (Exp1: *b*=1.62, *p*=.03; Exp2: *b*=1.31, *p*=.006), with GG participants continuing to produce DOs more frequently than CC participants. While priming effects were numerical smaller in Exp2 than Exp1, experiment did not emerge as a robust modulator in a combined analysis of data from both experiments.

Discussion. In both experiments, explicit instructions to predict a partner's utterance (**RQ1**) led to greater learning in terms of change from baseline to posttest. Notably, the effect of this manipulation (GG/CC) only became robust in the posttest, suggesting it affected longer-term adaptation, or learning, more strongly than short-term activation of a primed structure. Future studies including delayed posttests will need to examine the longevity of this effect, yet this finding presents preliminary evidence to suggest that applied approaches seeking to use priming as a tool for error-driven L2 learning may benefit from incorporating a forced prediction component. Meanwhile, no clear evidence for modulation of L2 priming by social factors (**RQ2**) emerged. This is unexpected in light of findings showing native speakers adapt more to talkers using a more standard variety (WeatherholtzEtAl2014), but aligns with the only previous study of social factors in L2 structural priming (Chun&Kaan2020), which suggested such effects may be more complex than predicted by models based on data from native language processing.

References

Brothers, T., Swaab, T. Y., & Traxler, M. J. (2017). Goals and strategies influence lexical prediction during sentence comprehension. *Journal of Memory and Language, 93*, 203-216.

Chang, F., Dell, G., & Bock, K. (2006) Becoming syntactic. *Psychological Review, 113*, 234–72.

Chun, E., & Kaan, E. (2020). The effects of speaker accent on syntactic priming in second-language speakers. *Second Language Research.* https://doi.org/10.1177/0267658320926563

Jackson, C. N., & Hopp, H. (2020). Prediction error and implicit learning in L1 and L2 syntactic priming. *International Journal of Bilingualism, 24*, 895-911.

Jaeger, T. F., & Snider, N. E. (2013). Alignment as a consequence of expectation adaptation. *Cognition, 127*, 57–83.

Kaan, E., & Chun, E. (2018). Priming and adaptation in native speakers and second-language learners. *Bilingualism: Language and Cognition, 21*, 228-242.

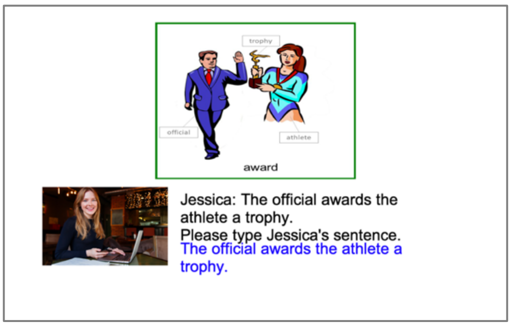
McDonough, K., Neumann, H., & Trofimovich, P. (2015). Eliciting production of L2 target structures through priming activities. *Canadian Modern Language Review, 71*, 75–95.

Weatherholtz, K., Campbell-Kibler, K., & Jaeger, T. F. (2014). Socially-mediated syntactic alignment. *Language Variation and Change, 26*, 387–420.

***Table 1**. Experiment design. (NB: no lexical boost)*

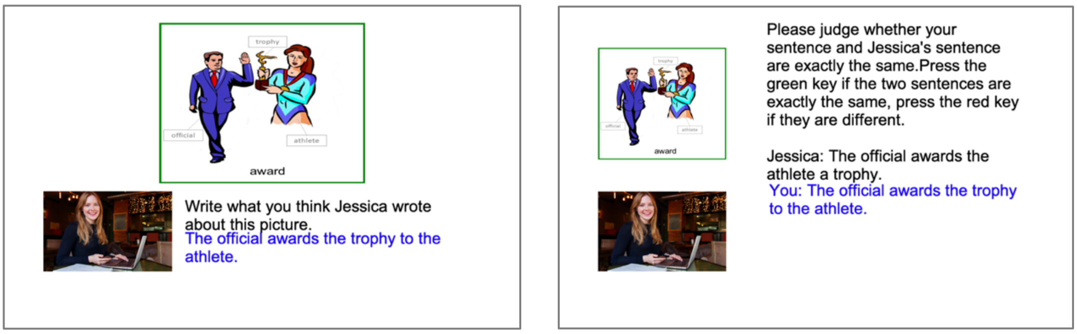
Phase	*Experimental items*	
	# and structure of prime-target pairs	
Baseline	6	prime: (in)transitive target: ditransitive
Priming	8	prime: ditransitive: DO target: ditransitive
Posttest	6	prime: (in)transitive target: ditransitive

***Figure 2**. Prime trial, CC condition (Exp1)*



Jessica: The official awards the athlete a trophy.
Please type Jessica's sentence.
The official awards the athlete a trophy.

***Figure 1.** Prime trial, GG condition (Exp1); sample participant responses in blue*

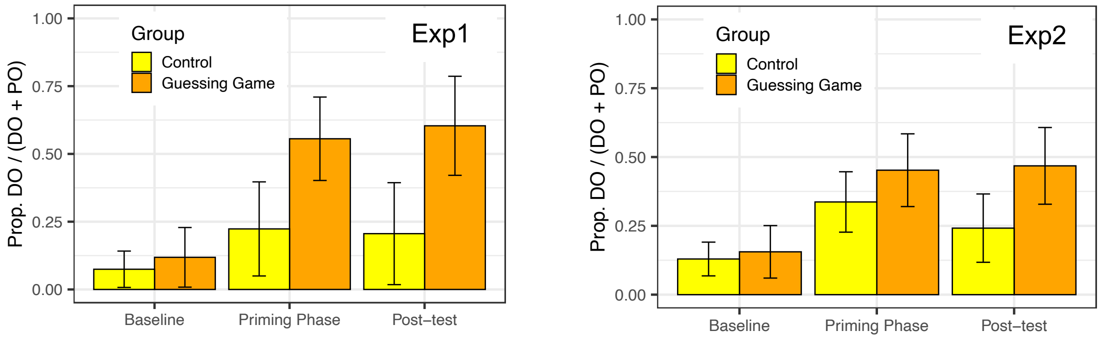


Write what you think Jessica wrote about this picture.
The official awards the trophy to the athlete.

Please judge whether your sentence and Jessica's sentence are exactly the same.Press the green key if the two sentences are exactly the same, press the red key if they are different.

Jessica: The official awards the athlete a trophy.
You: The official awards the trophy to the athlete.

***Figure 3.** Prop. DO/(DO+PO) by experiment, group and task phase. (Participant Ms, 95%CIs)*



Error-correction mechanisms in language learning: tracking individual differences

Adnane Ez-zizi, Dagmar Divjak and Petar Milin (University of Birmingham)

Over the past two decades, error-correction learning, and in particular the Rescorla-Wagner (1972) rule, has been successfully used to model a wide range of language phenomena (cf., Baayen et al., 2011; Milin et al., 2017; Baayen et al., 2019; Divjak, Milin et al., 2020). Previous studies have provided general support for the Rescorla-Wagner rule by using it to explain the behaviour of participants at the aggregate level, and most often by simulating it with default parameter values on a large-scale corpus. By contrast, our work (1) starts from general predictions generated by the model, then (2) tests them in a controlled semi-artificial language learning experiment, and finally (3) tracks how well the model captures the experimental data while taking into account individual differences in learning abilities as well as cognitive and personal characteristics.

Our experimental paradigm was inspired by the challenge of learning subject-verb agreement in the plural past tense in Polish. In the past tense, verbs are marked for the grammatical gender of the subject (*-li or -ły*) depending on the animacy (animate or personal) and grammatical gender (masculine or feminine) of the subject referents. For example, 'kobieta i koza chodziły' (the woman and goat were walking) but 'kobieta i Mężczyzna chodzili' (the woman and man were walking). Such a paradigm has the advantage of being straightforwardly modelled by the Rescorla-Wagner learning network.

Sixty-six native speakers of English participated in the experimental part of our study. The task consisted of a training and a test phase. In the training phase, participants were presented with 8 learning events, each repeated 15 times (for a total of 120 learning events). Each event consisted of a scene that represented the joint action of 'walking' performed by a group of human and/or animal characters, and for each learning event, participants saw a picture that depicts the scene, along with an audio recording of a Polish clause that describes it. In the test phase, each participant encountered 29 learning events that were either presented for the first time or already seen previously during the training phase. We collected both participants' choices and time latencies in the test phase, and assessed the Rescorla-Wagner model for its capacity to recover these, along with participants' levels of response agreement. We also compared the model to other plausible, yet rule-based response strategies.

Rather than fitting a one-for-all model using a single set of default parameters, we show that the model accurately captures participants' language learning when we adjust the learning rate parameter to fit the trial-by-trial behavioural choices of participants. The astounding success of the model is reflected across our three dimensions of interest: language-specific choices, time latencies and levels of response agreement. The model also outperforms all the other rule-based response strategies that we considered as competitors in capturing participants' behaviour in the task. Last but not least, we show that cognitive and personal characteristics such as working memory and gender affect the extent to which the Rescorla-Wagner rule captures language learning in our task.

References

Baayen, R. H., Milin, P., Đurđević, D. F., Hendrix, P., & Marelli, M. (2011). An amorphous model for morphological processing in visual comprehension based on naive discriminative learning. *Psychological review*, *118*(3), 438- 481.

Baayen, R. H., Chuang, Y. Y., Shafaei-Bajestan, E., & Blevins, J. P. (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. *Complexity*, *2019*.

Divjak, D., Milin, P., Ez-zizi, A., Józefowski, J., & Adam, C. (2020). What is learned from exposure: an error-driven approach to productivity in language. *Language, Cognition and Neuroscience*, 1-24.

Milin, P., Feldman, L. B., Ramscar, M., Hendrix, P., & Baayen, R. H. (2017). Discrimination in lexical decision. *PloS One*, *12*(2), e0171935.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning II: Current research and theory* (pp. 64-99). New York: Appleton-Century-Crofts.

How Distributional Context Solves the Variance Problem in Speech Sampling

Maja Linke, Michael Ramscar

Linguistics Department, University of Tuebingen
maja.linke@uni-tuebingen.de, michael.ramscar@uni-tuebingen.de

Numerous results have shown that language learners are sensitive to the probabilistic structure of the input they experience. These findings are often taken to imply that languages are themselves probabilistic systems. However, in combination with the fact that linguistic distributions tend to exhibit power-law behavior (Estoup, 1916; Zipf, 1949) and the questions this raises about the representativeness of language samples (Baayen, 2002), any notion of probabilistic presents a problem to both theoretical and computational accounts of learning. The idea that language is a probabilistic system implies that all users share a model of its probabilities. Simultaneously, the nature of linguistic distributions guarantees that individual speakers will experience and learn from very different samples of their linguistic environments. How are any two language users ever capable of converging on the same model of those parts of the system they have both been exposed to?

Speakers clearly learn to use language in context, and recent results have shown that when lexical distributions are considered in the communicative contexts in which they occur, their probabilistic structure is geometric (Ramscar, 2019, 2020). A critical property of the geometric (and other memoryless) distributions is that its structure appears to support a transmission process that is impervious to sampling differences, and simulation studies have shown that in contrast to other word distributions, when random samples are drawn from word categories that approximate geometric distributions they do in fact yield representative subsamples (Linke & Ramscar, 2020). A key implication of these findings is that they indicate that although individual word recurrence rates are highly irregular, the distributions of words in the contexts in which they are encountered (and learned) supports the learning of models of their probabilities that are largely independent of the samples individual learners experience. This allows speakers exposed to different samples of different sizes at different rates to nevertheless learn probabilistic models that enable them to establish and maintain similar linguistic expectations.

Previous analyses reveal that the distributions defined by context in a number of samples of English satisfy this property at multiple levels of description. To examine whether support for this model of probabilistic alignment could also be found in structurally simpler and seemingly more disorganized speech samples produced by and directed at children, we conducted a simulation study of the sampling properties on a set of nouns drawn from the CHILDES-EN corpus (MacWhinney, 2000).

We sampled nouns from 29 distributional clusters *family*, *mammal*, etc. used in learning simulations by Asr, Willits, and Jones (2016) and a random set of nouns drawn from the corpus, gradually increasing the sample size. For each of the 900 subsets we computed the distance between the sample distributions and the complete set by comparing linear model fits and the Kullback-Leibler divergence. The results of these simulations confirm that the distributions of nouns in child/caregiver speech are geometric and that random subsamples drawn from these speech samples yield nearly identical probability distributions independent of the subsample size. Our results show how the distributions of forms in child/caregiver speech solve the problem of alignment in the language learning process. Moreover, the distributions observed in the sample provide further support for the suggestion that human communicative codes are structured in a way that maximizes alignment between speakers independent of the time and the rate at which they are exposed to speech.

These findings provide further evidence that the power law distributions observed in human languages result from the aggregation of functionally distinct exponential distributions (Mitzenmacher, 2004; Newman, 2005; Ramscar, 2019; Reed, 2001). These findings further suggest that empirically, lexical distributions form a nested information structure (a geometric distribution of geometric distributions), offering a formal explanation of this phenomenon.

This helps to maintain the efficiency of speech and facilitate the diverse communicative repertoire of humans, while ensuring the transmission of communicative codes across generations.

References

Asr, F. T., Willits, J., & Jones, M. N. (2016). Comparing predictive and co-occurrence based models of lexical semantics trained on child-directed speech. In *Cogsci.*

Baayen, R. H. (2002). *Word frequency distributions* (Vol. 18). Springer Science & Business Media.

Estoup, J.-B. (1916). *Gammes sténographiques: méthode et exercices pour l'acquisition de la vitesse*. Institut sténographique.

Linke, M., & Ramscar, M. (2020). How the probabilistic structure of grammatical context shapes speech. *Entropy*, *22*(1), 90.

MacWhinney, B. (2000). *The childes project: The database* (Vol. 2). Psychology Press.

Mitzenmacher, M. (2004). Dynamic models for file sizes and double pareto distributions. *Internet Mathematics*, *1*(3), 305–333.

Newman, M. E. (2005). Power laws, pareto distributions and zipf's law. *Contemporary physics*, *46*(5), 323–351.

Ramscar, M. (2019). Source codes in human communication. *arXiv preprint arXiv:1904.03991*.

Ramscar, M. (2020). The empirical structure of word frequency distributions. *arXiv preprint arXiv:2001.05292*.

Reed, W. J. (2001). The pareto, zipf and other power laws. *Economics letters*, *74*(1), 15–19.

Zipf, G. K. (1949). Human behavior and the principle of least effort: an introd. to human ecology.

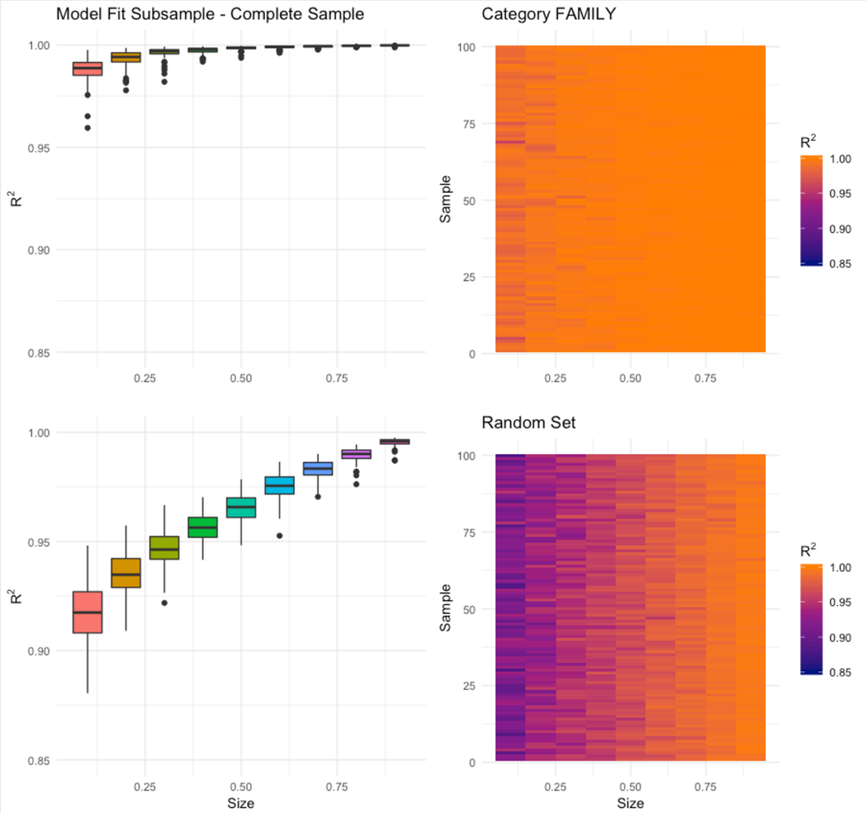


Figure 1. Alignment between subsample probability distributions and the complete sample for nouns from the distributional category *family* (top row) and a random set of nouns extracted from the corpus (bottom row). The average (left) and the individual (right) model fits show that while subsamples in context converge on a representative probabilistic structure after minimal exposure, random noun samples require to have seen 70% of the data to reach the threshold.

Error-driven learning as a mechanism for cross-language structural priming

Yung Han Khoe[1], Chara Tsoukala[1], Gerrit Jan Kootstra[1] and Stefan L. Frank[1]

[1]Centre for Language Studies (CLS), Radboud University, Nijmegen, The Netherlands

Structural priming is the tendency of speakers to reuse syntactic structures that they have recently encountered. It also occurs between different languages, as has been shown in behavioral experiments as well as corpus studies. The underlying mechanism is unclear, but error-driven learning has been proposed for within-language priming (Chang, 2002). Here, we investigate if the same mechanism can account for cross-language priming, by simulating priming of active and passive structures in bilingual models.

We implemented a Spanish-English and two Dutch-English models of balanced bilingual sentence production. The models are based on the Bilingual version (Tsoukala et al., 2021) of the Dual-path recurrent neural network (RNN) model of sentence-production (Chang, 2002), which is trained to incrementally generate sentences in an artificial, miniature language based on a natural language. Three of the artificial languages that we used (Spanish, English, and one of two versions of Dutch) had verb-medial passives, while the other version of Dutch had verb-final passives. The model implements priming as error-driven (backpropagation) learning from the prime sentence. For each of the three models, we performed a preregistered priming experiment with 80 model participants. These experiments consisted of 800 trials, which were balanced for prime and target languages, and for prime structure (active or passive).

Cross-language priming of transitives occurred in all three models, which provides evidence for the viability of an error-driven learning account of cross-language structural priming. Fig. 1 illustrates this for the Spanish-English experiment where more passive targets were produced after a passive than after an active prime, both for within-language and cross-language trials. Similar results for priming between verb-final Dutch and verb-medial English show that identical word order is not required for this priming effect, which has also been established in behavioral experiments. Our results revealed varying degrees of evidence for stronger within-language priming than cross-language priming. This is consistent with the conflicting human experimental findings where within-language priming is found to be significantly stronger in some studies but not in others.

To investigate whether these results critically depend on the artificial nature of the modelled languages, we also trained five RNNs for next-word prediction on a naturalistic corpus of ~18M sentences of Dutch and English (50% in each language). Following Van Schijndel & Linzen (2018), we then primed and tested them on locally ambiguous (garden-path) structures (1) or their non-ambiguous counterparts (2), using 120 Dutch sentence stimuli from Hoeks et al. (2006) and their English translations.

(1) The thief shot the jeweler and the cop **risked** his life.
(2) The thief shot the jeweler, and the cop risked his life.

As shown in Fig. 2, surprisal on the critical, disambiguating verb ("risked") was higher in the ambiguous sentences, simulating the garden-path effect. More importantly, the effect was stronger after priming with an ambiguous than unambiguous sentence, both within- and cross-language. This mirrors the Dual-Path model results, but with more realistic training and test data.

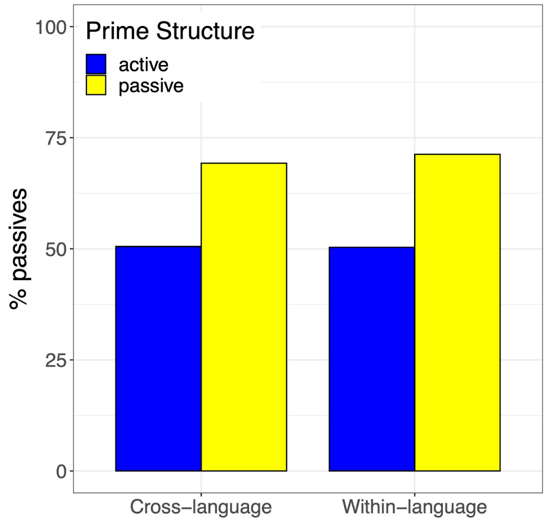


Fig. 1: Results from the Spanish-English model: Percentage of responses with a passive structure after either an active or a passive prime, split by within- or cross-language trials. Similar results were obtained from the two Dutch-English models.

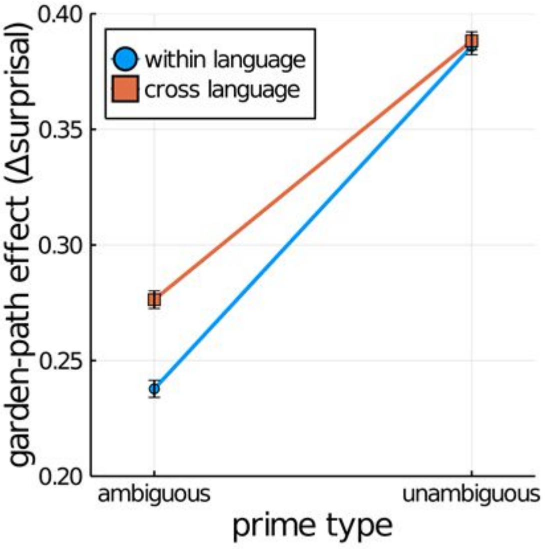


Fig. 2: Size of the garden-path effect (surprisal in ambiguous minus unambiguous structures) as a function of prime type (ambiguous or unambiguous structure) and language combination (within- or cross-language priming). Error bars denote 95% confidence intervals.

References

Chang, F. (2002). Symbolically speaking: A connectionist model of sentence production. *Cognitive Science*, *26*(5), 609–651.

Hoeks, J., Hendriks, P., Vonk, W., Brown, C., & Hagoort, P. (2006). Processing the noun phrase versus sentence coordination ambiguity: Thematic information does not completely eliminate processing difficulty. *The Quarterly J. of Experimental Psychology, 59*, 1581–1599.

Tsoukala, C., Broersma, M., Van den Bosch, A., & Frank, S.L. (2021). Simulating code-switching using a neural network model of bilingual sentence production. *Computational Brain & Behavior, 4,* 87–100.

Van Schijndel, M. & Linzen, T. (2018). A Neural Model of Adaptation in Reading. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*

Interaction of semantic and syntactic ambiguity in the light of discrimination learning

Ksenija Mišić & Dušica Filipović Đurđević
Department of Psychology, Faculty of Philosophy, University of Belgrade, Serbia
Laboratory for Experimental Psychology, Faculty of Philosophy, University of Belgrade, Serbia
ksenija.misic@f.bg.ac.rs, dusica.djurdjevic@f.bg.ac.rs

In this research we wanted to investigate whether the effects observed during simultaneous processing of semantic and syntactic lexical ambiguity could be interpreted within the framework of error-driven learning. To investigate the interaction of the two types of ambiguity we focused on polysemy in a highly inflected Serbian noun system and attempted to simulate their processing effects using naive discriminative learning (Baayen et al., 2011).

Polysemy is the type of lexical ambiguity where one word can have multiple related senses. For example, isolated word *paper* could refer to the writing paper, i.e., paper as the material, but also to scientific paper, and even a daily paper. Polysemous words are a complex phenomenon, whose processing is affected by number of senses, probability distribution of those senses, and degree of relatedness among the senses (Filipović Đurđević & Kostić, 2009; 2017, under revision; Klepousniotou, 2002; Rodd et al., 2002). Additionally, in Serbian, words can take up to seven inflected forms. Syntactic ambiguity of an isolated inflected form is reflected in the multitude of syntactic roles the given inflected form can take in the sentence. For example, inflected masculine noun *konja* (horse) can indicate the subject in the sentence (*Dva konja su trčala / Two horses were running*), but also the object (*Jahao sam konja / I rode the horse*). It has been demonstrated that different aspects of syntactic ambiguity affect lexical processing as well: information load based on relative frequency of the inflected form within its inflectional class and the number of syntactic functions and meanings (Kostić, 1991), inflectional entropy (Baayen et al, 2006), relative entropy (Milin et al., 2009), etc.

By definition, polysemous words are equally ambiguous in all of the inflected forms (Gortan-Premk, 2004). Hence, the inflected form does not serve as the cue for their true meaning. However, some research suggested that meaning can shape the way a noun is used in a sentence (Kostić et al., 2003).

We presented 35 polysemous nouns of masculine gender in a visual lexical decision task to 74 participants (data collection still ongoing). Each noun was presented in one of its seven forms in a latin-square design. Entropy of the sense frequency distribution was estimated in a norming study (Filipović Đurđević & Kostić, 2017). Relative frequencies of inflected forms and the number of syntactic functions and meanings were taken from Kostić (1965). Discrimination learning based predictors were derived from cue-outcome weights matrix calculated by equilibrium equations (Danks, 2003), implemented in the *ndl* package (Arppe, et al. 2015). Cues were bigrams of the polysemous words' inflected forms presented in the experiment. Outcomes were their lemmata and 1000 co-occurring context words.

We modelled processing times by applying GAMMs (Wood, 2006) and compared two models. Information-theoretic model revealed the expected facilitatory effect of entropy of word senses, but only in the nominative form. Discriminative predictors (Milin et al, 2017) affected processing in some word forms, again revealing the interaction. AIC comparison showed that the discrimination-based model was superior to the information-theoretic one. Correlations between the two sets of measures revealed some interesting relations. Activation was more sensitive to semantic ambiguity, whereas diversity captured both semantic and syntactic ambiguity. This suggests a more complex interplay of semantics and morpho-syntax than previously thought and the possibility of capturing such an interaction with discrimination-based diversity measures.

Arppe, A., Hendrix, P., Milin, P., Baayen, R. H., Sering, T., & Shaoul, C. (2015). *ndl: Naive Discriminative Learning* (R package version 0.2.17).

Baayen, R. H., Feldman, L. B., & Schreuder, R. (2006). Morphological influences on the recognition of monosyllabic monomorphemic words. *Journal of Memory and Language, 55*(2), 290–313.

Baayen, R. H., Milin, P., Đurđević, D. F., Hendrix, P., & Marelli, M. (2011). An amorphous model for morphological processing in visual comprehension based on naive discriminative learning. *Psychological Review, 118*(3), 438–481.

Danks, D. (2003). Equilibria of the Rescorla - Wagner model. J*ournal of Mathematical Psychology, 47*(2), 109–121.

Filipović Đurđević, D., Đurđević, Đ., & Kostić, A. (2009). Vector based semantic analysis reveals absence of competition among related senses. *Psihologija, 42*(1), 95–106.

Filipović Đurđević, D., & Kostić, A. (2017). Number, Relative Frequency, Entropy, Redundancy, Familiarity, and Concreteness of Word Senses: Ratings for 150 Serbian Polysemous Nouns. In S. Halupka-Rešetar & S. Martinez-Ferreiro (Eds.), *Studies in Language and Mind* (pp. 13–77). Filozofski fakultet u Novom Sadu.

Filipović Đurđević, D., & Kostić, A. We probably sense sense probabilities. Under review.

Gortan-Premk, D. (2004). *Polisemija i organizacija leksičkog sistema u srpskome jeziku*. [*Polysemy and the Organization of the Lexical System in Serbian Language*] Zavod za udžbenike i nastavna sredstva.

Klepousniotou, E. (2002). The Processing of Lexical Ambiguity: Homonymy and Polysemy in the Mental Lexicon. *Brain and Language, 81*(1–3), 205–223.

Kostić, A. (1991). Informational approach to the processing of inflected morphology: Standard data reconsidered. *Psychological Research, 53*(1), 62–70.

Kostić, A., Marković, T., & Baucal, A. (2003). Inflectional morphology and word meaning: Orthogonal or co-implicative cognitive domains?. In Baayen H. Schreuder R.[eds.] *Morphological Structure in Language Processing*, Berlin: Mouton de Gruyter.

Kostić, Đ. (1999). *Frekvencijski rečnik savremenog srpskog jezika*. [*Frequency Dictionary of the Contemporary Serbian Language*]

Milin, P., Feldman, L. B., Ramscar, M., Hendrix, P., & Baayen, R. H. (2017). Discrimination in lexical decision. *PLoS ONE*, 1–42.

Milin, P., Filipović Đurđević, D., & Moscoso del Prado Martín, F. (2009). The simultaneous effects of inflectional paradigms and classes on lexical recognition: Evidence from Serbian. *Journal of Memory and Language, 60*(1), 50–64.

Rodd, J. M., Gaskell, G., & Marslen-Wilson, W. (2002). Making Sense of Semantic Ambiguity: Semantic Competition in Lexical Access. *Journal of Memory and Language, 46*(2), 245–266.

Learning trajectories in L2 and bilingual language development: a structural priming investigation

Michaela Vann[1], Giulia Bencini[1], Virginia Valian[2,3]

[1] Ca' Foscari University, Venice
[2] Hunter College, City University of New York
[3] Graduate Center, City University of New York

Previous studies have shown that bilinguals exhibit within-language structural priming and greater sensitivity to lexical overlap between prime and target at lower proficiency levels suggesting that the development of syntax in a second language (L2) goes from lexically specific to shared abstract representations[2,5,6]. What is unknown is the extent to which L2 syntactic production is sensitive to semantic constraints on the mappings from thematic roles to syntactic positions throughout L2 development. We use a syntactic priming paradigm to examine this question.

Method.

Participants: We tested 375 participants (295 bilinguals and second language learners). English language proficiency was measured with an objective test (the grammar portion of the Michigan Test of English Language Proficiency) and with self-assessment ratings in the four modalities (speaking, listening, reading, writing).

Design and Procedure: Participants heard 16 transitive (8 active, 8 passive) and 12 ditransitive (6 double object dative (DO), 6 prepositional dative (PD)) priming sentences and described 16 images of transitive and 12 images of ditransitive events during a spoken to written cross-modal syntactic priming paradigm task[3]. Sensitivity to semantic-conceptual features was examined via an animacy manipulation with prototypical vs. non prototypical animacy mappings to thematic roles in passive sentences. We used non prototypical passive primes and targets with inanimate agents and patients (e.g., the milk is stirred by the spoon) and compared them to prototypical primes and targets with inanimate agents and animate patients (e.g., the man is chased by the dog). Sensitivity to lexical repetition was examined via a verb match manipulation for the ditransitive trials (same verb vs. different verb).

Results.

We analyzed participants' responses with logistic mixed-effects models in the lme4 package in R[1], predicting the logit-transformed likelihood of the production of passives and DOs. Proficiency was a continuous composite z-score obtained by combining the objective and subjective proficiency measures. Tables 1 and 2 show the best fit models for passives and DOs, respectively. The best fit model for passives included a significant interaction between proficiency and semantics (prototypical animacy vs. non prototypical animacy mappings). Figure 1 shows the priming effect (passives produced after passive primes minus passives after active primes) for the two animacy conditions. As can be seen from Figure 1, at lower proficiency levels L2 speakers' priming effects are similar for prototypical and non prototypical passives. As proficiency increases, L2 speakers produce fewer non prototypical passives. The best fit model for DOs included significant interactions between proficiency and verb match and verb match and prime structure. Figure 2 shows the proportion of DOs produced in the four cells of the design, as a function of proficiency. As can be seen from Figure 2, at the lowest levels of proficiency, participants only produce DOs after DO primes with the same verb. The figure also shows a larger lexical boost (DOs after DOs with the same verb minus DOs after DOs with a different verb) at lower proficiency levels.

Discussion.

Our results confirm previous studies with L2 speakers that report a larger lexical boost at lower proficiency levels. The same participants, however, showed that at lower proficiency levels, L2 speakers are less sensitive to the semantic constraints in prototypical passives and are more primed to produce non prototypical passives after non prototypical passive primes. We discuss our findings with respect to current models of language development and learning trajectories in L2 compared to L1 and with reference to different accounts of priming[4,5].

Table 1. Summary of fixed effects in the best fit mixed logit model for priming of passives.

Fixed effects	Estimate	SE	*z* value	95% CI	*p*-value
Intercept	-0.45	0.21	2.14	-.87 to -.04	< .01
Prime Structure	1.53	0.08	18.48	1.372 to 1.70	< .0001
Animacy	1.22	0.21	5.91	.82 to 1.63	< .0001
Proficiency	-.17	0.06	-2.58	-0.29 to -0.04	<.001
Prime Structure x Animacy	-0.08	0.08	-0.97	-.24 to 0.08	n.s.
Prime Structure x Proficiency	0.01	0.06	0.08	-0.12 to 0.13	n.s.
Animacy x Proficiency	0.13	0.06	2.39	0.02 to 0.24	< .01

Table 2. Summary of fixed effects in the best fit mixed logit model for priming of DOs.

Fixed effects	Estimate	Standard Error	z-value	95% CI	*p*-value
Intercept	-1.25	0.31	-3.98	-1.87 to -0.63	< .0001
Prime Structure	1.44	0.12	11.87	1.20 to 1.67	< .0001
Verb Match	.09	0.31	0.31	-0.51 to 0.7	n.s
Proficiency	.92	.11	8.47	0.71 to 1.13	< .0001
Prime Structure x Verb Match	.32	.11	2.9	.10 to .54	< .001
Prime Structure x Proficiency	-0.05	0.09	-0.55	-0.22 to 0.12	n.s.
Verb Match x Proficiency	-0.20	0.06	-2.97	-0.34 to -0.07	< .001

Figure 1. Priming effect of Passives across proficiency levels by animacy

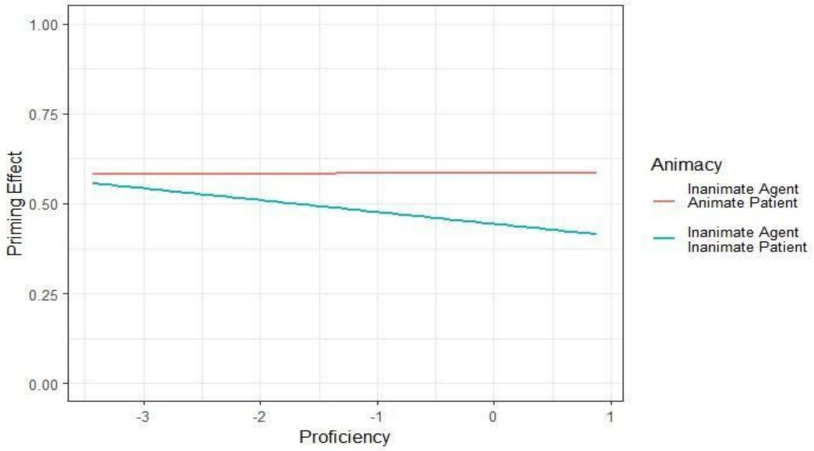
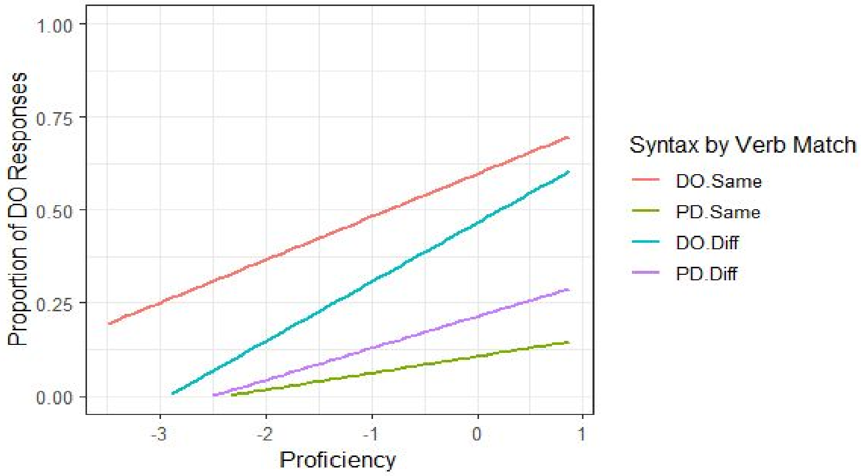


Figure 2. Proportion of DOs across proficiency by Priming Condition



References

[1] Bates, D. M. (2010). lme4: Mixed-effects modeling with R. Available online at<http://lme4.r-forge.r-project.org/book/> .

[2] Bernolet, S., Hartsuiker, R. J., & Pickering, M. J. (2013). From language-specific to shared syntactic representations: The influence of second language proficiency on syntactic sharing in bilinguals. *Cognition*, *127*(3), 287–306.

[3] Bock, J. K., Dell, G.S., Chang, F., & Onishi, K.H. (2007). Persistent structural priming from language comprehension to language production. Cognition, 104, 437-458.

[4] Chang, F., Dell, G.S., & Bock, J. K. (2006). Becoming syntactic. Psychological Review, 113, 234- 272.

[5] Hartsuiker, R. J., Pickering, M.J., & Veltkamp, E. (2004). Is syntax separate or shared between languages? Cross-linguistic syntactic priming in Spanish-English bilinguals. Psychological Science, 15, 409-414.

[6] Hartsuiker, R. J., & Bernolet, S. (2017). The development of shared syntax in second language learning. *Bilingualism*, *20*(2), 219–234.

Less is more? Language learning, between simple and deep embeddings

Harish Tayyar Madabushi[1] and Dagmar Divjak[2,3] and Petar Milin[2]
[1]School of Computer Science
[2]Department of Modern Languages
[3]Department of English Language and Linguistics
University of Birmingham,
UK
Harish@HarishTayyarMadabushi.com, D.Divjak@bham.ac.uk, P.Milin@bham.ac.uk

The past decade has seen dramatic improvements in the field of computational linguistics across several tasks such as question answering, translation and summarisation. These gains are a direct result of the recent infusion of Machine Learning (ML) methods into traditional Natural Language Processing (NLP). To make this integration a reality, NLP research has had to focus on methods of representing words as numbers (vectors), to be able to feed text to learning machines.

Numeric representations of text (embeddings) can be agnostic of context, as in the case of GloVe (Pennington et al., 2014), providing the same representation for words with more than one meaning (eg. bank as in riverbank or a financial institution). Alternatively, contextual or dynamic embeddings, a prime example of which is BERT (Devlin et al., 2018), provide dynamically changing embeddings based on context. Both types of embeddings have recently become notorious for their success in a range of applications. While these word embeddings reflect Firth's observation: “You shall know a word by the company it keeps” (Firth, 1957:11), they do not do so in a way that is cognitively plausible in a strict sense.

Conversely, one of the earliest biologically (i.e., neurologically) inspired models of learning is the Widrow-Hoff (1960) rule (WH; also known as the Delta or Least Mean Square rule). Arguably, it is the simplest of all learning rules, yet, over decades, it has shown versatility and has been successful ever since in a range of practical applications (e.g., noise cancellation in telephone lines which is used to date; cf., Haykin, 1999).

In our talk we first present the rule in some detail followed by a discussion on the implementation challenges posed by the need to model learning over large datasets, and with many inputs and outputs. Next, we pit this simple learning principle against the state-of-the-art embedding framework - GloVe. In two case studies, conducted on a set of Russian verbs and English connectors, we show that the WH rule does an exceptional job in modelling the learnability or predictability of said target forms, lemmata, and high-level notions (i.e., ‘meanings’).

Our results show that WH learning weights are reasonable representations of words even when compared with GloVe, a model trained on a significantly larger corpus. Crucially, however, the present results reflect learning that is biologically or cognitively plausible. We also show that Widrow-Hoff learning weights help filter the signal from noisy input. In a sense, they detect strong regularities in highly variable input. We conclude that the WH rule, as simple and shallow as it is, can in fact go far and reach deep. It appears eminently suited for the investigation of the emergence of abstractions in language, while remaining faithful to its inspirations from biological networks.

References

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP).

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of Deep Bidirectional Transformers for Language Understanding. preprint arXiv:1810.04805.

Firth, J. R. (1957). Papers in Linguistics, 1934-1951 Oxford University Press

Widrow, B., & Hoff, M. E. (1960). Adaptive switching circuits. Paper presented at the WESCON Convention Record Part IV.

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. The bulletin of mathematical biophysics, 5(4), 115-133.

Haykin, S. S. (1996). Adaptive filter theory. Upper Saddle River, NJ: Prentice Hall.

Uncertainty of polysemous word senses in the light of discrimination learning

Dušica Filipović Đurđević
Department of Psychology, Faculty of Philosophy, University of Belgrade, Serbia

dusica.djurdjevic@f.bg.ac.rs

This research will present an attempt to simulate the processing effects of polysemy using the model based on discrimination learning (Baayen, et al., 2011) thus showing that lexical-semantic processing can be described using the principles of error-driven learning.

Polysemous words denote multiple related senses (e.g. scientific paper, daily paper, wrapping paper, etc.; Eddington & Tokowicz, 2015). The starting point of this research is the finding that the processing of polysemous words is affected both by the number of senses and by the balance of sense probabilities. It has been shown that the processing latencies of the polysemous words decrease as the number of senses increases and as the redundancy of sense probability distribution decreases (i.e. as the balance of sense probabilities increases; Filipović Đurđević, 2007; Filipović Đurđević & Kostić, under review).

Here, 150 polysemous words for which processing latencies were previously collected were split into bigrams which served as the input to the model, i.e. the cues. At the output level, each set of bigrams constituting one word form was linked to its corresponding lemma and to the co-occurring context words, i.e. the context words which appeared within the -/+3 window surrounding the target word form. The context words were preselected from the Frequency Dictionary of the Contemporary Serbian Language (Kostić, 1999). We started with 3000 most frequent nouns, adjectives, and verbs (1000 each), and ended with 2383 context words after excluding the homographs. We started by building first-order co-occurrence vectors (Schütze, 1998) for 150 polysemous words which were presented in the experiment. Separate vectors were built for each occurrence of the word, each vector consisting of the zeros (0) for the context words that were not found within the seven-point window, and the ones (1) for the context words that co-occurred with the target word. This information was then used to represent the lemma followed by the co-occurring context words as the outcomes. The simulation was run in R (R CoreTeam, 2017), using ndl package (Arppe et al., 2015), following the procedure described in Baayen et al., (2011). The activations were calculated for each outcome by summing the strengths of all the bigrams present in the target word. Finally, the corresponding activations for the lemma and the co-occurring context words were summed. These activations were taken as the indicator of the strength of support for the given outcome by the cues which were present in the input. The given outcome consisted of lemma and co-occurring context words.

The calculated activations were significantly correlated both with processing latencies observed in the experiment and with descriptors of lexical ambiguity. We observed negative correlation between activations and processing latencies (r=-.42, t(144)=5.639, p<.001), positive correlation between number of senses and activation (r=.18, t(144)=2.229, p=.027), and negative correlation between redundancy of sense probability distribution and activation (r=-.23, t(144), p=.004). However, when we performed multiple linear regression with several lexical variables in addition to the number of senses and redundancy as predictors of activation, only redundancy accounted for the activation variance over and above the contribution of familiarity, concreteness, and orthographic neighborhood size.

This finding brings evidence that the effect of balance of sense probabilities can be simulated in a model based on the principles of discrimination learning. In other words, it demonstrates that semantic ambiguity effects can arise through error-driven learning.

Arppe, A., Hendrix, P., Milin, P., Baayen, R.H., Sering, T., & Shaoul, C. (2015). *ndl: Naive Discriminative Learning. R package version 0.2.17.* http://CRAN.R-project.org/package=ndl

Baayen, R. H., Milin, P., Filipović Đurđević, D., Hendrix, P., & Marelli, M. (2011). An amorphous model for morphological processing in visual comprehension based on naïve discriminative learning. *Psychological Review, 118*, 438–482. https://doi.org/10.1037/a0023851

Eddington, C. M., & Tokowicz, N. (2015). How meaning similarity influences ambiguous word processing: the current state of the literature. *Psychonomic Bulletin & Review*, *22*(1), 13–37. http://doi.org/10.3758/s13423-014-0665-7

Filipović Đurđević, D. (2007). *Polysemy effect in processing of Serbian nouns*. PhD thesis, University of Belgrade.

Filipović Đurđević, D., & Kostić, A. We probably sense sense probabilities. Under review.

Kostić Đ. (1999). *Frekvencijski rečnik savremenog srpskog jezika*. Beograd: Institut za eksperimentalnu fonetiku i patologiju govora i Laboratorija za eksperimentalnu psihologiju.

R Core Team (2017). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Second language learners' sensitivity to competing alternatives is modulated by proficiency: Evidence from L2 Mandarin

Yanxin (Alice) Zhu (U. of Hawai‘i), Yang Zhao (Peking U.), Theres Grüter (U. of Hawai‘i)

Tachihara & Goldberg (2020) provided evidence suggesting that L2 English speakers have reduced sensitivity to competing alternatives, and thus more readily accept novel verb-construction pairings than native speakers do. Tachihara & Goldberg suggested that this might be due to L2 learners' reduced capacity to generate expectations in sentence processing (e.g., Grüter et al., 2017; Kaan, 2014). If so, L2 learners will have less opportunity to learn from prediction errors that would potentially occur when the observed output does not match the predicted output.

Tachihara & Goldberg found only weak support for the hypothesis that L2 learners' sensitivity to competing alternatives would be modulated by L2 proficiency, using a self-report measure. The current study aims to extend these findings to L2 Mandarin speakers and to further explore the role of proficiency as measured through a cloze test. The study addressed two research questions:

RQ1: Do L2 Mandarin speakers rate novel combinations of verbs and ditransitive constructions (see Table 1) as more acceptable than native speakers do?

RQ2: Does Mandarin proficiency modulate their ratings?

Eighty Mandarin learners (40 L1-English and 40 L1-Japanese) and 20 native Mandarin speakers participated in the study.

As illustrated in Figure 1, L2 speakers judged novel combinations as more acceptable than L1 speakers did. However, L2ers also judged conventional combinations as less acceptable than L1 speakers did. Opposite to Tachihara & Goldberg's (2020) results, the difference between L1 and L2 speakers' judgements for conventional sentences (β = -4.35) was larger than the difference between the two groups for novel sentences (β = 1.54), suggesting that general uncertainty played a bigger role than lack of sensitivity to competing alternatives in explaining the divergence between judgements in the two groups in this study.

However, we also observed a significant interaction between sentence type and cloze test scores within the L2 group (β = -0.60, z = -4.21, p <.001., Figure 2). More detailed analysis revealed that L2 learners' acceptance for novel sentences significantly decreased as their proficiency increased (β = -0.35, z = -2.68, p = .007.), whereas their acceptance for conventional sentences did not increase significantly with increasing proficiency (β = 0.34, z = 1.44, p = .15). This suggests that sensitivity to competing alternatives may be more strongly modulated by proficiency than general uncertainty is. The overall stronger effects of proficiency in the present study compared to those in Tachihara & Goldberg (2020) suggest that cloze tests may better capture the relevant aspects of proficiency than self-report.

If it is true that learning to rule out novel formulations is dependent on error-based learning, while learning to accept conventional sentences just needs entrenchment, we can infer from the interaction between sentence type and cloze test scores that proficiency modulates error-based learning in particular. Robenalt & Goldberg (2016) also found that L2 learners at the highest proficiency levels showed sensitivity to competing alternatives, like native speakers did.

Table 1. Examples of novel sentences (i.e., unattested in Mandarin) and conventional competing alternatives with two ditransitive verbs *ji* ('send') and *gaosu* ('tell'). NB: While Mandarin has both prepositional (PO) and double-object (DO) dative constructions, SEND-type verbs can only appear in PO and TELL-type verbs only in DO construct*ions (Liu, 2006).*

Novel sentences	*Conventional competing alternatives*
DO *Mali **ji** le Dawei yi feng xin. Mary **send** ASP David a CL letter ‘Mary **sent** David a letter.’	PO Mali **ji** le yi feng xin gei Dawei. Mary **send** ASP a CL letter to David ‘Mary **sent** a letter to David.’
PO *Mali **gaosu** le yi ge mimi gei Dawei. Mary **tell** ASP a CL secret to David ‘Mary **told** a secret to David.’	DO Mali **gaosu** le Dawei yi ge mimi. Mary **tell** ASP David a CL secret ‘Mary **told** David a secret.’

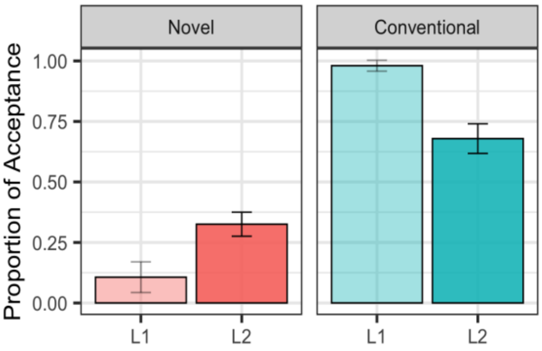


Figure 1. Proportion of acceptance for novel vs. conventional sentences by group.

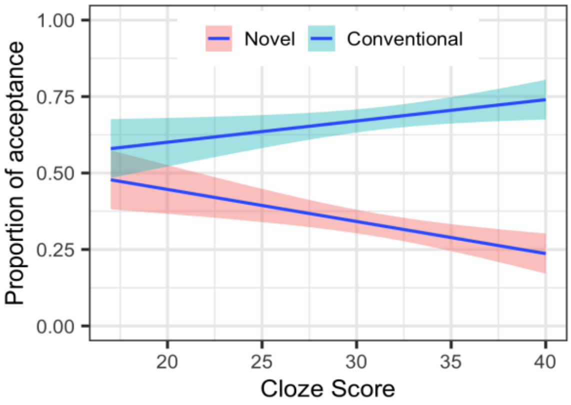


Figure 2. L2 learners’ acceptance for novel sentences significantly decreased as their proficiency increased.

References

Grüter, T., Rohde, H., & Schafer, A. J. (2017). Coreference and discourse coherence in L2: The roles of grammatical aspect and referential form. *Linguistic Approaches to Bilingualism, 7*(2), 199–299.

Kaan, E. (2014). Predictive sentence processing in L2 and L1: What is different? *Linguistic Approaches to Bilingualism, 4,* 257–282.

Liu, F. (2006). Dative constructions in Chinese. *Language and Linguistics 7(*4), 863–904.

Robenalt, C., & Goldberg, A. E. (2016). Nonnative speakers do not take competing alternative expressions into account the way native speakers do. *Language Learning, 66,* 60–93. https://doi.org/10.1111/lang.12149

Tachihara, K., & Goldberg, A. (2020). Reduced Competition Effects and Noisier Representations in a Second Language. *Language Learning, 70*(1), 219–265. https://doi.org/10.1111/lang.12375

Effects of input type frequency on structural priming and statistical preemption in the acquisition of L2 dative construction

Chi Zhang (Xi’an Jiaotong University; Ghent University), Min Wang (Xi’an Jiaotong University)
Chi.zhang.jarvis@outlook.com

Second language (L2) learners usually have difficulty in learning dative constructions, particularly in abstracting double object (DO) structure (e.g*., John gave Mary a necklace*) and prepositional dative (PD) structure (e.g., *John gave a necklace to Mary*). L2 learning of dative constructions requires learners to extend the acceptability of less preferred structure but avoid excessively accepting certain combinations between verbs and structures that are ungrammatical (Oh, 2010). However, it was still unclear how L2 learning results in overgeneralization in the first place. Overgeneralization might be caused by structural priming, that is the persistence of syntactic structures between language input and production/comprehension (Ivanova, Pickering, McLean, Costa, & Branigan, 2012). For example, after reading a DO sentence, speakers tend to produce an erroneous sentence where a non-generalizable PD verb combines with a DO structure (i.e., overgeneralized DO, e.g., **John drove Mary a car*) instead of using a grammatical PD structure (i.e., Non-generalizable PD, e.g., *John drove a car to Mary*). Second, it was suggested that L1 learners constrain the overgeneralized form via statistical preemption (Boyd & Goldberg, 2011), whereby they take the repeated input of structure X (e.g., non-generalizable PD) as indirect negative evidence of the appropriateness of a semantically related structure Y (DO) in the same context. However, it is still debatable whether statistical preemption affects L2 learners’ language generalization. Third, the effect of certain statistical feature, especially the type frequency of input, on L2 learning has not yet been elucidated.

To test the above questions, the present study investigated the effects of structural priming and statistical preemption on L2 learning of DO structure, and how the type frequency of the input modulates these effects. Two pretest-exposure-posttest experiments were conducted (see Figure 1). In both experiments, the experimental group received input of English dative sentences during the exposure session (DO in Experiment 1, DO and non-generalizable PD in Experiment 2). The type frequency of the DO input was manipulated between groups in Experiment 1 (HDZP vs. LDZP) and that of the non-generalizable PD input was manipulated between groups in Experiment 2 while the input type frequency of DO was kept high (HDHP vs. HDLP). Additionally, there was a control group in which subjects received no language input during the exposure session. The production of dative structures for each group was assessed before, immediately after, and two days after exposure. The findings were three-fold: first, there were both short-term and long-term effects of structural priming on well-formed and overgeneralized production (i.e. the likelihood of well-formed and overgeneralized DO production increased in the posttests, see Figure 2) and statistical preemption (i.e., the likelihood of overgeneralized DO production was lower in the posttests of HDHP vs. HDZP); second, in terms of the DO overgeneralization, the short-term effect of type frequency was found on structural priming (i.e., more overgeneralized DO responses in the immediate posttest of HDZP vs. LDZP) and statistical preemption (i.e., less overgeneralized DO responses in the immediate posttest of HDHP vs. HDLP), while the long-term effect of type frequency was found on preemption; third, in terms of the well-formed DO production, a significant short-term and a marginal long-term modulating effect of type frequency on preemption was found, while type frequency also showed marginal short-term and long-term effects on priming.

In sum, in two experiments we showed that L2 learners’ generalization, overgeneralization, and avoidance of overgeneralization in dative learning can be driven by structural priming and statistical preemption. Both processes are sensitive to input type frequency. These findings provide evidence in support that statistical-driven processes can facilitate L2 learners to recover from conservativity and overgeneralization.

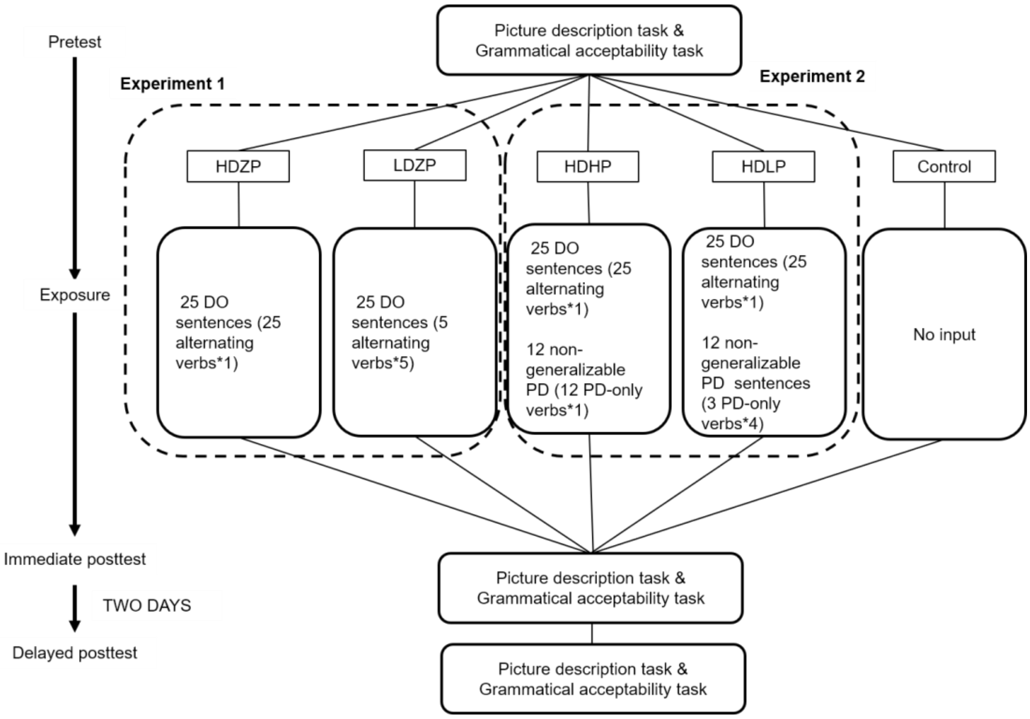

Pretest
Picture description task & Grammatical acceptability task
Experiment 1
Experiment 2
HDZP
LDZP
HDHP
HDLP
Control
Exposure
25 DO sentences (25 alternating verbs*1)
25 DO sentences (5 alternating verbs*5)
25 DO sentences (25 alternating verbs*1) 12 non-generalizable PD (12 PD-only verbs*1)
25 DO sentences (25 alternating verbs*1) 12 non-generalizable PD sentences (3 PD-only verbs*4)
No input
Immediate posttest
TWO DAYS
Delayed posttest
Picture description task & Grammatical acceptability task
Picture description task & Grammatical acceptability task

Figure 1. Experiment design in Experiment 1 and Experiment 2. HDZP = High type frequency DO, Zero PD, LDZP = Low type frequency DO, Zero PD, HDHP = High type frequency DO, Low type frequency PD, HDLP = High type frequency DO, Low type frequency PD.

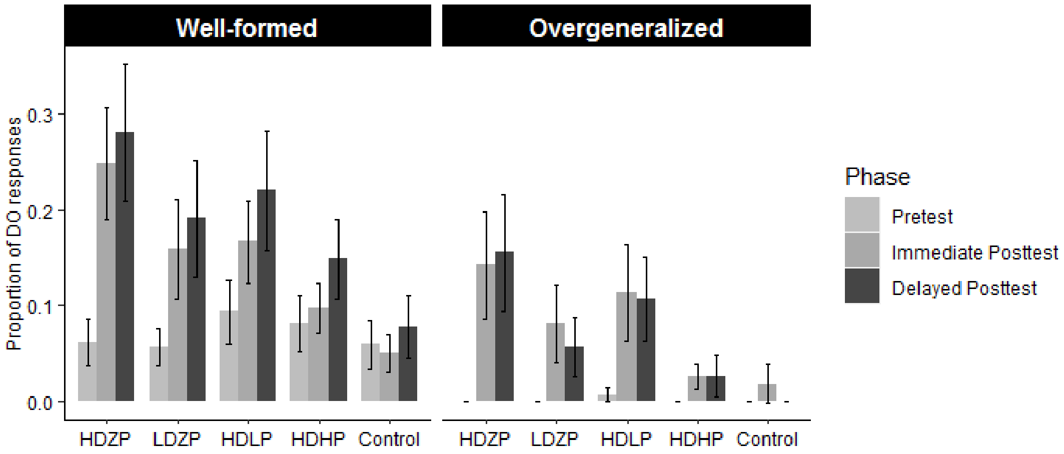


Figure 2. Proportion of DO responses collapsed by group, phase, and production type. HDZP = High type frequency DO, Zero PD, LDZP = Low type frequency DO, Zero PD, HDHP = High type frequency DO, Low type frequency PD, HDLP = High type frequency DO, Low type frequency PD.

References

Boyd, J. K., & Goldberg, A. E. (2011). Learning what not to say: The role of statistical preemption and categorization in a-adjective production. *Language, 87*(1), 55-83.

Ivanova, I., Pickering, M. J., McLean, J. F., Costa, A., & Branigan, H. P. (2012). How do people produce ungrammatical utterances? *Journal of Memory and Language, 67*(3), 355-370.

Oh, E. (2010). Recovery from first-language transfer: The second language acquisition of English double objects by Korean speakers. *Second Language Research, 26*(3), 407-439.

Modeling Maltese broken and sound plurals with Naive Discriminative Learning

Jessica Nieder[a], Fabian Tomaschek[b] & Ruben van de Vijver[a]

nieder@phil.hhu.de, fabian.tomaschek@uni-tuebingen.de, Ruben.Vijver@hhu.de

[a]Department of General Linguistics, Heinrich-Heine-Universität Düsseldorf, Germany;
[b]Department of General Linguistics, Eberhard-Karls-Universität Tübingen, Germany

According to a word-based approach of morphology, such as the Word and Paradigm model, the word and its inflectional paradigm are the central units of contrast and are therefore used for generalizing to new word forms (Blevins, 2016). With the whole word as the basic unit of morphology, the Word and Paradigm approach avoids problems that are related to the notion of the morpheme. In the present study we test a word-based approach by using NDL to computationally model the Maltese plural formation without morphemes.

Maltese, a Semitic language spoken in the island country of Malta, is a language that shows a rich variety of inflected forms. Its complex noun plural system is split between a great number of concatenative (*sound plurals*, sg. *nazzjon* - pl. *nazzjonijiet* 'nation') and non-concatenative (*broken plurals*, sg. *kelb* - pl. *klieb* 'dog') plural forms.

To answer the question as to what information is necessary to classify Maltese nouns we used a data set of 3190 singular-plural pairs (2406 sound, 784 broken) and manually transcribed them such that every phone is represented as exactly one letter or symbol. We trained NDL on 90% of the data set to predict Maltese plural classes on the basis of the phonological forms. Based on the experimental background presented in Nieder et al. (2020) we distinguished between 8 types of plurals: the three most frequent Maltese sound plurals suffixes (*-ijiet*, *-iet* and *-i*) and one category that contains all other, less frequent, sound plural forms (*sound (rest)*) and the three most frequent broken plural patterns (*CCVVCVC, (C)CVCVC, CCVVC*) and one category that contained all other broken plural forms (*broken (rest)*).

NDL predicted the outcomes, the different plural classes, on the basis of different cues: singular forms coded as 2-phones or 3-phones and singular-plural pairs, coded as sets of 2-phones or 3-phones. Table 1 below shows the results of the best performing model:

	CCVVCVC	(C)CVCVC	CCVVC	broken	sound iet	sound ijiet	sound i	sound
CCVVCVC	**6** (27%)	4 (18%)	7 (32%)	1 (5%)	2 (9%)	0 (0%)	0 (0%)	2 (9%)
(C)CVCVC	1(5%)	**14** (70%)	0 (0%)	4 (20%)	0 (0%)	0 (0%)	1 (5%)	0 (0%)
CCVVC	1 (8%)	3 (23%)	**7** (54%)	2 (15%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
broken (rest)	1(11%)	0 (0%)	0 (0%)	**6** (67%)	0 (0%)	0 (0%)	0 (0%)	2 (22%)
sound iet	1 (3%)	1 (3%)	4 (10%)	6 (15%)	**26** (67%)	1 (3%)	0 (0%)	0 (0%)
sound ijiet	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1(2%)	**44** (96%)	1 (2%)	0 (0%)
sound i	0 (0%)	1 (1%)	2 (1%)	4 (3%)	0 (0%)	1 (1%)	**127** (94%)	0 (0%)
sound (rest)	2 (6%)	0 (0%)	5(15%)	3 (9%)	1 (3%)	0 (0%)	6(18%)	**17** (50%)

Table 1: Confusion matrix of the NDL model with the singular-plural paradigm coded as 2-phone cues. Rows represent the input category, columns represent their classification. Overall accuracy: 78%.

NDL shows the best predictions for Maltese plurals if the paradigm is coded as 2-phones with errors mainly caused by confusing the different broken plural types or sound plural suffixes with each other in their respective plural group. Both versions using singulars as cues failed to correctly predict broken plurals. This shows that information about the plural is necessary and needs to be stored in order to predict the plural form of a novel word.

Our results are then in line with a word-based approach of morphology. Information about whole words and their paradigms is needed to build Maltese sound and broken plurals.

References

Blevins, J. (2016). *Word and Paradigm Morphology*. Oxford University Press.

Nieder, J., van de Vijver, R., & Mitterer, H. (2020). Knowledge of Maltese singular-plural mappings. Analogy explains it best. *Morphology*. https://doi.org/https://doi.org/10.1007/s11525-020-09353-7

Linear Discriminative Learning in Julia

Xuefeng Luo, Yu-Ying Chuang, and Harald Baayen

Linear Discriminative Learning (LDL)[1] is a computational framework that learns and makes predictions about word meanings and forms. For comprehension, the model learns to predict word meanings from word forms. For production, it predicts word forms from word meanings. The networks' weights are estimated using matrix algebra, the underlying mathematics of which is the same as that of multivariate multiple regression. Given the matrices of word form and meaning representations $\boldsymbol{C}$ and $\boldsymbol{S}$, the mappings $\boldsymbol{F}$ and $\boldsymbol{G}$, equivalent to the comprehension and production networks, are obtained by solving the equations $\boldsymbol{CF} = \boldsymbol{S}$ and $\boldsymbol{SG} = \boldsymbol{C}$, respectively. On the comprehension side, the model predicts word meanings $\hat{s}$ by multiplying the form vector $\boldsymbol{c}$ with $\boldsymbol{F}$. It then evaluates the predicted semantic vectors by taking the gold label semantic vector with which it is most strongly correlated. For production, the model first predicts the form vector $\hat{c}$ by multiplying $\boldsymbol{s}$ with $\boldsymbol{G}$. But then it proceeds to assemble n-gram cues in $\hat{c}$ in the proper order. Model accuracy is evaluated by comparing the predicted word forms to the corresponding gold standard forms.

`LDL` was initially implemented in `R`. Despite high accuracy, the `R` implementation, `WpmWithLdl`, is extremely time- and computational power-consuming, making it difficult to run the model on larger datasets. `Julia` is another programming language that optimizes numeric operations. We made a completely new implementation of the `LDL` model in Julia, named `JudiLing`, which reduces computation time and memory dramatically, in part because we implemented Cholesky decomposition when calculating the inverse of matrices. Speed of calculation and accuracy are also considerably improved for the two new sequencing algorithms for production, compared to the original algorithm for phone sequencing used by `WpmWithLdl`. In addition, to speed up matrix operation, we made use of a specific matrix format in Julia to deal with sparse matrices.

Benchmarking studies show that `JudiLing` not only reduces computation time and makes it possible to process larger datasets, but also offers increased cross-validation accuracies. For example, for a dataset containing 6,440 inflectional forms of Estonian nouns, the cross-validation takes more than 1 day in `WpmWithLdl`, but only 48.2 seconds in `JudiLing`, while the accuracy increases from 0.695 to 0.899. For another dataset of 21,360 short French utterances with auxiliary and verb combinations, we found that `WpmWithLdl` cannot process it because it takes too much memory. However, `JudiLing` only takes 20.6 hours of training and evaluation on cross-validation and achieves 0.67 accuracy.

Other features like multi-threading computing and incremental learning using the Widrow-Hoff learning rule[2] are also implemented in `JudiLing`. The next step we are now considering, is to design several high-level wrapper functions to simplify the modeling steps for the ease of the user. The faster and energy-saving `Julia` implementation of LDL `JudiLing` has now made it possible to study a wider range of morphological systems and larger

[1] R. H. Baayen et al. "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 (2019), pp. 1–39. DOI: `10.1155/2019/4895891`.

[2] Bernard Widrow and Marcian E. Hoff. "Adaptive switching circuits". In: *1960 WESCON Convention Record Part IV* (1960), pp. 96–104. DOI: `10.21236/ad0241531`.

datasets, while at the same time making the model more environmentally friendly.

Triphone meanings co-determine tongue shape during articulation: An ultrasound study

Motoki Saito, Fabian Tomaschek, R. Harald Baayen
Eberhard Karls Universität Tübingen

motoki.saito@, fabian.tomaschek@, harald.baayen@uni-tuebingen.de

Discriminative learning (Rescorla, 1988; Rescorla & Wagner, 1972) has been successfully employed to model a wide range of experimental data (e.g. Baayen, Milin, & Ramscar, 2016; Tomaschek, Plag, Ernestus, & Baayen, 2019). However, most of these studies have focused on the comprehension side of language processing (e.g. Baayen, Chuang, Shafaei-Bajestan, & Blevins, 2019; Shafaei-Bajestan & Baayen, 2018) or on the speech signal produced (Tomaschek et al., 2019). What is not well understood at present is to what extent words' meanings help shape the way in which words are articulated.

Saito (2020a) begins to address this question by investigating word-final triphones that contained the stem vowel [a(ː)] and the word-final [t]. These authors observed lower tongue tip and higher tongue body positions for word-final triphones that were more strongly implicated in the mappings between form and meaning. Triphones' degree of semantic support was estimated using a Linear Discriminative Learning (LDL) (Baayen et al., 2019) model; the degree of semantic support can be understood as an operationalization of the classical concept of *functional load* (Saito, 2020a). The tongue movement data in their study was recorded with electromagnetic articulography (EMA), where only a few sensors on the tongue can be tracked. To consolidate this effect of functional load on articulation, the present study made use of ultrasound as experimental method, instead of EMA. Ultrasound offers the possibility to study the movement of much larger parts of the tongue, especially when analyzed with GAMMs (Saito, 2020b).

The ultrasound data in the present study consist of 20 participants articulating 126 German inflected verbs with the stem vowel [aː]. These verbs were combined with two types of pronouns ([ziː] vs [(v)ːr]) and the two types of suffixes ([t] vs [n]). The verbs were selected subject to the criterion that at most one segment intervened between the stem vowel and the suffix (e.g. *malt* [maːlt]).

Functional load, the semantic contribution of sublexical units (triphones) to the target meaning, was first pitted against word frequency and three commonly used measures from Naive Discriminative Learning (NDL) (Tomaschek et al., 2019), i.e. activation, prior, and activation diversity, in order to replicate the previous finding that functional load outperforms frequency (Saito, 2020a) and the simple NDL measures (Baayen et al., 2019). To this end, a Random Forest model was fitted with brightness values in the ultrasound image as response variable. The functional load of the word-final triphone emerged with the highest variable importance.

A Generalized Additive Mixed-effects Model (GAMM) (Wood, 2006) was then fitted to model ultrasound images with x- and y-coordinate values and functional load as predictors. The ultrasound images fitted with the GAM model are shown in Figure 1 for a high (quantile=0.9) and a low (quantile=0.1) value of the functional load of the word-final triphone. As expected and consistent with Saito (2020a), the tongue body is positioned higher for the high functional load (in the leftmost plot), compared to a low functional load (second plot). Conversely, the tongue tip is higher in the second plot than in the first plot. The differences between these two plots are presented in the third plot. The rightmost panel highlights where differences are significant.

The present finding indicates that a greater functional load of the word-final triphone induces a more bulged shape of the tongue. This result provides further support for the hypothesis that semantics influences fine details of phonetic realizations (Saito, 2020a) and challenges classical views of speech production such as WEAVER++ (e.g. Levelt, Roelofs, & Meyer, 1999; Levelt & Wheeldon, 1994).

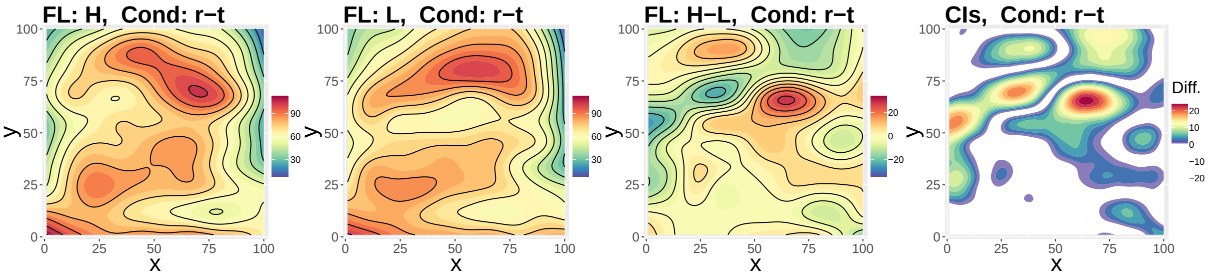


Figure 1: Fitted ultrasound images with GAM. The mouth front is to the right of each image. Warmer colors represent higher values and colder colors represent lower values. The first (leftmost) and second plots are when the functional load is high and low for each. The third plot shows how the two surfaces differ. The rightmost visualizes where the two surfaces differ significantly. Warmer colors indicate larger differences between confidence regions.

References

Baayen, R. H., Chuang, Y.-Y., Shafaei-Bajestan, E., & Blevins, J. P. (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. *Complexity*, 1–39. doi: 10.1155/2019/4895891

Baayen, R. H., Milin, P., & Ramscar, M. (2016). Frequency in lexical processing. *Aphasiology*, *30*(11), 1174–1220. doi: 10.1080/02687038.2016.1147767

Levelt, W. J. M., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences*, *22*, 1–75.

Levelt, W. J. M., & Wheeldon, L. (1994). Do speakers have access to a mental syllabary? *Cognition*, *50*, 239–269.

Rescorla, R. A. (1988). Pavlovian Conditioning: It's Not What You Think It Is. *American Psychologist*, *43*(3), 151–160.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning ii: Current research and theory* (pp. 64–99). New York: Appleton-Century-Crofts.

Saito, M. (2020a). Relative functional load determines co-articulatory movements of the tongue tip. In *Proceedings of issp 2020 - 12th international seminar on speech production.*

Saito, M. (2020b). An ultrasound study of frequency and co-articulation. In *Proceedings of issp 2020 - 12th international seminar on speech production.*

Shafaei-Bajestan, E., & Baayen, R. H. (2018). Wide learning for auditory comprehension. In *Proceedings of the annual conference of the international speech communication association, interspeech* (pp. 966–970). doi: 10.21437/Interspeech.2018-2420

Tomaschek, F., Plag, I., Ernestus, M., & Baayen, R. H. (2019). Phonetic effects of morphology and context: Modeling the duration of word-final S in English with naïve discriminative learning. *Journal of Linguistics*, 1–39. doi: 10.1017/S0022226719000203

Wood, S. N. (2006). *Generalized Additive Models: An Introduction with R*. Boca Raton, Florida, U.S.A.: CRC Press.