In Search of Abstract Morphological Structure

Micha Elsner, Ohio State University

2021

Morphology: grammatical information within the word

English: *cat* (SG) ~ *cats* (PL)

Morphology: grammatical information within the word

English: *cat* (SG) ~ *cats* (PL)

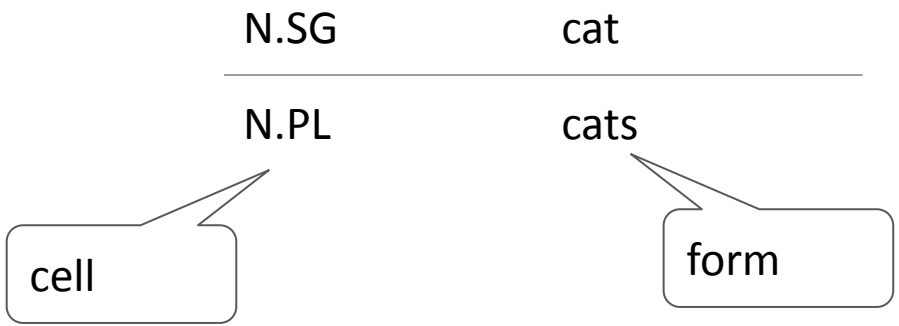
But also:

Latin:
laudāverītis
“You all would have praised”

Murrinhpatha (Forshaw et al 2016):
bamngingkardungintha
“The two non-siblings (either ♀ ♀ or ♀ ♂) saw me”

A **lexeme** (“a word”) has a number of related **forms**, which make up its **paradigm**.

The set of properties that a form realizes is its **cell**.



Of course, some of these tables are bigger than others...

What goes in the cells?

Language users constantly face data sparsity when trying to remember what goes where (Bybee 2003, Ackerman and Malouf 2013)

Challenges from:

- Unfamiliar lexemes
- Rare cells
- Less-dominant language with lower exposure

Both lexemes and cells are Zipfian

(Lignos and Yang 2016)

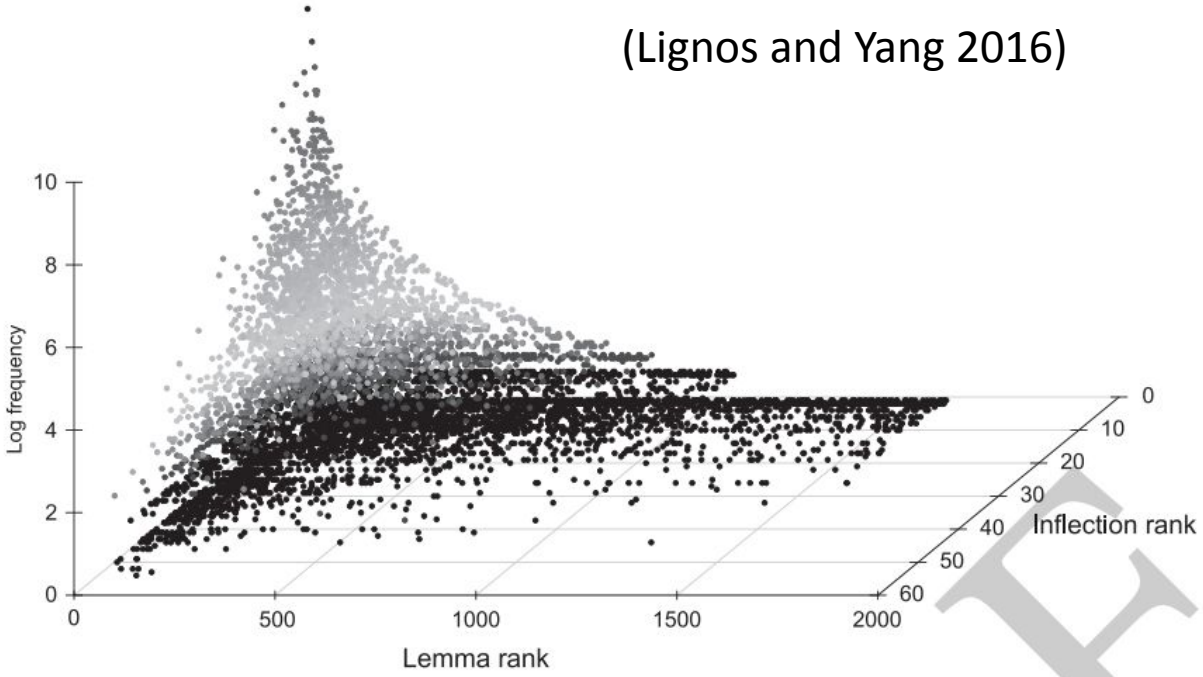


Figure 27.1. Frequencies of CHILDES Spanish lemmas across inflection categories

Luckily, analogy can fill in some of the gaps

	Across words	
N.SG	cat	mink
N.PL	cats	**?**

Luckily, analogy can fill in some of the gaps

Across words

N.SG	cat	mink
N.PL	cats	**?**

Across cells

“You praise”	laudās
“You all praise”	laudātis
“You would praise”	laudēs
“You all would praise”	laudētis
“You would have praised”	laudāverīs
“You all would have praised”	**?**

Abstraction

These analogies copy bits of structure (the *-s* and *-ti-*) within the word.

But not all analogies work this way.

Latin	PL “woman”	PL “fire”
NOM	fēmin-ae	ign-ēs
GEN	fēmin-ārum	ign-ium
DAT	fēmin-īs	ign-ibus
ACC	fēmin-ās	ign-ēs
ABL	fēmin-īs	ign-ibus

Generalizations about paradigms (Wurzel 1989)

Which cells are the same/different

Which cells are easy/hard to predict

Where and how the morphological marking appears in the word

	English	
N.SG	cat	ox
N.PL	cat**s**	ox**en**

	Irish	
	“flat land”	“gap”
N.SG	mín	gáibéal
N.PL	mín**te**	gáib**éi**l

Paradigm discovery

Alex Erdmann (Ph.D. 2020)

The paradigm discovery problem: Erdmann, Elsner, Wu, Cotterell and Habash, ACL 2020

Modeling morphological learning, typology, and change: What can the neural sequence-to-sequence framework contribute?: Elsner, Sims, Erdmann, et al, JLM 2019

The problem: paradigms from raw text

The cat **watched** me **watching** it .
I **followed** the show but she hadn’t **seen** it .
Let’s **see** who **follows** your logic .

Which types fill the same cells?

Which ones are forms of the same lexeme?

What fills in the unattested cells?

The problem: paradigms from raw text

The cat **watched** me **watching** it .
I **followed** the show but she hadn’t **seen** it .
Let’s **see** who **follows** your logic .

Which types fill the same cells?

Which ones are forms of the same lexeme?

What fills in the unattested cells?

C1	see
C2	follows
C3	watching
C4	watched, followed
C5	seen

The problem: paradigms from raw text

The cat **watched** me **watching** it .
I **followed** the show but she hadn't **seen** it .
Let's **see** who **follows** your logic .

Which types fill the same cells?

Which ones are forms of the same lexeme?

What fills in the unattested cells?

"watch"	"follow"	"see"
?	?	see
?	follows	?
watching	?	?
watched	followed	?
?	?	seen

The problem: paradigms from raw text

The cat **watched** me **watching** it .
I **followed** the show but she hadn't **seen** it .
Let's **see** who **follows** your logic .

Which types fill the same cells?

Which ones are forms of the same lexeme?

What fills in the unattested cells?

"watch"	"follow"	"see"
watch	**follow**	see
watches	follows	**sees**
watching	**following**	**seeing**
watched	followed	**saw**
watched	**followed**	seen

Step 1: finding cells

Represent the word types with FastText (Bojanowski et al 2017), a Word2Vec-like method that incorporates substring information.

Ideally, things with similar distribution and string markers will be embedded close together.

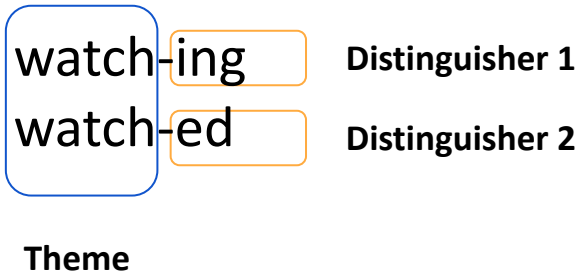
Use *k*-means to extract hard clusters.

Results: clustering into cells is challenging

	F-score of paradigm cell mates
Arabic nouns	39.9
German nouns	35.2
English verbs	64.0
Latin nouns	38.8
Russian nouns	44.5

Step 2: finding lexemes

Basic idea: forms of the same lexeme *usually* consist of some shared content (the **theme**) and some cell-specific marking (**distinguishers**)

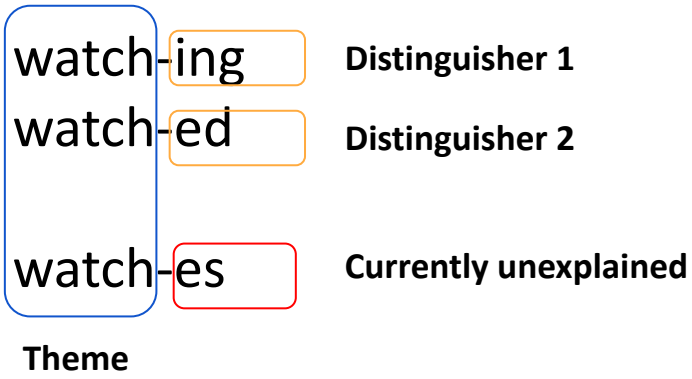


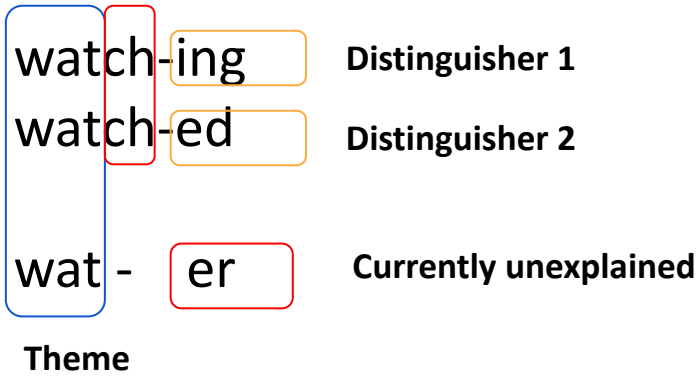
Building lexemes incrementally

Consider adding a new form to the lexeme…

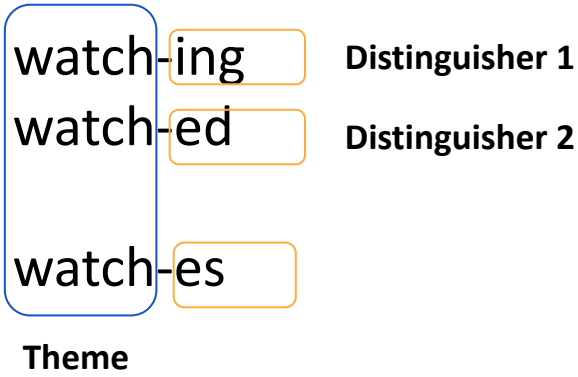
Adding “watches” adds 5 theme characters and 2 unexplained characters.

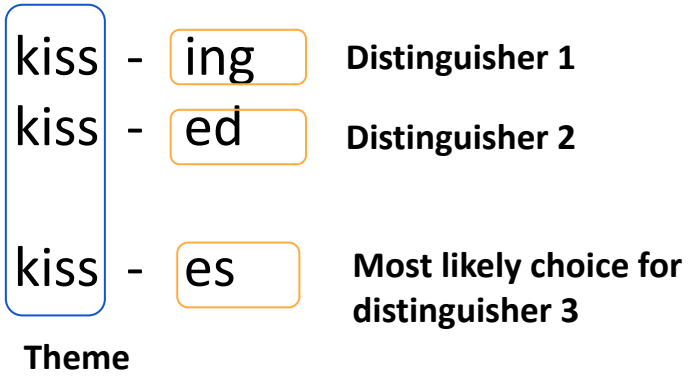
Adding “water” adds 3 theme characters and 6 unexplained characters.





After each pass, re-estimate the distinguishers





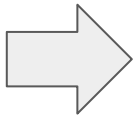
Results: clustering into lexemes is not too bad

	F-score of lexeme mates
Arabic nouns	48.5
German nouns	59.4
English verbs	80.1
Latin nouns	73.2
Russian nouns	72.2

Filling the gaps

Use a small Transformer network (Wu and Cotterell 2020) to fill the empty slots

“watch”	“follow”	“see”
?	?	see
?	follows	?
watching	?	?
watched	followed	?
?	?	seen



“watch”	“follow”	“see”
watch	**follow**	see
watches	follows	**sees**
watching	**following**	**seeing**
watched	followed	**saw**
watched	**followed**	seen

Results: filling empty slots is also difficult

	% missing forms predicted (anywhere)	Supervised
Arabic nouns	49.5	87.0
German nouns	56.7	74.9
English verbs	67.5	80.7
Latin nouns	72.9	88.0
Russian nouns	86.2	96.8

Why is this so tough?

Things that look the same don't always belong together:

	SG				PL					
NOM	GEN	DAT	ACC	ABL	NOM	GEN	DAT	ACC	ABL	**Gloss**
serv-us	*i*	*o*	*um*	*o*	*i*	*orum*	*is*	*os*	*is*	"slave.M"
serv-a	*ae*	*ae*	*am*	*a*	*ae*	*arum*	*is*	*as*	*is*	"slave.F"
frat-er	*ris*	*ri*	*rem*	*re*	*res*	*rum*	*ribus*	*res*	*ribus*	"brother"

This can cause cascading errors. If we decide *servum* and *servorum* belong in the same cell, we must assign them to different lexemes.

If we put *fratrum* and *servum* in the same cell, we hide the real correspondence *fratrum* ~ *servorum*.

What’s missing

More sensitive use of context (but this isn’t a cure-all, especially in languages with relatively free word order)

Better use of paradigm structure; use internal consistency of the tables to avoid being misled by surface overlap.

Better integration of the transformer: can feedback from the neural learner help to exclude implausible guesses?

Work is still ongoing

A SigMorphon 2021 shared task (Wiemerslage et al) investigates a similar problem.

(Bayesian) clustering-based systems are still state-of-the-art (McCurdy et al)

And scores are still comparatively low...

Looking for paradigm structure

Grace LeFevre
(B.A. 2021, now Ph.D. program at Northwestern)

Formalizing Inflectional Paradigm Shape with Information Theory: Lefevre, Elsner and Sims, SCiL 2021

“Paradigm shape” in Spanish

LEXEME	GLOSS	PRS.1SG	PRS.2SG	PRS.3SG	PRS.1PL	PRS.2PL	PRS.3PL
CANTAR	‘sing’	canto	cantas	canta	cantamos	cantáis	cantan
SUBIR	‘rise’	subo	subes	sube	subimos	subís	suben
SENTIR	‘feel’	siento	sientes	siente	sentimos	sentís	sienten
PENSAR	‘think’	pienso	piensas	piensa	pensamos	pensáis	piensan
MOVER	‘move’	muevo	mueves	mueve	movemos	movéis	mueven

Traditional analysis of the Spanish verbs, based on the vowels in suffixes.

“Paradigm shape” in Spanish

LEXEME	GLOSS	PRS.1SG	PRS.2SG	PRS.3SG	PRS.1PL	PRS.2PL	PRS.3PL
CANTAR	‘sing’	canto	cantas	canta	cantamos	cantáis	cantan
SUBIR	‘rise’	subo	subes	sube	subimos	subís	suben
SENTIR	‘feel’	siento	sientes	siente	sentimos	sentís	sienten
PENSAR	‘think’	pienso	piensas	piensa	pensamos	pensáis	piensan
MOVER	‘move’	muevo	mueves	mueve	movemos	movéis	mueven

But it’s also well-known that some verbs have stem alternations.
*This distribution is what Maiden (2005) terms the N-pattern

Measure similarity in a flexible way

What is the underlying structure of the verb classes?

Cantar and *pensar* share their stem vowel, but not their alternation pattern

Do the verbs form real “clusters”, and if so, which organizing principles matter?

“Confusion sets”

PRS.1SG	PRS.2SG
canto	cantas
subo	subes
siento	sientes
pienso	piensas
muevo	mueves

theme	distinguishers
cant	o, as
sub	o, es
sient	o, es
piens	o, as
muev	o, es

Alignment-based segmentation, following Beniamine et al. (2017)

Segmenting Stem Alternations

PRS.1SG	PRS.1PL
subo	subimos
muevo	movemos

theme	distinguishers
subo	--, ims
mvo	ue, oems

Method sketch

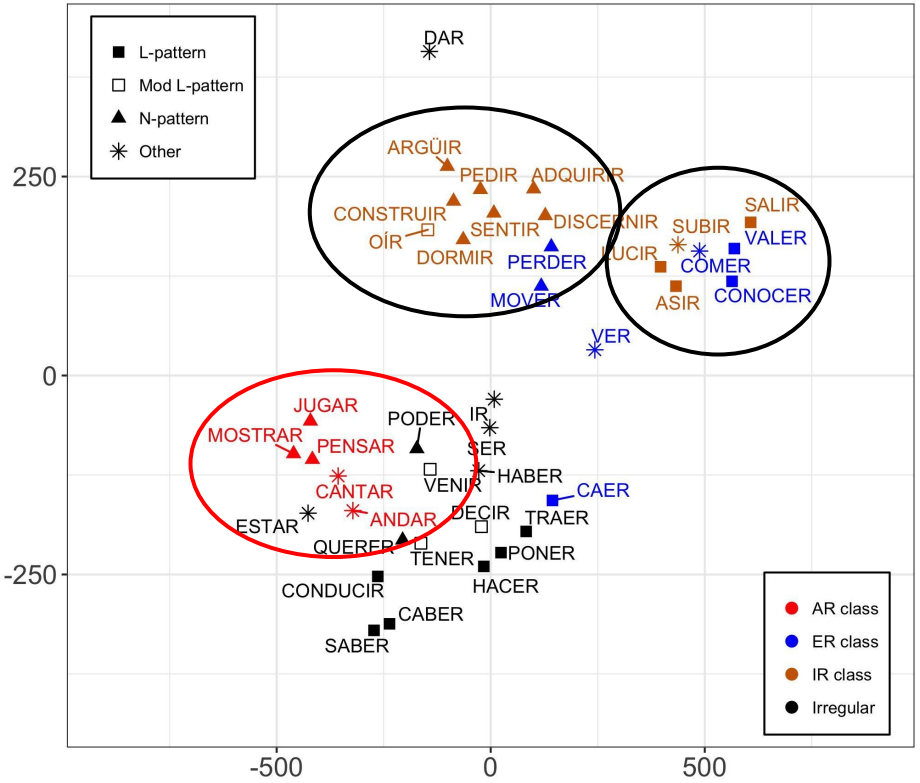
Generate matrix of entropy values quantifying relations in the inflectional system

- Segment locally for subsets of cells
- Divide lexemes into “confusion sets” with identical distinguishers
- Calculate entropy values: larger confusion sets get higher values

Use these entropies as coordinates in an embedding space

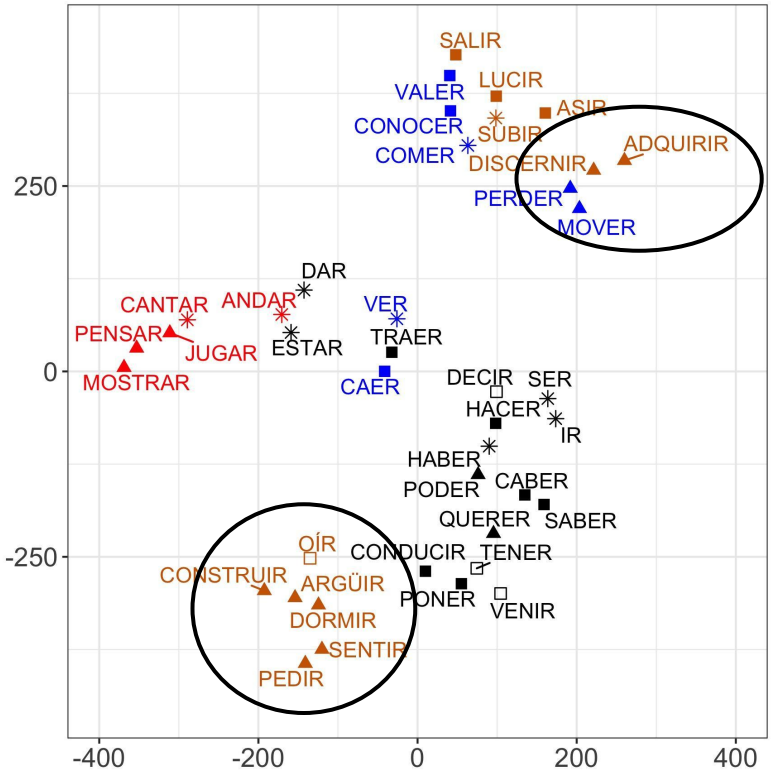
Verbs are **similar** for a set of columns if their forms are **similarly easy or hard to predict**

t-SNE Visualization



- Allomorphic groupings based on inflectional exponents
- Distributional groupings based on Maiden's (2005) stem alternations

Follow-up “deidentified” analysis



Replacing individual characters with abstract identifiers focuses more on which cells contain alternations

Circled clusters have an additional vowel change in the preterite

Lessons

Spanish verbs have a cross-cutting organization (Baerman 2010); they are clustered differently according to suffix and stem alternation

This complicates the clustering landscape of verb types

But there is still a lot of information there for the learner

Generalizing across languages

Andrea D. Sims (Associate Professor, OSU)

What transfers in morphological inflection? Experiments with analogical models: Elsner, SigMorphon 2021

Analogical modeling of morphology for L1 effects in language contact: Elsner and Sims, AIMM5 (presentation only) 2021

Language transfer

Lots of previous work (McCarthy et al 2019, Kann 2020 …) on using knowledge from one language to learn another

Important in low-resource settings; also forces model to focus on **abstraction** rather than memorizing specific markers

How to encourage models to do this?

An old idea: analogical learning

Train a model to solve four-part proportional analogy problems like this one (to produce a Maori passive, from waiata “sing”)

`waiata karanga:karangatia waiatatia`

source exemplar : exemplar form prediction target

*also investigated by Liu and Hulden 2020: “cross-table examples”

Issue: choosing exemplars

Good exemplar (matches target)

waiata karanga:karangatia waiatatia

Bad exemplar (inflects differently from target)

waiata kaukau:kaukauria waiatatia

Two basic options

Random exemplars: Easy. Training matches test.

Similar exemplars: Pick training exemplar that uses the same alignment-based edit rule (for example, *+tia*). Training mismatches test.

Always use **random exemplars** at test time.

Setup

Transformer (Wu and Cotterell 2020) as learner again

Staged training:

- Learn to copy
- Multilingual model
- Fine-tuning

SigMorphon 2020 development languages as training set

Supervised results

Random exemplars	Similar exemplars	Non-analogical transformer
89	57	90

Test-train mismatch makes “similar” quite poor

“Random” underperforms where inflection class structure is important

One-shot mode

No fine-tuning; apply trained multilingual model to an unseen language

Provide an exemplar from the target language at test time

One-shot results

	Random exemplars	Similar exemplars	Fine-tuned Baseline
Known family	29	38	80
Novel family	11	28	92
Overall	14	30	90

Similar exemplars performs much better; much more faithful copying from exemplars

One-shot results: Catalan

Lemma	Exemplar	Rand. sys	Sim. sys	Correct form
arrissar	posar : posarien	arrissaren	arrissarien	arrissarien
disputar	descriure : descriuria	disputarta	disputaria	disputaria
repetir	cremar : cremo	repetirer	repetio	repeteixo
engolir	forjar : forjava	engolire	engoliva	engolia
llevar-se	terminar : termino	llevar-se	llevor-se	llevo

Model seems to “know” an abstract suffixation rule

What transfers?

Use synthetic data to probe what the model has learned about specific languages and processes:

`modi  gobu : gogobu  ???`

When example is labeled as Tagalog, output is: `momodi`

But when labeled as Swahili, output is: `gomodi`

Probe tasks

Modeled on real morphological processes with varying representation in the SigMorphon dataset

Process	**What families (in data)?**	**Example**
Prefixing	Niger-Congo, Austronesian	semet ~ **igo**semet
Suffixing	Germanic, Uralic, some others	semet ~ semet**igo**
Reduplication	Austronesian	semet ~ **se**semet
Gemination	None	semet ~ sem**m**et

Results (similar exemplars)

	Prefix	Suffix	Reduplicate	Geminate
Tagalog (Austronesian)	30	75	88	0
German (Germanic)	86	99	0	5
Swahili (Niger-Congo)	99	88	0	0
Mezquital Otomi (Oto-Manguean)	96	59	5	0
Finnish (Uralic)	52	98	0	12

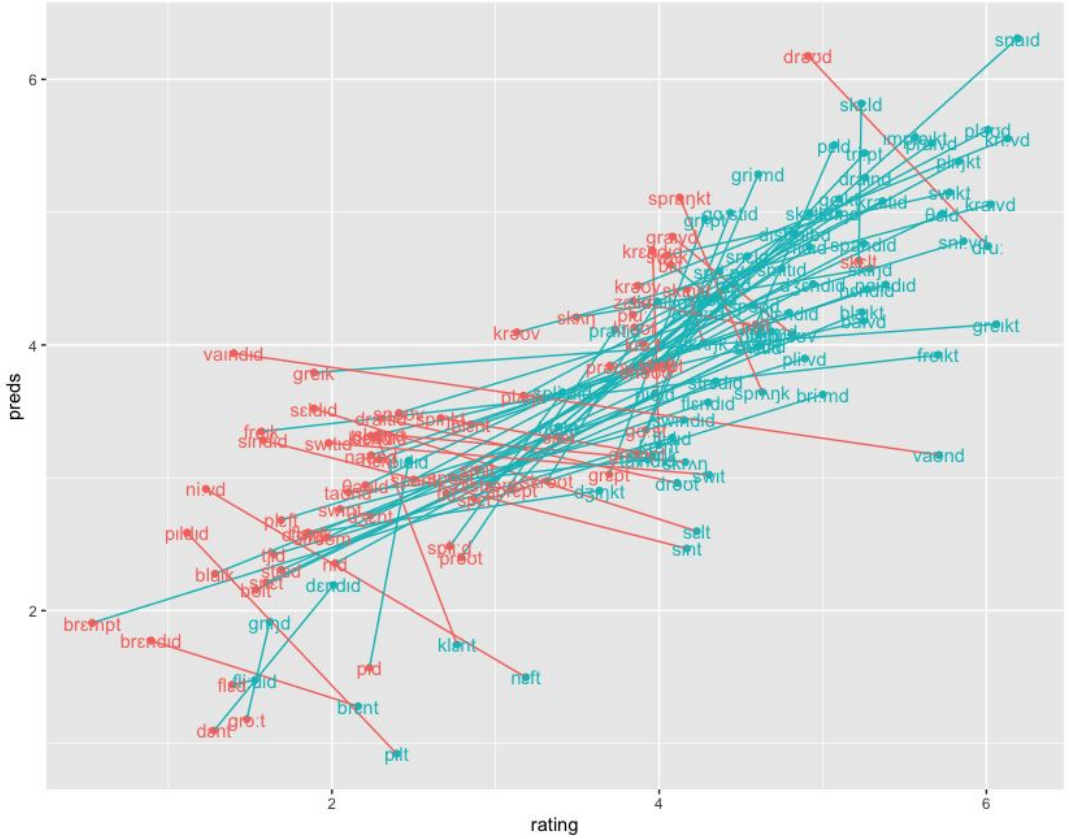
Future work on language transfer

Currently: making model less sensitive to how exemplars are chosen

Teach the model simulated L1, then look at acquisition trajectories in L2

Also model frequency effects; how are exemplars really selected?

Matching human ratings of “wug” words



2021 SigMorphon shared task

Conclusions

Morphology is about more than just what affixes go where…

There are all sorts of abstract structures that help speakers predict:

- Patterns of which cells are the same/different (Latin nouns)
- Commonalities in what is predictable/shared across cells (Spanish verbs)
- Systemic generalizations about how to mark morphological information (Catalan suffixation)
- And more…

Current models learn surface structure well

But we're still learning how to **measure** their knowledge of abstract structures

And how to encourage them to **learn more** and **generalize better**

Thank you!

A closer look: reduplication

	Reduplicated *momodi*	Prefixed *gomodi*	Infixed *mogodi*	Unaltered *modi*	Other *moodi*
Tagalog (Austronesian)	87.5	0	0	0	12.5
German (Germanic)	0	0	7.5	10	82.5
Swahili (Niger-Congo)	0	100	0	0	0
Mezquital Otomi (Oto-Manguean)	5	87.5	0	2.5	5
Finnish (Uralic)	0	0	5	2.5	92.5

A closer look: suffixation

	Suffixed *semet-igo*	All but first *semet-go*	All but last *semet-ig*	Unaltered *semet*	Other *semet-g*
Tagalog (Austronesian)	75	16	~0	3	6
German (Germanic)	100	~0	0	0	0
Swahili (Niger-Congo)	88	5	0	1	6
Mezquital Otomi (Oto-Manguean)	59	5	8	12	16
Finnish (Uralic)	98	2	0	0	0