Learning Morphophonology From Morphology and MDL

John A Goldsmith
The University of Chicago
http://linguistica.uchicago.edu

17 July 2011

1 *Unsupervised learning as a way of doing linguistic theory*

1. Hypothesis generation. *Today's focus*.
2. Hypothesis testing (evaluation).

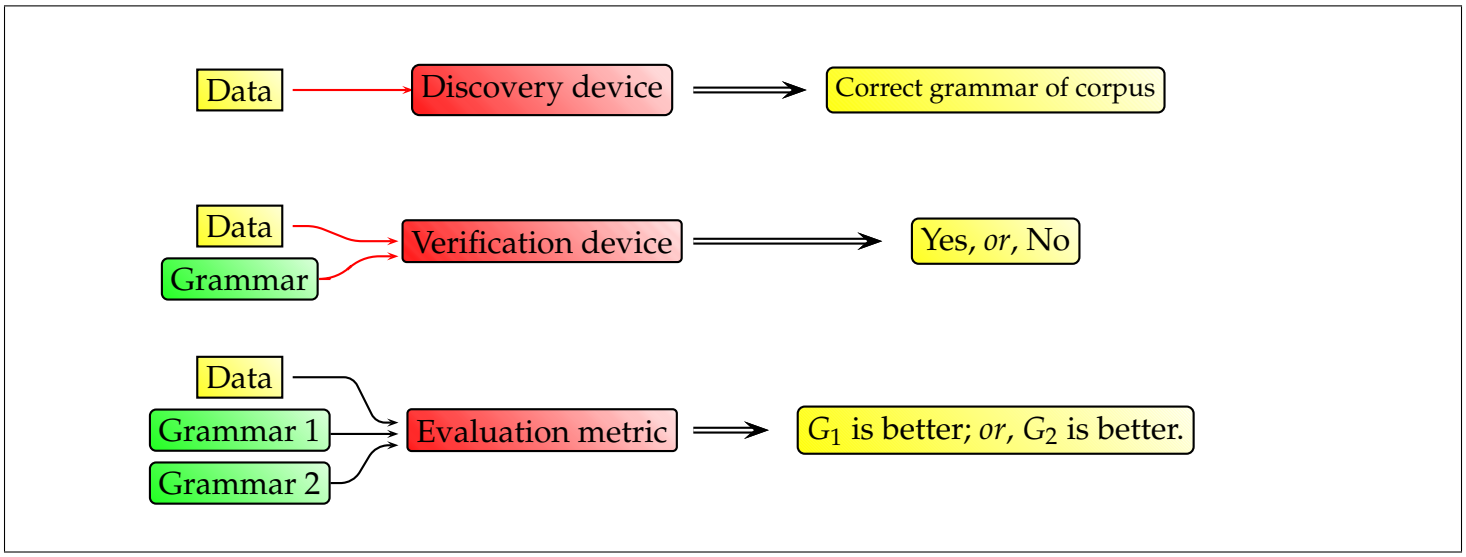


Figure 1: Chomsky's three conceptions of linguistic theory

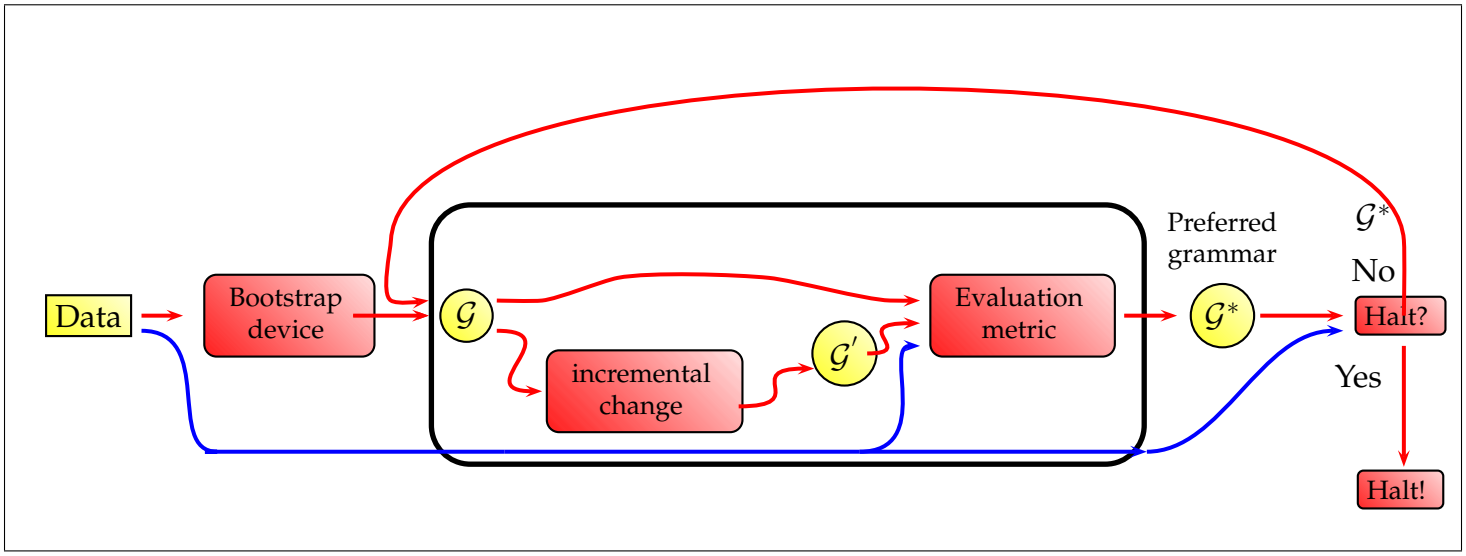


Figure 2: Unsupervised learning of grammars

2 *Unsupervised learning of morphology: the Linguistica project (2001)*

2.1 *Working on the unsupervised learning of natural language morphology. Why?*

What is the task, then? Take in a raw corpus, and produce a morphology. What is a morphology? The answer to that depends on what linguistic problems we want to solve. Let's start with the simplest: analysis of words into morphs (and eventually into morphemes). Solution looks like an FSA, then. Examples: English, French, Swahili. An FSA is a set of vertices (or nodes), a set of edges, and for each edge a label and a probability, where the sum of the probabilities of the edges leaving each node sums to 1.0.

1. **English morphology: morphemes on edges of a finite-state automaton**

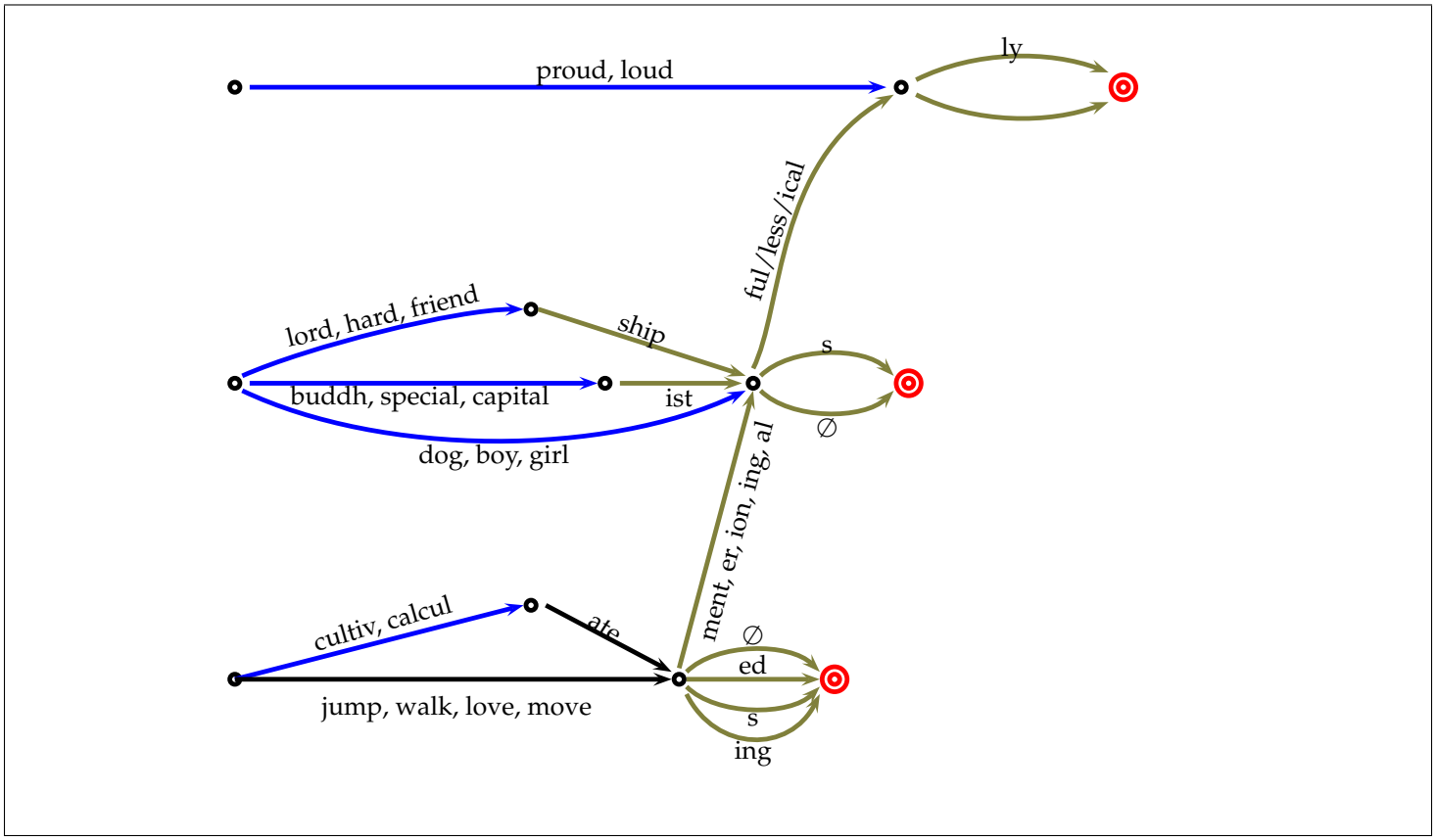


Figure 3: English morphology: morphemes on edges

Pose the problem as an optimization problem: quantitative data that can be measured, but provides qualitatively special points in a continuous world of measurement.

Turning this into a linguistic project

Some details on the MDL model; no time to talk about the search methods.

We can use the term *length* (of something) to mean the *number of bits = amount of information* needed to specify it. Except where indicated, the probability distribution(s) involved are from maximum likelihood models. The *length* of an FSA is the number of bits needed to specify it, and it equals the sum of these things:

1. List of morphemes: assigning the phonological cost of establishing a lean class of morphemes. Avoid redundancy; minimize multiple use identical strings. The probability distribution here is over phonemes (letters).

$$\sum_{t \in morphemes} \sum_{i=1}^{|t|+1} -log\ pr_{phono}(t_i|t_{i-1})$$

2. List of nodes v: the cost of morpheme classes

$$\sum_{v \in Vertices} -log\ pr(v)$$

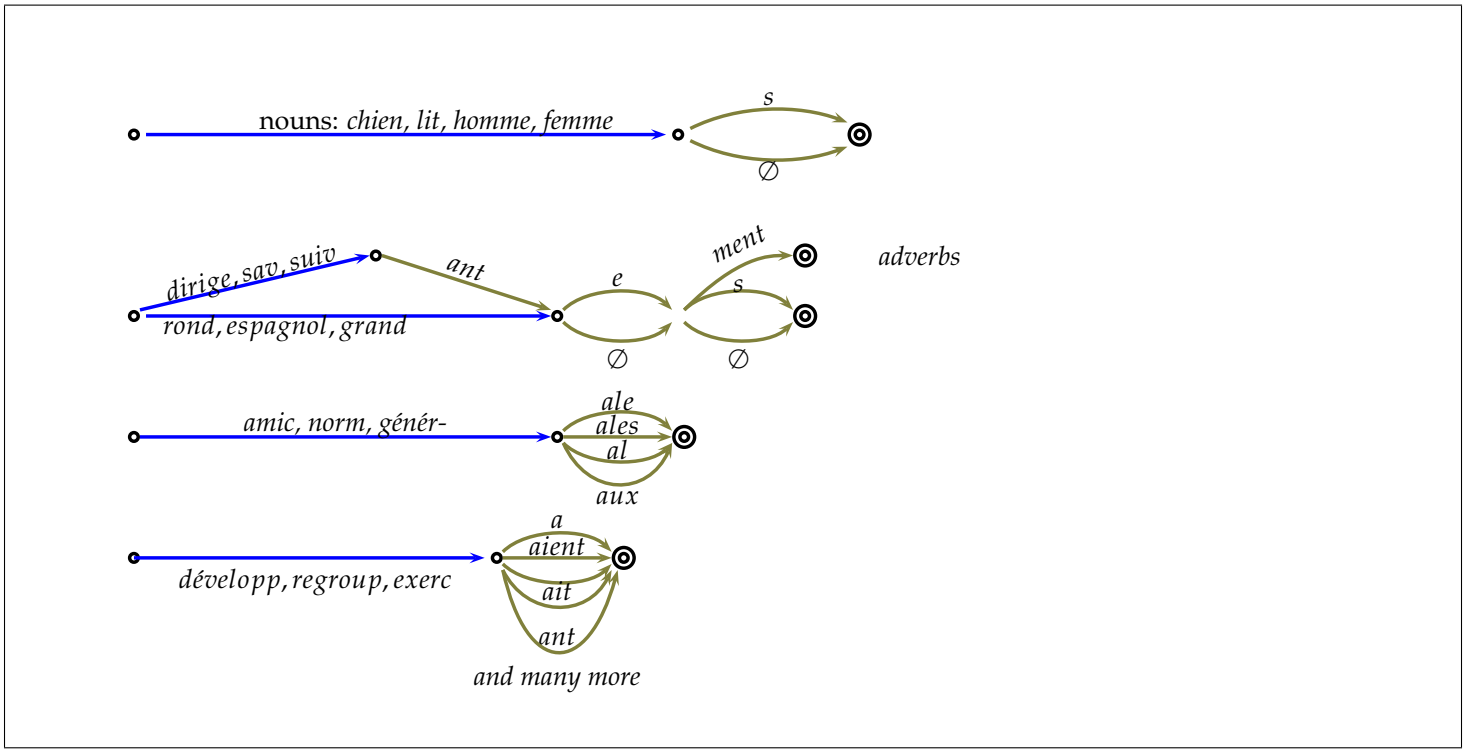


Figure 4: French

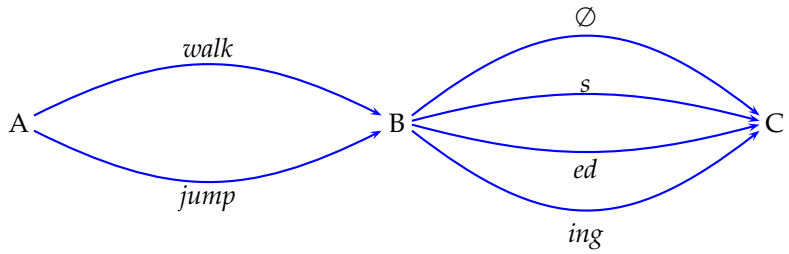
3. List of edges e: the cost of morphological structure: avoid morphological analysis except where it is helpful.

$$\sum_{e(v_1,v_2,m)\in Edges} -log\ pr(v_1) - log\ pr(v_2) - log\ pr(m)$$

(I leave off the specification of the probabilities on the FSA itself, which is also a cost that is specified in bits.)

In addition, a *word* generated by the morphology is the same as a *path* through the FSA. $Pr(w)$ = product of the choice probabilities of for w's path.

So: **for a given corpus, Linguistica seeks the FSA for which the description length of the corpus given the FSA is minimized,** which is something that can be done in an entirely language-independent and unsupervised fashion.



Interpreting this graph: The x-axis and y-axis both quantities measured in *bits*. The x-axis marks how many bits we are allowed to use to write a grammar to describe the data: the more bits we are allowed, the better our description will be, until the point where we are over-fitting the data. Thus each point along the x-axis represents a possible grammar-length; but for any given length l, we care only about the grammar g that assigns the highest probability to the data, i.e., the *best* grammar. The red line indicates how many bits of data are left unexplained by the grammar, a quantity which is equal to -1 * log probability of the data as assigned by the grammar. The blue line shows the sum of these two qunantities (which is the conditional *description length* of the data). The black line gives the length of the grammar.

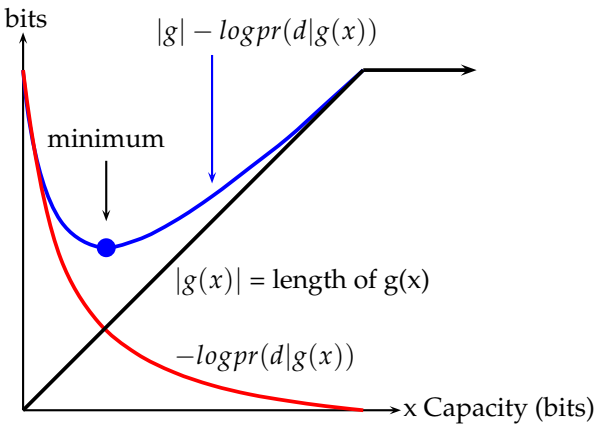


Figure 5: MDL optimization

3cTop part of *Linguistica*'s output from 600,000 words of English:

Signature	exemplar	count	Stem count
$\emptyset - s$	pagoda	20,615	1330
$'s - \emptyset$	Cambodia	30,100	683
$\emptyset - ly$	zealous	14,441	479
$\emptyset - ed - ing - s$	yield	6,235	123
$'s - \emptyset - s$	youngster	4,572	121
$e - ed - es - ing$	zon	3,683	72
$ies - y$	weekl	2,279	124
$\emptyset - ly - ness$	wonderful	2,883	64
$\emptyset - es$	birch	2,472	96
$\emptyset - ed - er - ing - s$	pretend	957	19
$ence - ent$	virul	571	37
$\emptyset - ed - es - ing$	witness	638	18
...			

3 *Learning (morpho)phonology from morphology*

> It never ceases to amaze me how hard it is to develop an explicit algorithm to perform a simple linguistic task, even one that is purely formal. Surely succeeding in that task is a major goal of linguistics.

Morphology treats the items in the lexicon of a language (finite or infinite; let's assume finite to make the math easier). Any given analysis divides the lexicon up into a certain number of subgroups. If there are n subgroups, each equally likely, in a lexicon of size V (V for *vocabulary*), then marking each word costs $-log_2 \frac{n}{V}$. (If the groups are not equally likely, and the i^{th} group has n_i members, then marking a word as being in that group costs $-log_2 \frac{n_i}{V} = log_2 \frac{V}{n_i}$. Each word in the i^{th} group needs to be marked, and all of those markings together costs $n_i \times log_2 \frac{V}{n_i}$. If we can collapse two subgroups analytically, then we savea lot of bits. How many? If the two groups are equal-sized, then we save 1 bit for each item.

Why? Suppose we have two groups, g_1 and g_2 of 100 words out of a vocabulary of 1000 words. Each item in those two groups is marked in the lexicon at a cost of $log_2 \frac{1000}{100} \approx 3.3\,bits$; 200 such words costs us $200 \times 3.32\,bits$ = 664 bits. If they were all treated as part of a single category, the cost of pointing to the larger category would be $-log_2 \frac{200}{1000} = 2.32\,bits$, so we would pay a total of $200 \times 2.32 = 464\,bits$. for a total saving of 200 bits. We actually compute how complex an analysis is. And the morphological analysis that Linguistica provides can be made "cheaper" by decreasing the number of distinct patterns it contains, by adding a (morpho)phonology component after the morphology.

But how can we discover it automatically?

3.1 English verb

(1)

Regular verbal patterns	
jump	walk
jumped	walked
jumping	walking
jumps	walks

(2)

e-final verbal pattern		
move	love	hate
moved	loved	hated
moving	loving	loved
moves	loves	loves

(3)

s-final pattern		
push	miss	veto
pushed	missed	vetoed
pushing	missing	vetoing
pushes	misses	vetoes

(4)

C-doubling pattern		
tap	slit	nag
tapped	slitted	nagged
tapping	slitting	nagging
taps	slits	nags

(5)

y-final pattern		
try	cry	lie*
tried	cried	lied
trying	crying	lying
tries	cries	lies

Figure 6: Some related paradigms

string S	string T	$\Delta_R(S,T)$
jumped	jumping	$\frac{ed}{ing}$
jump	jumping	$\frac{\emptyset}{ing}$
walk	jump	$\frac{walk}{jump}$
walked	jumped	$\frac{walked}{jumped}$

Definition (loose): Given two strings S and T whose longest common initial string is m;
$S = m + s_1$;
$T = m + t_1$.
Then
$\Delta_R(string_1, string_2) = \frac{s_1}{t_1}$

Definition (tight): Given an alphabet A. Define a *cancellation* operation and an *inverse alphabet* A^{-1}: For each $a \in A$ there is an element a^{-1} in A^{-1} such that $aa^{-1} = a^{-1}a = e$. Define an *augmented alphabet* $\mathcal{A} \equiv A \cap A^{-1}$. $\mathcal{A}^*$ is the set of all strings drawn from $\mathcal{A}$. If we add the cancellation operation to $\mathcal{A}^*$, then we get a free group $\mathcal{G}$ in which (e.g.) $ab^{-1}cc^{-1}b = a$. We normally denote the elements in $\mathcal{G}$ by the shortest strings in $\mathcal{A}^*$ that correspond to them.

$\boxed{\Delta_R(S,T) \equiv T^{-1}S.}$

$\boxed{\Delta_L(S,T) \equiv ST^{-1}.}$

E.g. $\Delta_R(jumped, jumping) \equiv (jumping)^{-1}jumped = (ing)^{-1}(jump)^{-1}(jump)(ed) = (ing)^{-1}(ed) = \frac{ed}{ing}$

Still, these matrix are quite similar to one another. We can formalize that observation, if we take advantage of the notion of string difference we defined just above. We extend the definition of Δ_L to $\Sigma^* \times \Sigma^*$ in this way:

$$\Delta_L(\frac{a}{b}, \frac{c}{d}) = \frac{\Delta_L(a,c)}{\Delta_L(b,d)} \tag{6}$$

If we define Δ_L on a matrix as the item-wise application of that operation on the individual members, then we can express the difference between 6 and 7 in this way (where we indicate $\frac{\emptyset}{\emptyset}$ with a blank). See Figures 7,8 on next two pages.

3.2 Hungarian

See Figure 10 below.

3.3 Spanish

See Figure 9 below.

4 Conclusion

Let P be a sequence of words (think *P[aradigm]*) of length n.
We define the quotient $P \div Q$ of two sequences P, Q of the same length n as a 2×2 matrix, where

$$P \div Q(i,j) \equiv \Delta_L(p_i, q_j)$$

	jump	jumps	jumped	jumping	
jump		$\frac{\emptyset}{s}$	$\frac{\emptyset}{ed}$	$\frac{\emptyset}{ing}$	$\emptyset$
jumps	$\frac{s}{\emptyset}$		$\frac{s}{ed}$	$\frac{s}{ing}$	**s**
jumped	$\frac{ed}{\emptyset}$	$\frac{ed}{s}$		$\frac{ed}{ing}$	**ed**
jumping	$\frac{ing}{\emptyset}$	$\frac{ing}{s}$	$\frac{ing}{ed}$		**ing**
	$\emptyset$	**s**	**ed**	**ing**	

	push	pushes	pushed	pushing	
push		$\frac{\emptyset}{es}$	$\frac{\emptyset}{ed}$	$\frac{\emptyset}{ing}$	$\emptyset$
pushes	$\frac{es}{\emptyset}$		$\frac{s}{d}$	$\frac{es}{ing}$	**es,s**
pushed	$\frac{ed}{\emptyset}$	$\frac{d}{s}$		$\frac{ed}{ing}$	**d,ed**
pushing	$\frac{ing}{\emptyset}$	$\frac{ing}{es}$	$\frac{ing}{ed}$		**ing**
	$\emptyset$	**es, s**	**d, ed**	**ing**	

	move	moves	moved	moving	
move		$\frac{\emptyset}{s}$	$\frac{\emptyset}{d}$	$\frac{e}{ing}$	**e,** $\emptyset$
moves	$\frac{s}{\emptyset}$		$\frac{s}{d}$	$\frac{es}{ing}$	**es, s**
moved	$\frac{d}{\emptyset}$	$\frac{d}{s}$		$\frac{ed}{ing}$	**d, ed**
moving	$\frac{ing}{e}$	$\frac{ing}{es}$	$\frac{ing}{ed}$		**ing**
	e,$\emptyset$	**es, s**	**d, ed**	**ing**	

	slit	slits	slitted	slitting	
slit		$\frac{\emptyset}{s}$	$\frac{\emptyset}{ted}$	$\frac{\emptyset}{ting}$	$\emptyset$
slits	$\frac{s}{\emptyset}$		$\frac{s}{ted}$	$\frac{s}{ting}$	**s**
slitted	$\frac{ted}{\emptyset}$	$\frac{ted}{s}$		$\frac{ed}{ing}$	**ed, ted**
slitting	$\frac{ting}{\emptyset}$	$\frac{ting}{s}$	$\frac{ing}{ed}$		**ing, ting**
	$\emptyset$	**s**	**ed, ted**	**ing, ting**	

	try	tries	tried	trying	
try		$\frac{y}{ies}$	$\frac{y}{ied}$	$\frac{\emptyset}{ing}$	**y,** $\emptyset$
tries	$\frac{ies}{y}$		$\frac{s}{d}$	$\frac{ies}{ying}$	**ies, s**
tried	$\frac{ied}{y}$	$\frac{d}{s}$		$\frac{ied}{ying}$	**ied, d**
trying	$\frac{ing}{\emptyset}$	$\frac{ying}{ies}$	$\frac{ying}{ied}$		**ing, ying**
	y,$\emptyset$	**ies, s**	**d, ied**	**ing, ying**	

Figure 7: Matrix of string differences

In particular

$$P \div P(i,j) \equiv \Delta_L(p_i, p_j)$$

We may compare two paradigms then as the *second difference*:

$$\bigtriangledown(P,Q) \equiv (P \div P) \div (Q \div Q)$$

This is what we have explored in this handout.

Many morphophonological changes emerge as the second difference of sets ('paradigms') of words.

jump:move	1	2	3	4
1. ∅			$\frac{\emptyset}{e}$	$\frac{\emptyset}{e}$
2. *s*			$\frac{\emptyset}{e}$	$\frac{\emptyset}{e}$
3. *ed*	$\frac{e}{\emptyset}$	$\frac{e}{\emptyset}$		
4. *ing*	$\frac{e}{\emptyset}$	$\frac{e}{\emptyset}$		

jump:push	1	2	3	4
1. ∅		$\frac{e}{\emptyset}$		
2. *s*	$\frac{\emptyset}{e}$		$\frac{\emptyset}{e}$	$\frac{\emptyset}{e}$
3. *ed*		$\frac{e}{\emptyset}$		
4. *ing*		$\frac{e}{\emptyset}$		

jump:split	1	2	3	4
1. ∅		$\frac{t}{\emptyset}$	$\frac{t}{\emptyset}$	
2. *s*		$\frac{t}{\emptyset}$	$\frac{t}{\emptyset}$	
3. *ed*	$\frac{\emptyset}{t}$	$\frac{\emptyset}{t}$		
4. *ing*	$\frac{\emptyset}{t}$	$\frac{\emptyset}{t}$		

jump:try	1	2	3	4
1. ∅	$\frac{ie}{y}$	$\frac{i}{y}$		
2. *s*	$\frac{y}{ie}$	$\frac{y}{ie}$	$\frac{y}{ie}$	
3. *ed*	$\frac{y}{i}$	$\frac{ie}{y}$		
4. *ing*		$\frac{ie}{y}$		

Figure 8: Difference of differences: English verb

	emberem	embered	embere	emberünk	emberetek	emberük
emberem		$\frac{m}{d}$	$\frac{m}{\emptyset}$	$\frac{em}{ünk}$	$\frac{m}{tek}$	$\frac{em}{ük}$
embered	$\frac{d}{m}$		$\frac{d}{\emptyset}$	$\frac{ed}{ünk}$	$\frac{d}{tek}$	$\frac{ed}{ük}$
embere	$\frac{\emptyset}{m}$	$\frac{\emptyset}{d}$		$\frac{e}{ünk}$	$\frac{\emptyset}{tek}$	$\frac{e}{ük}$
emberünk	$\frac{ünk}{em}$	$\frac{ünk}{ed}$	$\frac{ünk}{e}$		$\frac{ünk}{etek}$	$\frac{nk}{k}$
emberetek	$\frac{tek}{m}$	$\frac{tek}{d}$	$\frac{tek}{\emptyset}$	$\frac{etek}{ünk}$		$\frac{etek}{ük}$
emberük	$\frac{ük}{em}$	$\frac{ük}{ed}$	$\frac{ük}{e}$	$\frac{k}{nk}$	$\frac{ük}{etek}$	

	dögöm	dögöd	döge	dögünk	dögötek	dögük
dögöm		$\frac{m}{d}$	$\frac{öm}{e}$	$\frac{öm}{ünk}$	$\frac{m}{tek}$	$\frac{öm}{ük}$
dögöd	$\frac{d}{m}$		$\frac{öd}{e}$	$\frac{öd}{ünk}$	$\frac{d}{tek}$	$\frac{öd}{ük}$
döge	$\frac{e}{öm}$	$\frac{e}{öd}$		$\frac{e}{ünk}$	$\frac{e}{ötek}$	$\frac{e}{ük}$
dögünk	$\frac{ünk}{öm}$	$\frac{ünk}{öd}$	$\frac{ünk}{e}$		$\frac{ünk}{ötek}$	$\frac{nk}{k}$
dögötek	$\frac{tek}{m}$	$\frac{tek}{d}$	$\frac{ötek}{e}$	$\frac{ötek}{ünk}$		$\frac{ötek}{ük}$
dögük	$\frac{ük}{öm}$	$\frac{ük}{öd}$	$\frac{ük}{e}$	$\frac{k}{nk}$	$\frac{ük}{ötek}$	

Differences of differences

emberük		∅	$\frac{e}{ö}$	$\frac{e}{ö}$	∅	$\frac{e}{ö}$
emberük	∅		$\frac{e}{ö}$	$\frac{e}{ö}$	∅	$\frac{e}{ö}$
emberük	$\frac{ö}{e}$	$\frac{ö}{e}$		∅	$\frac{ö}{e}$	∅
emberük	$\frac{ö}{e}$	$\frac{ö}{e}$	∅		$\frac{ö}{e}$	∅
emberük	∅	∅	$\frac{e}{ö}$	$\frac{e}{ö}$		$\frac{e}{ö}$
emberük	$\frac{ö}{e}$	$\frac{ö}{e}$	∅	∅	$\frac{ö}{e}$	

Figure 9: Hungarian vowel harmony: commutative free group

	hablar	hablo	hablas	habla	hablamos	hablan	hablé	hable	hables
hablar		$\frac{ar}{o}$	$\frac{r}{s}$	$\frac{r}{\emptyset}$	$\frac{r}{mos}$	$\frac{r}{n}$	$\frac{ar}{é}$	$\frac{ar}{e}$	$\frac{ar}{es}$
hablo	$\frac{o}{ar}$		$\frac{o}{as}$	$\frac{o}{a}$	$\frac{o}{amos}$	$\frac{o}{an}$	$\frac{o}{é}$	$\frac{o}{e}$	$\frac{o}{es}$
hablas	$\frac{s}{r}$	$\frac{as}{o}$		$\frac{s}{\emptyset}$	$\frac{s}{mos}$	$\frac{s}{n}$	$\frac{as}{é}$	$\frac{as}{e}$	$\frac{as}{es}$
habla	$\frac{\emptyset}{r}$	$\frac{a}{o}$	$\frac{\emptyset}{s}$		$\frac{\emptyset}{mos}$	$\frac{\emptyset}{n}$	$\frac{a}{é}$	$\frac{a}{e}$	$\frac{a}{es}$
hablamos	$\frac{mos}{r}$	$\frac{amos}{o}$	$\frac{mos}{s}$	$\frac{mos}{\emptyset}$		$\frac{mos}{n}$	$\frac{amos}{é}$	$\frac{amos}{e}$	$\frac{amos}{es}$
hablan	$\frac{n}{r}$	$\frac{an}{o}$	$\frac{n}{s}$	$\frac{n}{\emptyset}$	$\frac{n}{mos}$		$\frac{an}{é}$	$\frac{an}{e}$	$\frac{an}{es}$
hablé	$\frac{é}{ar}$	$\frac{é}{o}$	$\frac{é}{as}$	$\frac{é}{a}$	$\frac{é}{amos}$	$\frac{é}{an}$		$\frac{é}{e}$	$\frac{é}{es}$
hable	$\frac{e}{ar}$	$\frac{e}{o}$	$\frac{e}{as}$	$\frac{e}{a}$	$\frac{e}{amos}$	$\frac{e}{an}$	$\frac{e}{é}$		$\frac{\emptyset}{s}$
hables	$\frac{es}{ar}$	$\frac{es}{o}$	$\frac{es}{as}$	$\frac{es}{a}$	$\frac{es}{amos}$	$\frac{es}{an}$	$\frac{es}{é}$	$\frac{s}{\emptyset}$	

	buscar	busco	buscas	busca	buscamos	buscan	busqué	busque	busques
buscar		$\frac{ar}{o}$	$\frac{r}{s}$	$\frac{r}{\emptyset}$	$\frac{r}{mos}$	$\frac{r}{n}$	$\frac{car}{qué}$	$\frac{car}{que}$	$\frac{car}{ques}$
busco	$\frac{o}{ar}$		$\frac{o}{as}$	$\frac{o}{a}$	$\frac{o}{amos}$	$\frac{o}{an}$	$\frac{co}{qué}$	$\frac{co}{que}$	$\frac{co}{ques}$
buscas	$\frac{s}{r}$	$\frac{as}{o}$		$\frac{s}{\emptyset}$	$\frac{s}{mos}$	$\frac{s}{n}$	$\frac{cas}{qué}$	$\frac{cas}{que}$	$\frac{cas}{ques}$
busca	$\frac{\emptyset}{r}$	$\frac{a}{o}$	$\frac{\emptyset}{s}$		$\frac{\emptyset}{mos}$	$\frac{\emptyset}{n}$	$\frac{ca}{qué}$	$\frac{ca}{que}$	$\frac{ca}{ques}$
buscamos	$\frac{mos}{r}$	$\frac{amos}{o}$	$\frac{mos}{s}$	$\frac{mos}{\emptyset}$		$\frac{mos}{n}$	$\frac{camos}{qué}$	$\frac{camos}{que}$	$\frac{camos}{ques}$
buscan	$\frac{n}{r}$	$\frac{an}{o}$	$\frac{n}{s}$	$\frac{n}{\emptyset}$	$\frac{n}{mos}$		$\frac{can}{qué}$	$\frac{can}{que}$	$\frac{can}{ques}$
busqué	$\frac{qué}{car}$	$\frac{qué}{co}$	$\frac{qué}{cas}$	$\frac{qué}{ca}$	$\frac{qué}{camos}$	$\frac{qué}{can}$		$\frac{é}{e}$	$\frac{é}{es}$
busque	$\frac{que}{car}$	$\frac{que}{co}$	$\frac{que}{cas}$	$\frac{que}{ca}$	$\frac{que}{camos}$	$\frac{que}{can}$	$\frac{e}{é}$		$\frac{\emptyset}{s}$
busques	$\frac{ques}{car}$	$\frac{ques}{co}$	$\frac{ques}{cas}$	$\frac{ques}{ca}$	$\frac{ques}{camos}$	$\frac{ques}{can}$	$\frac{es}{é}$	$\frac{s}{\emptyset}$	

	hables	hables	hables	hables	hables	hables	hables	hables	hables
hables							$\frac{qu}{c}$	$\frac{qu}{c}$	$\frac{qu}{c}$
hables							$\frac{qu}{c}$	$\frac{qu}{c}$	$\frac{qu}{c}$
hables							$\frac{qu}{c}$	$\frac{qu}{c}$	$\frac{qu}{c}$
hables							$\frac{qu}{c}$	$\frac{qu}{c}$	$\frac{qu}{c}$
hables							$\frac{qu}{c}$	$\frac{qu}{c}$	$\frac{qu}{c}$
hables							$\frac{qu}{c}$	$\frac{qu}{c}$	$\frac{qu}{c}$
hables	$\frac{c}{qu}$	$\frac{c}{qu}$	$\frac{c}{qu}$	$\frac{c}{qu}$	$\frac{c}{qu}$	$\frac{c}{qu}$			
hables	$\frac{c}{qu}$	$\frac{c}{qu}$	$\frac{c}{qu}$	$\frac{c}{qu}$	$\frac{c}{qu}$	$\frac{c}{qu}$			
hables	$\frac{c}{qu}$	$\frac{c}{qu}$	$\frac{c}{qu}$	$\frac{c}{qu}$	$\frac{c}{qu}$	$\frac{c}{qu}$			

Figure 10: Difference of differences: Spanish verb