naclo

SOLUTIONS

NAACL

YAHOO!

LSA Linguistic Society of America

DARPA

NSF

ARL

Carnegie Mellon University Language Technologies Institute

Brandeis University Computational Linguistics

MICHIGAN M
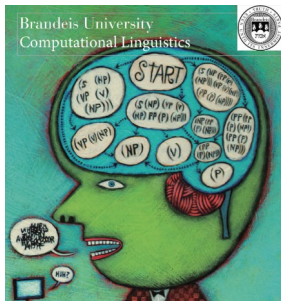
W UNIVERSITY of WASHINGTON

acm

SIGIR Special Interest Group on Information Retrieval

Choosito! © search and learn

LDC

UNIVERSITY OF **MARYLAND**

The Tenth Annual

North American Computational Linguistics Olympiad 2016

www.nacloweb.org

Invitational Round March 10, 2016

Serious language puzzles that are surprisingly fun!
-Will Shortz, Crossword editor of The New York Times and Puzzlemaster for NPR

(I) Deriving Enjoyment (1/2)

The derivations, which are formed according to a uniform set of rules, indicate diminutive, including derived names is shown by the suffix, -ič, or feminine shown by the suffix -ica. If the noun is already feminine (i.e. ends in an –a), this suffix has the diminutive meaning (at least in the data given).

In addition there is consonant change (palatalization) of k to č, g to ž, h to š. This happens whether there is an –a ending or not. Not all cases are exemplified in the given data (volk, roka, -g, knjiga, menih, -ha) so it has to be inferred that the rule applies to both genders.

And with a polysyllabic stem ending in e+consonant, drop the e.

I1.

| (a) bivol	buffalo	bivolica	female buffalo
bog	god	(b) božič	small god
grm	bush	(c) grmič	small bush
knjiga	book	(d) knjižica	booklet
muha	fly	(e) mušica	midge (small fly)
orel	eagle	orlica	female eagle
		(f) orlič	eaglet
osel	donkey	oslič	donkey foal
		(g) oslica	jenny (i.e. she-donkey)
otrok	child	(h) otročič	baby
oven	sheep	(i) ovnič	lamb
Pavel	Paul	(j) Pavlič	Paulson
(k) rak	crab	račič	baby crab
(l) Štefan	Stephen	Štefanič	Stephenson
Tomaž	Thomas	(m) Tomažič	Thomson
(n) trn	thorn	trnič	small thorn
veter	wind	(o) vetrič	draft (current of air)
vrh	peak	(p) vršič	small peak
zid	wall	(q) zidič	small wall
žep	pocket	(r) žepič	small pocket

I2. It might be rož or rog: you can't tell from the diminutive whether the stem has undergone palatalization (like bog) or ended in ž anyway (like Tomaž). In fact it is rog.

I3. It might be čoln or čolen: you can't tell from the diminutive whether or not there is an e in the final syllable of the stem. In fact it is čoln.

With example (a), you might think that bivola is a possible answer. But in the data there is no example of the –ica suffix meaning 'feminine' attaching to a word ending in –a: in all such cases, the derived word is a diminutive. This is because (in this data at least), all feminine stems end in –a (though that is not a general rule for Slovene). For example the word živalica means 'small animal', not 'female animal' because the root žival is actually (already) feminine.
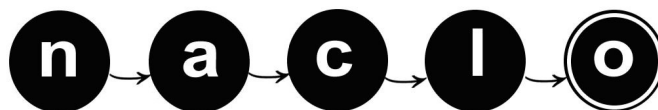
(J) Get Edumacated! (1/1)

J1.

	Alan	Barbara	Chris
octet	(a) (F) OcamaTET	(b) (B) OCtemaTET	I dunno…
purple	(c) (A) PURpamaPLE	(d) (E) PURplemaPLE	(e) (G) PURRRmaPLE
tuba	(f) (C) TUbamaBA	TUbamaBA	(g) (D) TUUUmaBA

J2.

	Alan	Barbara	Chris
antiseptic	(a) ANtimaSEPtic	(b) ANtimaSEPtic	(c) ANtimaSEPtic
Canada	(d) CAnamaDA	(e) CAnamaDA	(f) CAnamaDA
feudalism	(g) FEUdamaLISm	FEUdamaLISm	(h) FEUdamaLISm
optics	(i) OPamaTICS	(j) OPtimaTICS	(k) I dunno; (OPPmaTICS also permitted)
party	PARtamaTY	(l) PARtymaTY	(m) PAARRmaTY
table	(n) TAbamaBLE	(o) TAblemaBLE	(p) TAAmaBLE
water	(q) WAtamaTER	(r) WAtermaTER	WAAAmaTER

N.B. Some degree of variance from these answers is expected and permitted, especially regarding syllabification, stress, expressing length for Chris, etc. The important part is that Alan always tries to insert <a> and will fall back to a partial reduplicant <Ca> when necessary, Barbara reduplicates CV/CR/CL, and Chris lengthens.

Alan putting in an additional consonant before -ama- is a minor mistake, to be graded as a minor deduction or no deduction at all. Same with Barbara reduplicating a fuller syllable like TIC or TICS instead of just TI in "optics".

J3. Various kinds of answers would be acceptable here; what we're looking for is nontrivial insight into what might be going on rather than an answer according to what a phonologist might answer.
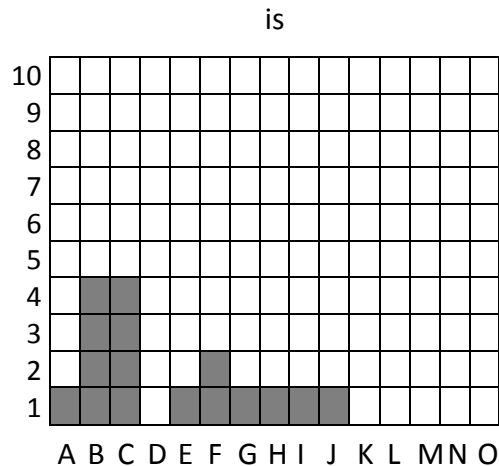
Nonetheless, a top-quality answer should probably bring up the idea that the respondents are trying to achieve contradictory goals, and that different respondents' strategies here are prioritizing different goals. (Other top quality answers are possible, but I can't anticipate what they might be.)

Points are available for observing that (a) there's something contradictory/impossible/problematic about asking this for two-syllable words, (b) suggesting why this might be, (c) observing that the speakers are systematic in how they solve this, (d) describing what each speaker does, and (e) making an attempt at explaining why speakers strategies are differing.
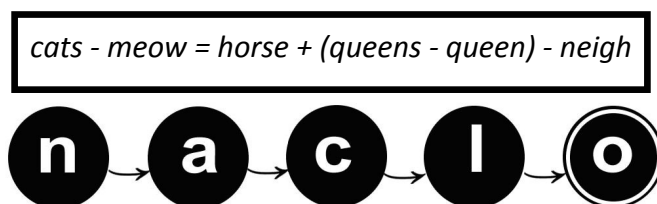
(K) Kings, Queens, and Counts (1/2)

K1.

is

The graph is filled in by looking at a window of two words on either side of each occurrence of is, giving the graph above. For example, whether occurs once within two words of is, the occurs four times within two words of is, etc.

K2.

a. antismartnessesquely	mystery word #10
b. aunt	mystery word #11
c. big	mystery word #5
d. can	mystery word #4
e. cats	mystery word #2
f. Kenya	mystery word #3
g. Kenyan	mystery word #9:
h. meow	mystery word #7
i. strange	mystery word #1
j. strangest	mystery word #8
k. the	mystery word #6

The key insight for this part is that analogies between words can be expressed by adding and subtracting graphs. For example, the analogy "*queen* is to *king* as *woman* is to *man*" is reflected in the fact that the difference between the graphs of *queen* and *king* is roughly equal to the difference between the graphs of *woman* and *man* (e.g., *queen* has 6 fewer co-occurrences with A than *king*; 2 more co-occurrences with C than *king*; 7 more co-occurrences with E than *king*; 1 fewer co-occurrence with H than *king*; 3 more co-occurrences with M than *king*; and 3 fewer co-occurrences with N than *king*. *Woman* and *man* have roughly the same differences in co-occurrences). I say "roughly" because there is a margin of error of plus or minus one in all cases to reflect the fact that the addition and subtraction of distributional vectors is by no means exact. Thus, for example, *aunt* can be identified as 11 because the difference between *uncle* and graph 11 is similar to the difference between *king* and *queen* or the difference between *man* and *woman*. Similarly, *cats* and *meow* can be identified as the pair that satisfies:

$$\text{cats - meow = horse + (queens - queen) - neigh}$$

and *Kenya* and *Kenyan* are the pair such that:

> *Kenya - Kenyan*
> =
> *India - (rupee - (ariary - (Antananarivo - (Berlin - (Merkel - (Roussef - Brazilian))))))*

Lastly, *the* and *antismartnessesquely* can be identified as the words that occur with other words extremely frequently and not at all, respectively.

K3. *Mystery word #4* is *can*. You might expect *can* to have the graph labeled *expected graph for mystery word #4* because that graph reflects the analogy "*king* is to *kings* as *can* is to *cans*" or "*queen* is to *queens* as *can* is to *cans*." However, in addition to being the singular form of *cans* as in "a **can** of soup," *can* also is an auxiliary verb as in "Nothing **can** stop me now!" Thus, when the graph is formed for *can*, it will include counts for the noun *can* but also the (much more common) auxiliary verb *can* (plus the verb *can*, as in "I love to **can** vegetables") which muddies the waters even further). That is why the actual graph for *can* has much higher counts than the *expected graph for mystery word #4*.

(I know the solution provided above is kind of hypocritical for being more than the requested maximum of two sentences. An actual answer could just be something like "Mystery word #4 is *can*, which has multiple meanings." Anything that mentions how *can* has more than one meaning will get full points.)

(L) The Short Hand of the Law (1/1)

L1. B, E, C, A, D

L2. THE DEFENDANT: Absolutely one hundred percent not guilty.

(M) Sound Judgments (1/2)

M1. The correct historical order is: D, C, G, F, B, A, E

And the rules are:
> c > k
> j > g
> b^h > Ø / m __
> b^h > p
> g > k
> s > Ø / __ #

Working out an order of the steps relative to one another should be straightfoward by any method, e.g. by observing that the first word begins either b^h- or p-, and that the simplest conclusion is that this represents a historical change from one to the other (so all the forms with b^h- will be at 'one end' of the timeline).

This will give an ordering of the elements relative to one another (e.g. "this stage goes in between these other two"), but the next step is to tell which part is the beginning and which is the end. The trick to this is to realize that some changes are irreversible, specifically the mergers of c with already existent k (this is why the word kóro- is included) and later g with k. The statement in the rubric that sound changes *must* be regular means that the reverse of this can't happen; that would require an irregular change of k to g in some words but k in others. (A logical possibility is that k becomes g before on but not or; candidates should realize that this is not the most parsimonious analysis.)

M2. The rules given in the question are lettered below.

(A) o > ë	(B) nt > Ø / __ #	(C) e > ə	(D) or > ur / __ #
(E) ē, ī, ū > e, i, u	(F) m > n / __ #	(G) t > c / __ e, ē	(H) d > ś / __ e, ē
(I) n > Ø / __ #	(J) m > n / __ t	(K) u > ə	(L) k > ś / __ e, ē
(M) w > wʸ / __ i, ī	(N) g > k	(O) d > Ø / __ #	(P) s > Ø / __ #
(Q) ti > Ø / n __ #			

The first thing to realize here is that (as the question suggests) only a partial order can be deduced. For example, 15 can be put anywhere in the ordering, because it doesn't have an effect on any other rules. The rules that can be ordered are the ones that make reference to each other's inputs and outputs.

The orderings are possible to work out (by a kind of *modus tollens*).

(D) comes before both (A) and (K)
In *akëntər, we have: -or > -ur (by (D)) > -ər (by (K)). If the order were (A) > (D), we would have: -or > -ër (by (A)) and (D) would fail to apply. If the order were (K) > (D), we would have: -or > -ur, but then no subsequent change to -ər because (K) has already applied earlier.

n → a → c → l → o

(G) and (H) both come before (C)
In *cə́kʷë, we have: té- > cé- (by (G)) > cə́ (by (C)). If the order were (C) > (G), we would have: té- > tə́ (by (C))
and (G) would fail to apply. The same reasoning applies to (H) in the word *śə́kə.

(C) comes before (E)
In *pacér, we have: -ế- > -é- (by (E)), where (C) has already applied earlier. If the order were (E) > (C), we
would have: -ế- > -é- (by (E)) > -ə́- (by (C)).

(F) comes before (I)
In *kənté̋, we have: -m > -n (by (F)) > Ø (by (I)). If the order were (I) > (F), we would have: -m > -n (by (F)), but
then no subsequent deletion because (I) has already applied earlier.

(J) comes before (B)
In *śə́kə, we have: -əmt > -ənt (by (J)) > -ə (by (B)). If the order were (B) > (J), we would have: -əmt > -ənt (by
(J)), but then no subsequent deletion of -nt because (B) has already applied earlier.

(I) comes before (Q)
In *akën, we have: -nti > -n (by (Q)), where (I) has already applied earlier. If the order were (Q) > (I), we would
have: -nti > -n (by (Q)) > Ø (by (I)).

(L), (M), (N), (O), and (P) are unorderable with respect to any other rules.

A neat diagram of a solution might look like this, where a downward arrow means 'precedes'.

(L)   (M)   (N)   (O)   (P)
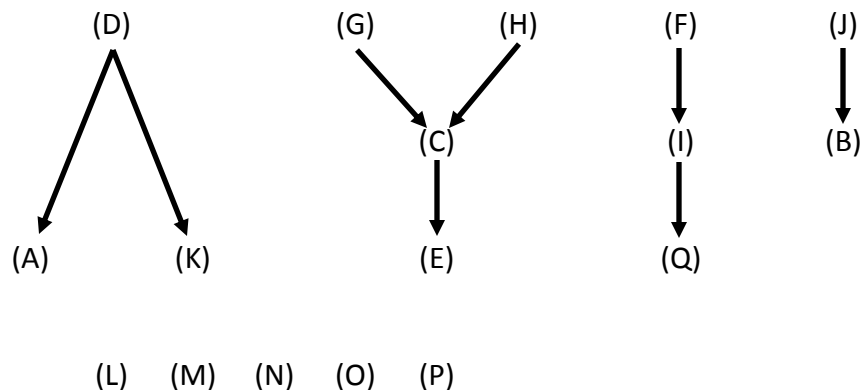
(N) What happened at the chess tournament? (1/1)

N1. Valaki megvert valakit.

N2. Kit vert meg valaki?

N3. Senki nem verte meg a Petyát.

N4. Valakit senki nem vert meg.

N5. Senki nem vert meg mindenkit.

N6. Senkit nem vert meg a Petya.

N7. Ki nem vert meg senkit?

N8. Valaki senkit nem vert meg.

N9. Mindenki megvert valakit.

N10. Valaki megverte a Marcit.

N11. Senkit nem vert meg senki.

N12. Kit nem vert meg senki?

N13. Nem vertem meg senkit.

(H) Someone beat someone.

(G) Who got beaten by someone?

(F) No one beat Peter.

(M) There is someone who didn't get beaten.

(A) No one beat everyone (at e.g. chess).

(L) Peter beat no one.

(J) Who didn't beat anyone?

(K) There's someone who didn't beat anyone.

(I) Everyone beat someone.

(D) Someone beat Martin.

(C) No one got beaten.

(B) Who wasn't beaten by anyone?

(E) I didn't beat anyone.

n a c l o

(O) Don't Sell the House! (1/2)

O1. Cáu cháhn đày non.   I can truly sleep.
O2. Da páy non!   Don't go sleep!
O3. Mưhn bô sạhm mi slờng hẻht hơn mi?   Was she also unwilling to build the house?
O4. Mưhn ngám bô sạhm páy hơn.   She also went home just now.
O5. I wasn't about to eat it just previously.   Cáu náhc-thày ca mi kíhn.
O6. She didn't have to eat it alone like that just now.   Mưhn ngám mi slày tảhng kíhn.
O7. The house truly can't eat you.   Hơn mi cháhn đày kíhn mưhng.
O8. Then were you also about to go just previously?   Mưhng náhc-thày chớng ca bô sạhm páy mi?

Explanation:

In Nung, word order is fixed, meaning that for a given sentence, there is only one correct way to order the words. We can figure out from comparing sentences with the same subject, same adverb, same main verb, or same object that in Nung, sentences are constructed in subject-adverb-verb-object-question.marker order, where there can be many adverbs and 'it' as an object is left unsaid but implied. 'Like that' is treated as an object in terms of word order.

The correct order of adverbs is as follows, listed from first to last:

náhc-thày	just previously
ngám	just now
chớng	then
ca	about to
vửhn nhahng	continue to
bô sạhm	also
mi	not
cháhn	truly
slờng / slày / fải / đày	want / [don't] have to / have to / can
tảhng	alone

In addition, we know that da 'don't' and tan đohc 'only' both occur before the verb, but in our data, no other adverbs occur in the same sentence as da or tan đohc, so it is unclear what the correct order should be.

The main question and challenge of the problem consists of identifying the correct positioning of adverbs, which always occur in a fixed order. A strategy for discovering this is to list all the orderings attested, and merge those to create a comprehensive order for all the adverbs provided. Crucially, it is necessary to also consider the Nung sentences provided without translations in order to clarify where in the pattern the adverb ngám lies.
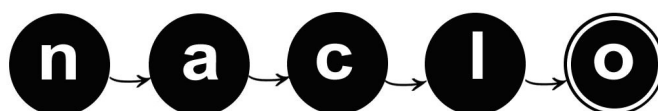
(O) Don't Sell the House! (2/2)

The secondary challenge of the problem is to map Nung sentences onto English ones: not only the word order but also the meaning may differ slightly. For example, in Nung 'build a house' is expressed as 'do house'. For the purposes of this problem, small differences in translation that represent equally valid interpretations of a Nung or English sentence are allowed as long as they follow proper Nung or English word order.

List of words not previously defined:

cáu	I
khải	sell
mưhng	you
hảhn	see
kíhn	eat
non	sleep
hểht	do, build
mi	question marker
páy	go
hơn	house, home
mưhn	she
pehn tế	like that

(P) A Matter of Horn Clauses (1/1)

P1. Rules:

 V(help).
 V(let).
 N(John).
 N(Kim).
 S(xy) :- *V*(x), *N*(y).
 S(xy) :- *S*(x), *S*(y).

Sample derivation:

 S(let Kim) :- *V*(let), *N*(Kim).
 S(let John) :- *V*(let), *N*(John).
 S(let Kim let John) :- *S*(let Kim), *S*(let John).
 S(help John) :- *V*(help), *N*(John).
 S(help John let Kim let John) :- *S*(help John), *S*(let Kim let John).

P2. Rules:

 N(Jan).
 N(Hans).
 Ddat(em).
 Dacc(de).
 Vdat(hälfe).
 Vacc(laa).
 V P(xy, z) :- *Ddat*(x), *N*(y), *Vdat*(z).
 V P(xy, z) :- *Dacc*(x), *N*(y), *Vacc*(z).
 V P(wy, xz) :- *V P*(w, x); *V P*(y, z).
 S(xy) :- *V P*(x, y).
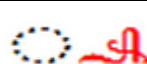
Sample derivation:

 V P(em Jan, hälfe) :- *Ddat*(em), *N*(Jan), *Vdat*(hälfe).
 V P(em Hans, hälfe) :- *Ddat*(em), *N*(Hans), *Vdat*(hälfe).
 V P(em Jan em Hans, hälfe hälfe) :- *V P*(em Jan, hälfe), *V P*(em Hans, hälfe).
 V P(de Hans, laa) :- *Dacc*(de), *N*(Hans), *Vacc*(laa)..
 V P(de Hans em Jan em Hans, laa hälfe hälfe) :- *V P*(de Hans, laa), *V P*(em Jan em Hans, hälfe hälfe).
 S(de Hans em Jan em Hans laa hälfe hälfe) :- *V P*(de Hans em Jan em Hans, laa hälfe hälfe).
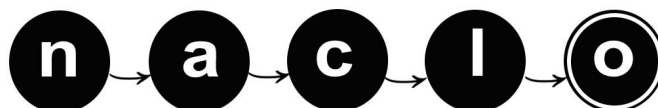
Javanese script is read from left to right, and each consonant has an inherent vowel 'a'. Here are the consonants when they are C1 in C1(C2)V(C3) and C2 in C1C2V(C3).

Latin Script	C1	C2 (suppresses the vowel of C1)
Øa (ha)*	ꦲ	-**
na	ꦤ	-
ra	ꦫ	
re***	ꦉ	
ka	ꦏ	-
ta	ꦠ	
sa	ꦱ	
la	ꦭ	-
pa	ꦥ	-
nya	ꦚ	-
ma	ꦩ	-
ga	ꦒ	-

(Q) A Cup of Javanese (2/5)

Javanese script is read from left to right, and each consonant has an inherent vowel 'a'. Here are the consonants when they are C1 in C1(C2)V(C3) and C2 in C1C2V(C3).

Latin Script	C1	C2 (suppresses the vowel of C1)
ba		
nga		-

*The consonant is either 'Ø' (no consonant) or 'h,' but the problem contains only the former.
**The '-' means that the form exists, but not in this problem.
***The CV combination 're' (historical remnant of /ɽ/) has its own special letters.

'ng,' 'h,' and 'r' must be C3 in (C1)(C2)VC3 before another C or at the end of a word. All other consonants after V must be C1 of the next syllable. If these consonants end a word, a 'vowel suppressor' must be added to suppress the inherent 'a.'

Latin Script	C3
-ng	
-h	
-r	
-C (vowel suppressor)	

Consonants can be modified to change the inherent vowel 'a' in C1(C2)V(C3).

Latin Script	V*
e**	

n → a → c → l → o
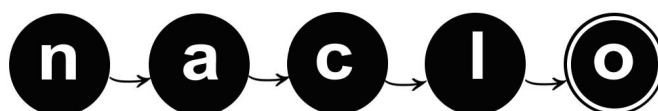
Latin Script	V*
i	
é	
u	
o	

* If C2 is on the right side of C1, then 'e,' 'i,' and 'u' modify C2. 'é' is always placed before C1, and 'o' always surrounds C1 and C2.
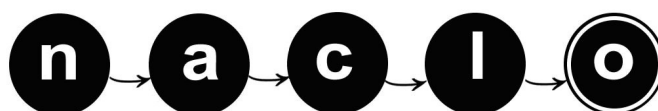
** 'e' is not marked in the special combination 're'.

When a underline proper noun begins with a vowel, a special letter is used for the initial vowel.

Latin Script	Special Letter
A	
I	

Answer key

Q1.

Javanese script	Latin Script	Meaning
	a. kerep	often

	b. aksara	letter / script
ꦮꦴꦲꦺꦴꦄꦮ	c. mbongkar	to unload
ꦤꦗꦵ2ꦲꦺꦴ	d. mrikso	to examine
ꦿꦤꦲꦺꦴꦄ2	e. murungaké	to cancel
ꦧꦮꦾꦤꦲ		

Q2.

	Javanese script	Latin Script	Meaning
a.	ꦤꦠꦺꦴ2ꦤꦭꦵ2	nyolong	to steal
b.	ꦩꦶꦮꦭꦶ	sepalih	half
c.	ꦕꦾꦭꦶꦲꦱ꧀	trengginas	lively

	Javanese script	Latin Script	Meaning
d.		Antartika	Antarctica
e.		Istanbul	Istanbul

R1.

(a) baray	I taught		(i) gashay	He entered
(b) baajiday	He prevented		(j) go'day	He cut
(c) bi'iday	He destroyed		(k) helay	I found
(d) bilaabtay	He began		(l) kacday	He rose
(e) cunay	I ate		(m) qaaday	He took
(f) daaqday	He grazed		(n) xidhay	He closed
(g) dhacday	He fell		(o) walaaqday	He stirred
(h) faraxday	He was happy			

The 1st person is formed by adding –*ay* to the stem, while the 3rd person is formed by adding –*tay* to the stem, but with some morphophonemic changes:

- *t* changes to *d* after *q* (6, 13, 30), *c* (14,24), *x* (7,18), and ' (4, 21), the so-called guttural consonants; the examples (g) and (l) require the solver to recognize that *c* acts like its voiceless counterpart *x*, as no example ending in *c* is given.
- stem apparently ending in *y*, drop the *y* (actually, *y* is added to a stem ending in a vowel to form the 1st person, but you cannot tell this from the data) and change *t* to *d* (1, 5, 8)
- stem ending in *l*, change *lt* to *sh* (17, 20, 22)
- stem ending in *d* or *dh*, just add –*ay* (16, 19, 23, 26)

Since we don't ask for an explanation, we need not worry about other, consistent, solutions, for example that the 3rd person is formed by adding –*day*, or by inserting a *d* before the *ay* with a list of exceptions, which would be something like the following:

- *d* changes to *t* after *b,g,n,r,s*
- drop the (stem-final) *y*
- stem ending in *l*, change *ld* to *sh* (or, -*lay* becomes -*shay*)
- stem ending in *d* or *dh*, just add –*ay*