Background for Hundred Sentences and Morphology Assignments: Part 1

February 3, 2016

Next two assignments

- One hundred sentences in your language
- Build a Finite State Transducer that parses words into morphemes.

Review from linguistics class

- Inflectional
- Derivational
- Isolating
- Agglutinating
- Fusional
- Polysynthetic
- Bound morphemes
- Free morphemes
- Prefixes
- Suffixes
- Clitics

What is Linguistic Morphology?

- Morphology is the study of the internal structure of words.
  - **Derivational morphology.** How new words are created from existing words.
    - *[grace]*
    - *[[grace]ful]*
    - *[un[grace]ful]]*
  - **Inflectional morphology.** How features relevant to the syntactic context of a word are marked on that word.
    - This example illustrates number (singular and plural) and tense (present and past).
    - Green indicates irregular.   Blue indicates zero marking of inflection.   Red indicates regular inflection.
    - This student walks.
    - These students walk.
    - These students walked.
  - ***Compounding.*** Creating new words by combining existing words
    - With or without spaces: surfboard, golf ball, blackboard

Morphemes

- **Morphemes.** Minimal pairings of form and meaning.
  - **Roots.** The “core” of a word that carries its basic meaning.
    - *apple* : ‘apple’
    - *walk* : ‘walk’
  - **Affixes** (**prefixes**, **suffixes**, **infixes**, and **circumfixes**). Morphemes that are added to a base (a root or stem) to perform either derivational or inflectional functions.
    - *un-* : ‘NEG’
    - *-s* : ‘PLURAL’

An English Word

- **From David Crystal (Cambridge Encyclopedia of English)**
- Grace (noun): graces
  - Graceful
    - Ungraceful
      - Ungracefully
      - Ungracefulness
    - Gracefully
    - Gracefulness
  - Grace (verb): graces, graced, gracing
  - Disgrace (noun): disgraces
    - Disgraceful
      - Disgracefully
      - Disgracefulness
    - Disgrace (verb): disgraces, disgraced, disgracing
  - Graceless
    - Gracelessly
    - Gracelessness
  - Gracious
    - Graciously
    - Graciousness
    - Ungracious
      - Ungraciously
      - Ungraciousness

Isolating Languages:

Little morphology other than compounding

- Chinese inflection
  - few affixes (prefixes and suffixes):
    - 们： 我们， 你们， 他们， 。。。同志们
      mén: *wǒmén*, *nǐmén*, *tāmén*, *tóngzhìmén*
      plural: we, you (pl.), they comrades, LGBT people
    - “suffixes” that mark aspect: 着 -*zhě* ‘continuous aspect’
- Chinese derivation
- 艺术家 ***yìshù**jiā* **‘artist’**
- Chinese is a champion in the realm of compounding—up to 80% of Chinese words are actually compounds.

毒	+	贩	→	毒贩
dú		*fàn*		*dúfàn*
‘poison, drug’		‘vendor’		‘drug trafficker’

Agglutinative Languages: Swahili

Verbs in Swahili have an average of 4-5 morphemes, http://wals.info/valuesets/22A-swa

Swahili	English
m-tu a-li-lal-a	'The person slept'
m-tu a-ta-lal-a	'The person will sleep'
wa-tu wa-li-lal-a	'The people slept'
wa-tu wa-ta-lal-a	'The people will sleep'

- Words written without hyphens or spaces between morphemes.
- Orange prefixes mark noun class (like gender, except **Swahili** has nine instead of two or three).
  - Verbs agree with nouns in noun class.
  - Adjectives also agree with nouns.
  - Very helpful in parsing.
- Black prefixes indicate tense.

Turkish

Example of extreme agglutination

But most Turkish words have around three morphemes

uygarlaştıramadıklarımızdanmışsınızcasına

“(behaving) as if you are among those whom we were not able to civilize”

uygar	“civilized”
+laş	“become”
+tır	“cause to”
+ama	“not able”
+dık	past participle
+lar	plural
+ımız	first person plural possessive (“our”)
+dan	ablative case (“from/among”)
+mış	past
+sınız	second person plural (“y’ all”)
+casına	finite verb → adverb (“as if”)

Fusional Languages: A New World Spanish

	Singular			Plural		
	1st	2nd	3rd formal 2nd	1st	2nd	3rd
Present	*am-o*	*am-as*	am-a	*am-a-mos*	*am-áis*	*am-an*
Imperfect	*am-ab-a*	*am-ab-as*	*am-ab-a*	*am-áb-a-mos*	*am-ab-ais*	*am-ab-an*
Preterit	*am-é*	*am-aste*	*am-ó*	*am-a-mos*	*am-asteis*	*am-aron*
Future	*am-aré*	*am-arás*	*am-ará*	*am-are-mos*	*am-aréis*	*am-arán*
Conditional	*am-aría*	*am-arías*	*am-aría*	*am-aría-mos*	*am-aríais*	*am-arían*

Polysynthetic Languages

- Polysynthetic morphologies allow the creation of full “sentences” by morphological means.
- They often allow the incorporation of nouns into verbs.
- They may also have affixes that attach to verbs and take the place of nouns.
- **Yupik Eskimo**
  untu-ssur-qatar-ni-ksaite-ngqiggte-uq
  reindeer-hunt-FUT-say-NEG-again-3SG.INDIC
  ‘He had not yet said again that he was going to hunt reindeer.’

Properties of Iñupiac

Arctic Ocean
Pacific Ocean
🟥 Aleut
🟦 Eskimo

- Long, multi-morphemic words
  - Tauqsiġñiaġviŋmuŋniaŋitchugut.
  - ‘We won’t go to the store.’

- Kalaallisut (Greenlandic, Per Langgaard, p.c.)
  - Pittsburghimukarthussaqarnavianngilaq
  - Pittsburgh+PROP+Trim+SG+kar+tuq+ssaq+qar+nav iar+nngit+v+IND+3SG
  - "It is not likely that anyone is going to Pittsburgh"

Mapudungun morphemes → Spanish words

- Mapudungun
  - *treka-lü-la-n*
  - walk-CAUS-NEG-1.sg.IND
  - ‘I didn’t make someone walk’
- Spanish
  - *no hice caminar*
  - not made walk
  - ‘I didn’t make someone walk’

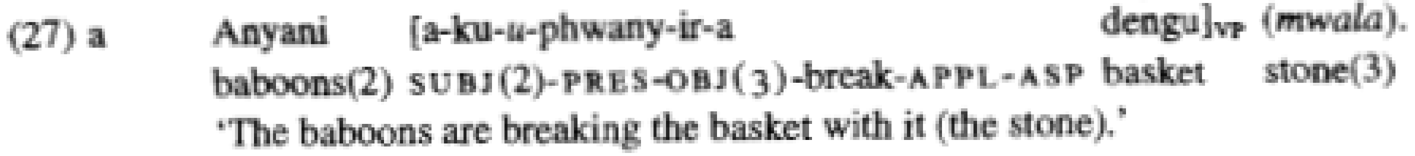
Kofketun → I eat bread

- Mapudungun
  - *iñche kofke-tu-n*
  - I bread-VERB-1.sg.IND
  - ‘I ate bread’

- Spanish
- *yo com-í pan*.

Templatic system

- Chichewa (Bresnan and Mchombo via Kroeger)
- SM-TNS-OM-ROOT-CAUS-APPL-PASS-ASP
  - – (causative and passive not shown in this example)



(27) a	Anyani	[a-ku-*u*-phwany-ir-a	dengu]$_{VP}$	(*mwala*).
	baboons(2)	SUBJ(2)-PRES-OBJ(3)-break-APPL-ASP	basket	stone(3)

‘The baboons are breaking the basket with it (the stone).’

Recursion

- Operationalization
- Oper+ate+ion+al+ize+ate+ion
- Happinesslessnesslessness
- Made Ada make Bertrand make Carl go

Root-and-Pattern Morphology

- **Root-and-pattern**. A special kind of fusional morphology found in Arabic, Hebrew, and their cousins.
- Root usually consists of a sequence of consonants.
- Words are derived and, to some extent, inflected by patterns of vowels intercalated among the root consonants.
  - **kitaab** ‘book’
  - **kaatib** ‘writer; writing’
  - **maktab** ‘office; desk’
  - **maktaba** ‘library’

Other Non-Concatenative Morphological Processes

- **Non-concatenative morphology** involves operations other than the concatenation of affixes with bases.
  - Infixation. A morpheme is inserted inside another morpheme instead of before or after it.
  - Reduplication. Can be prefixing, suffixing, and even infixing.

  - Tagalog:
    - sulat (write, imperative)
    - susulat (reduplication) (write, future)
    - sumulat (infixing) (write, past)
    - sumusulat (infixing and reduplication) (write, present)

  - Internal change (tone change; stress shift; apophony, such as umlaut and ablaut).
  - Root-and-pattern morphology.
  - And more...

Can you make a list of all the words in a language?

Productivity

In the Oxford English Dictionary (OED)

(www.oed.com, accessible for free from CMU machines)

- drinkable
- visitable

Not in the OED

- mous(e)able
- stapl(e)able

In NLP, you need to be able to process words that are not in the dictionary.

But could you make a list of all possible words, taking productivity into account?

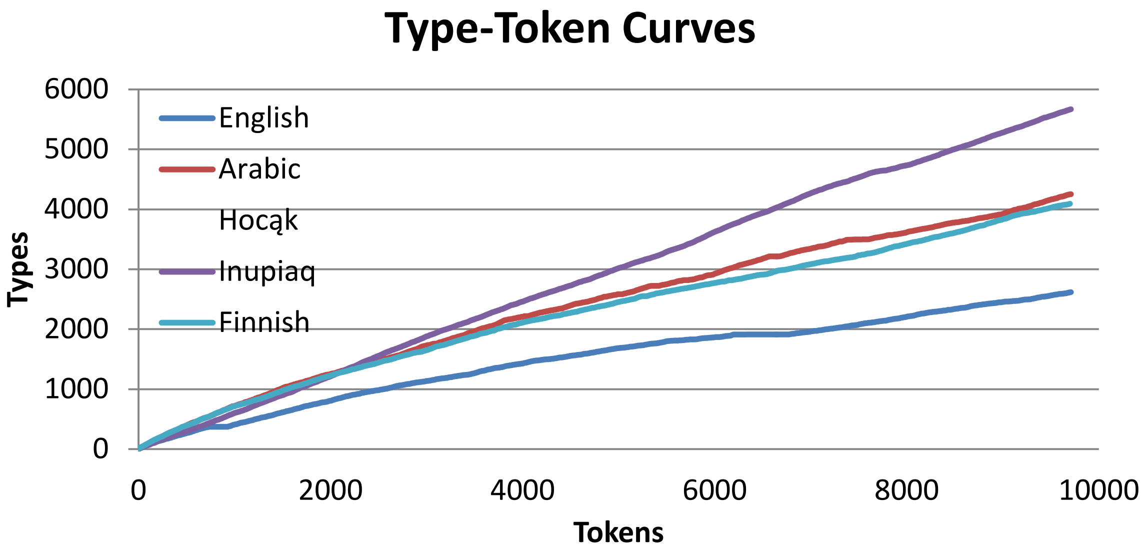
Type-Token Curves

Finnish is agglutinative
Iñupiaq is polysynthetic

Types and Tokens:
“I like to walk. I am walking now. I took a long walk earlier too.”

The type *walk* occurs twice. So there are two tokens of the type *walk*.

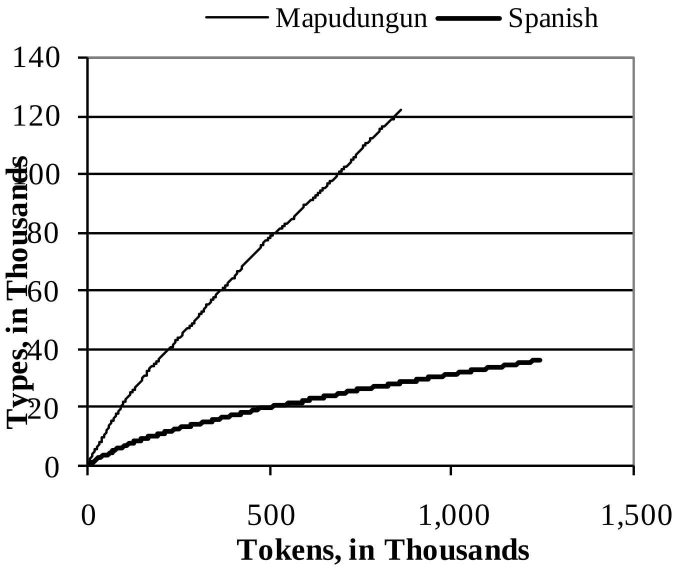
Walking is a different type that occurs once.



Mapudungun compared to Spanish

Mapudungun is polysynthetic

Spanish is fusional



Productivity and compositionality

- Productive morphemes result in words with compositional meanings.
  - The meaning of the word is predictable from the meanings of the parts.
- We will eat around ten-ish.
- She is nice-ish.

Semantic drift

- Via semantic drift, the word takes on a meaning that is more specific than you would predict from the meanings of the parts.
- childish
- boyish
- girlish

Compositionality Alert

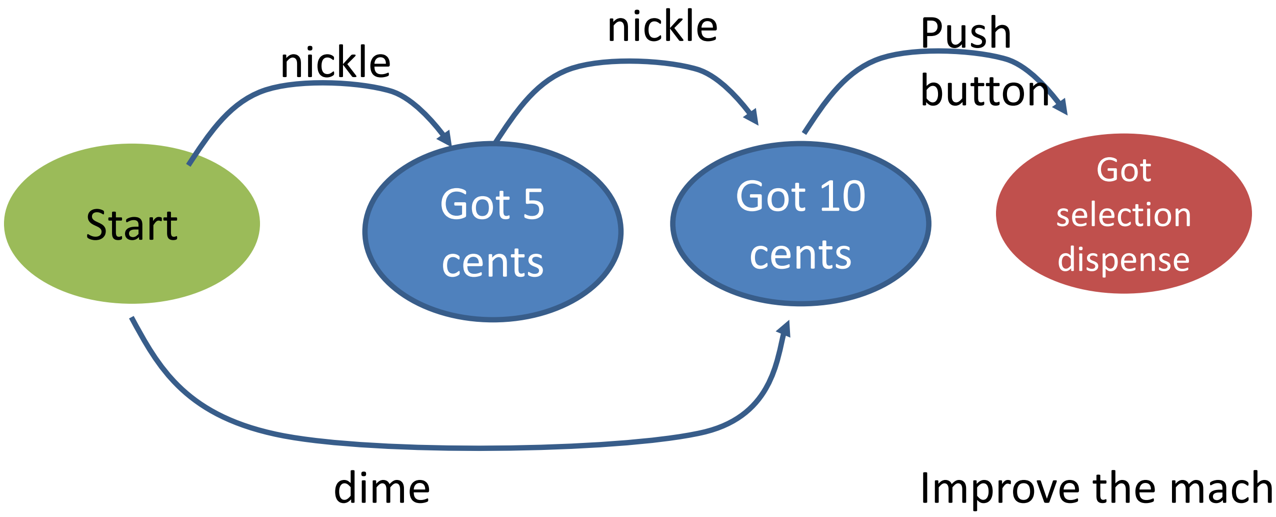
- http://www.newyorker.com/magazine/2012/12/24/utopian-for-beginners
- John Quijada
- Ithkuil language

Compositionality Alert

- “Hunched over the dining-room table, Quijada showed me how he would translate “gawk” into Ithkuil. First, though, since words in Ithkuil are assembled from individual atoms of meaning, he had to engage in some introspection about what exactly he meant to say.
- For fifteen minutes, he flipped backward and forward through his thick spiral-bound manuscript, scratching his head, pondering each of the word’s aspects, as he packed the verb with all of gawking’s many connotations. As he assembled the evolving word from its constituent meanings, he scribbled its pieces on a notepad. He added the “second degree of the affix for expectation of outcome” to suggest an element of surprise that is more than mere unpreparedness but less than outright shock, and the “third degree of the affix for contextual appropriateness” to suggest an element of impropriety that is less than scandalous but more than simply eyebrow-raising. As he rapped his pen against the notepad, he paged through his manuscript in search of the third pattern of the first stem of the root for “shock” to suggest a “non-volitional physiological response,” and then, after several moments of contemplation, he decided that gawking required the use of the “resultative format” to suggest “an event which occurs in conjunction with the conflated sense but is also caused by it.” He eventually emerged with a tiny word that hardly rolled off the tongue:*apq’uxasiu*. He spoke the first clacking syllable aloud a couple of times before deciding that he had the pronunciation right, and then wrote it down in the script he had invented for printed Ithkuil:”

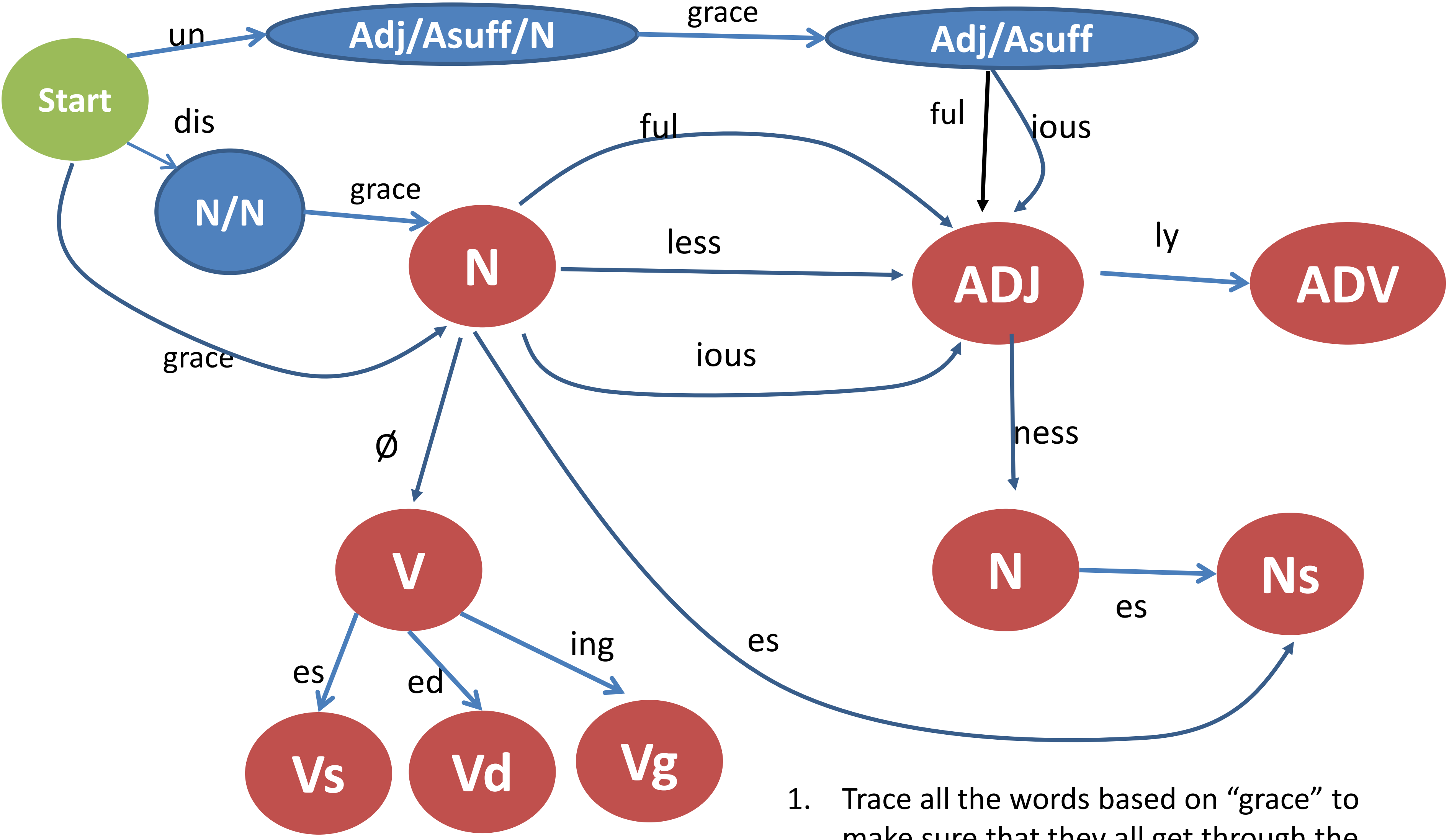
Stop here

Finite State Machine
Consume input; move to next state



Improve the machine?

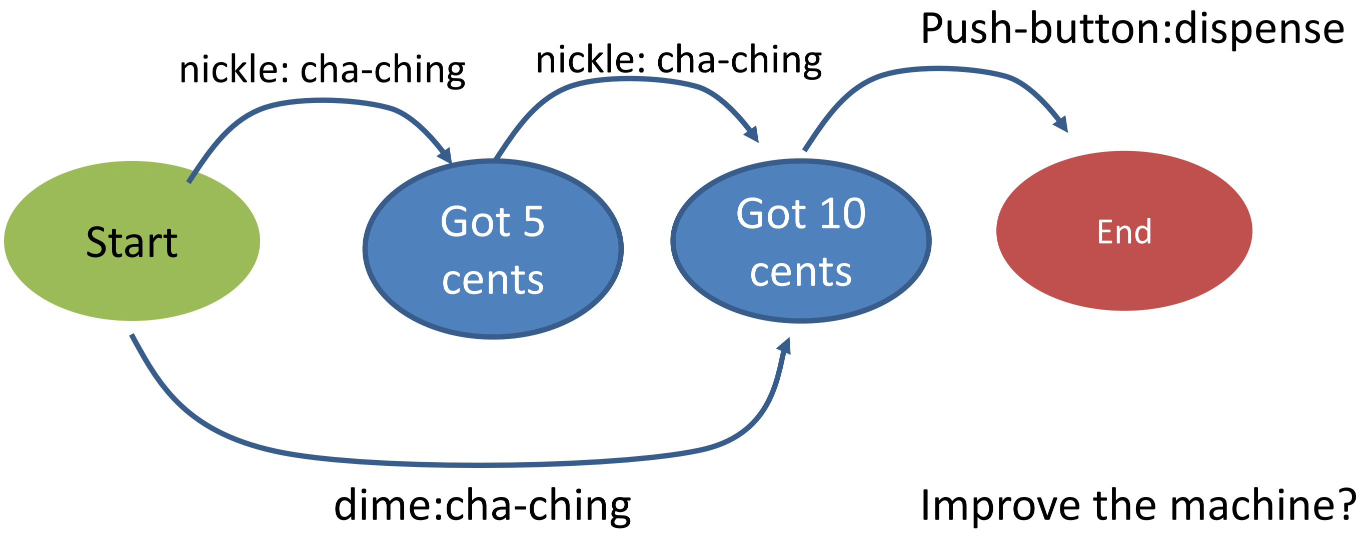
- Eliminate one state (more elegant)
- Add error message if you push the button too soon
- What to do if someone keeps putting money in after the final state.



1. Trace all the words based on “grace” to make sure that they all get through the machine.
2. What is wrong with the machine? (Overgenerate? Undergenerate? Redundant? Doesn’t capture structure and ambiguity?)

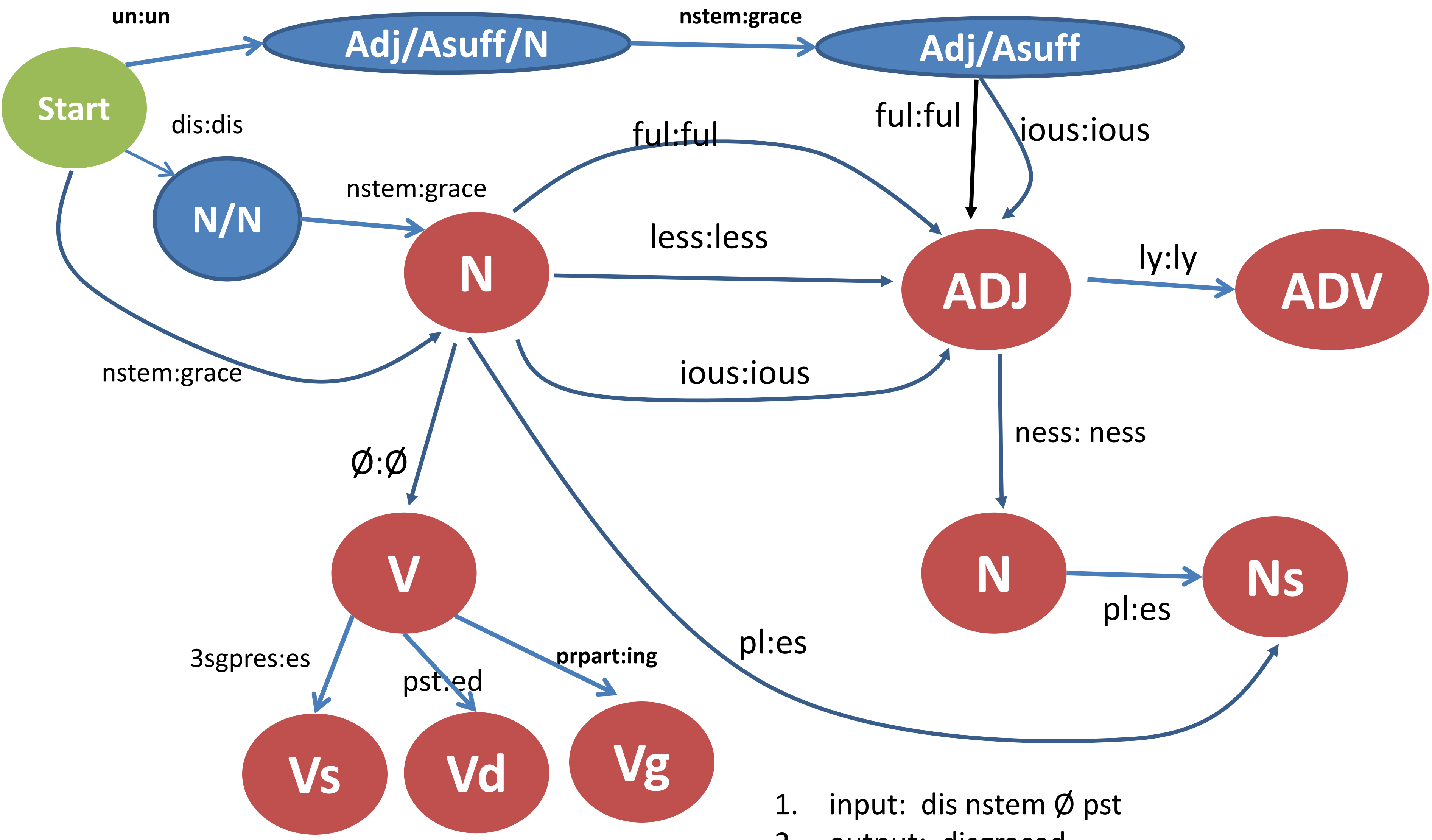
Finite State Transducer

Consume input; produce output: move to next state



Improve the machine?

- Eliminate one state (more elegant)
- Add error message if you push the button too soon
- What to do if someone keeps putting money in after the final state.



1. input: dis nstem Ø pst
2. output: disgraced

What will you do for this assignment?

- You will use a program called XFST (Xerox Finte State Transducer).
- If you give it an underlying form, it will give you the surface form.
- If you give it a surface form, it will give you the underlying form.

- XFST software: http://www.stanford.edu/~laurik/.book2software/
- Tutorial slides by Ken Beesley and Lauri Karttunen: http://www.stanford.edu/~laurik/fsmbook/lecture-notes/Beesley2004/index.html

It will look like code instead of circles and arrows

```
Multichar_Symbols  [NSg] [NPl] [Adj]
[AdjCmpr] [AdjSupr] [NDer]

LEXICON Root
0:0     NRoot ;
0:0     AdjRoot ;

LEXICON AdjRoot
happy[Adj]:happy    # ;
happy:happi          AdjSuffs ;

LEXICON AdjSuffs
[AdjCmpr]:er         # ;
[AdjSupr]:est         # ;
[NDer]:ness          NSuffs-es ;
```

```
LEXICON NRoot
book:book               NSuffs-s ;
pencil:pencil           NSuffs-s ;
child[NSg]:child       #;
child[NPl]:children     #;
rash:rash               NSuffs-es;

LEXICON NSuffs-s
[NSg]:0    # ;
[NPl]:s     # ;

LEXICON NSuffs-es
[NSg]:0    # ;
[NPl]:es   # ;
```

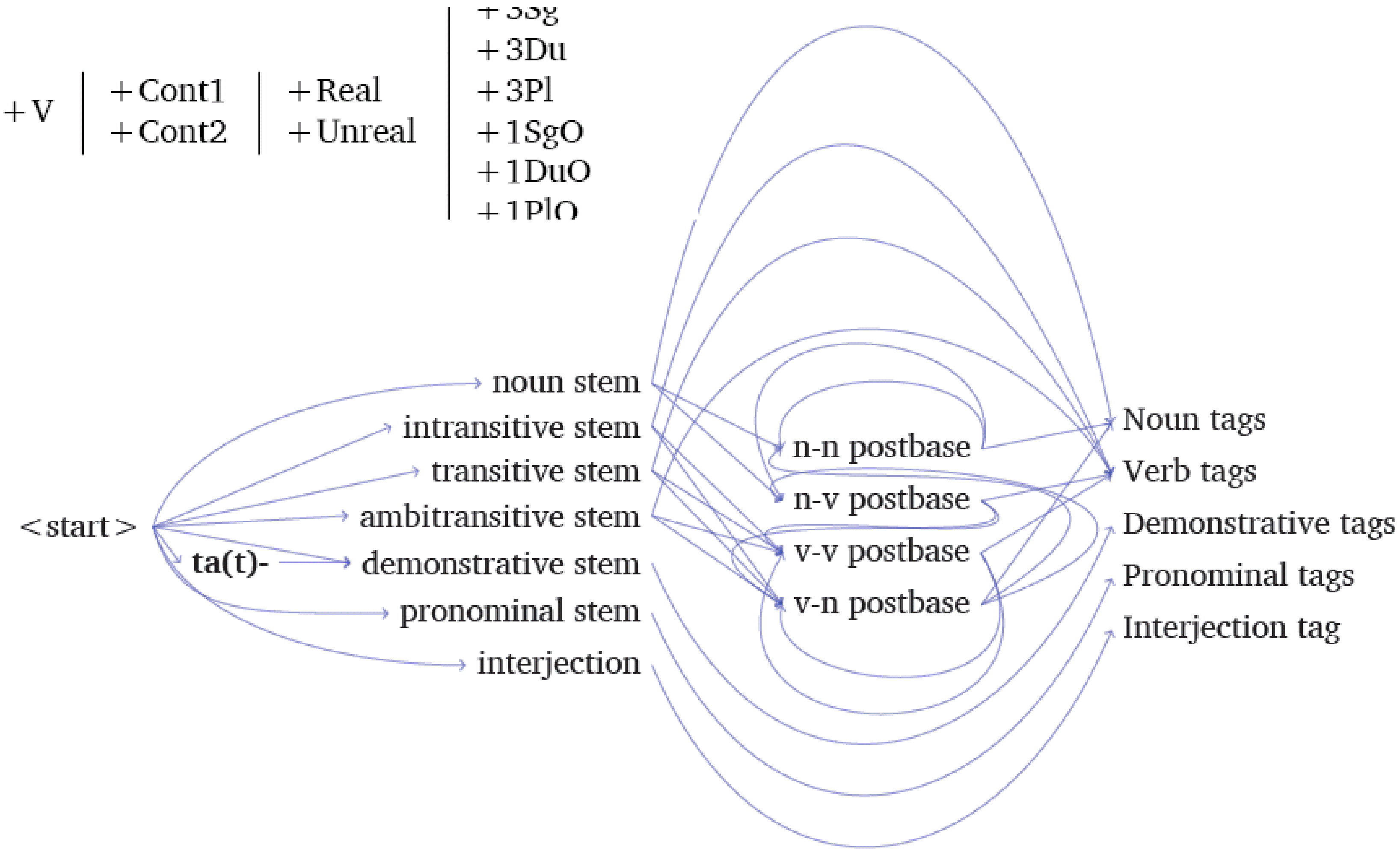
Another way

```
LEXICON NRoot
book;book   NSuffs;
pencil:pencil  NSuffs;
child[NSg]:child  #;
child[NPl]:children #;
rash;rash  NSuffs;

LEXICON NSuffs
[NSg]:0 # ;
[NPl]:s # ;
```

```
s -> e s ||  s h _ .#. ;
```

Morphotactics



Aric Bills, masters thesis, University of Alaska, Fairbanks;
LREC 2010