When *Push* Comes to *Shove*: A Computational Model of the Role of Motor Control in the Acquisition of Action Verbs

David Bailey
University of California, Berkeley

OUTLINE

- An Action-Verb Learning Task
- The Bodily Grounding Perspective
- X-schemas for Action Control
- Linguistic Linking Features
- Learning Verbs by Model Merging
- Results for English, Farsi and Russian

CROSSLINGUISTIC VARIATION IN HAND-ACTION VERBS

- Tamil ***thallu/ilu:*** push/pull, but implies jerkiness or suddenness (smooth motion requires a directional specifier)
- Tamil ***pudi:*** covers clutch, hold, restrain, catch; implies use of high force
- Spanish ***pulsar/presionar***: press with the index finger *vs.* press with the palm
- Farsi ***hol-daadan/feshaar-daadan:*** two senses of pushing: move an object away from body *vs.* apply pressure to an unmoving object
- Farsi ***zadan:*** hit; strum, or play any musical instrument; object manipulation using quick motion

A COMPUTATIONAL TASK

- Given:
  - Training examples which pair an action with a verb
- Learn to:
  - label novel actions
  - obey verbal commands
- *Actions are those of the language learner. Thus, their representation includes goals and motor commands.*
- Data collection and evaluation:
  - Must understand x-schemas

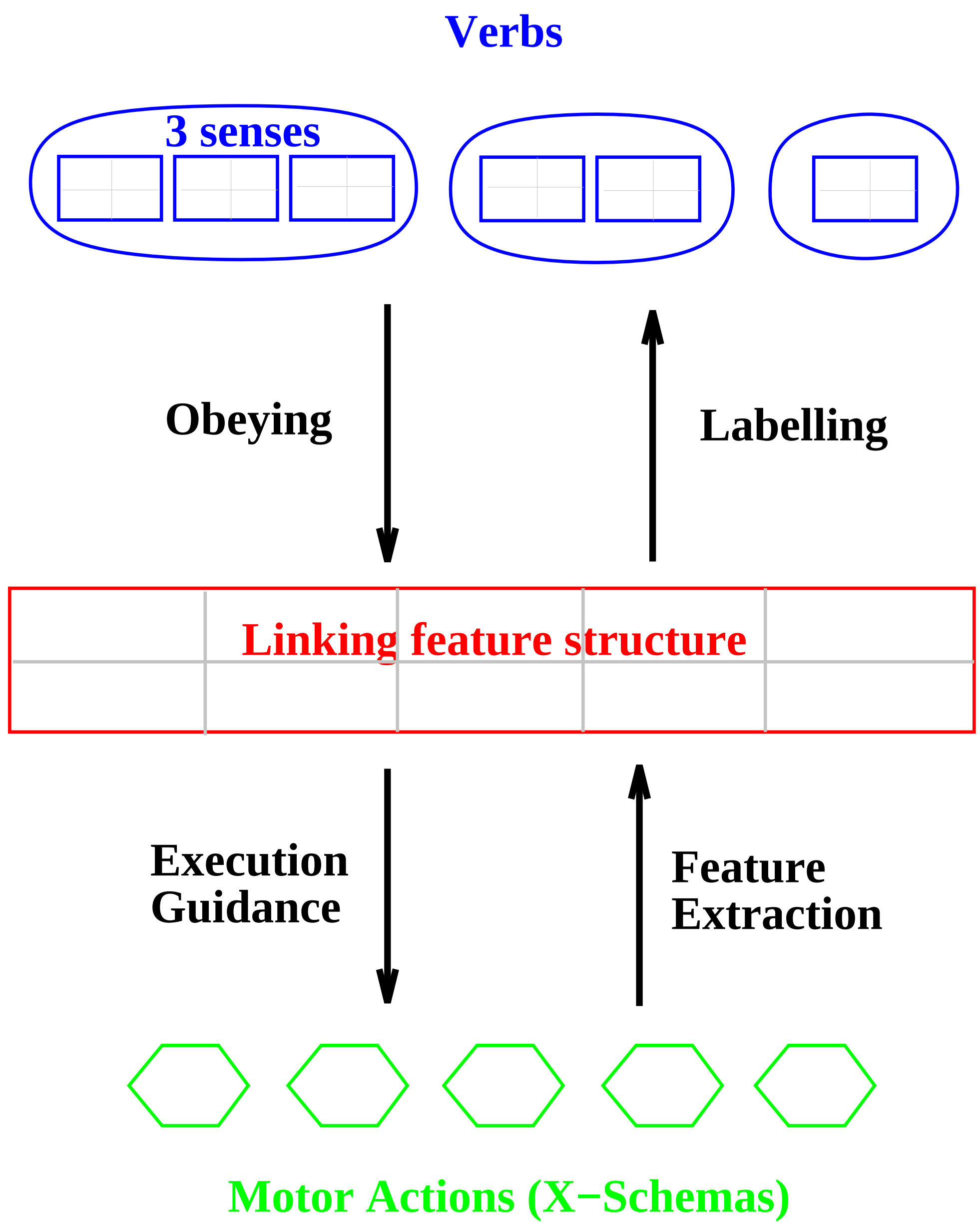
OR

- Use *Jack* animation package:
  - Realistic body and motions
  - Natural interface to x-schemas

BODILY GROUNDING

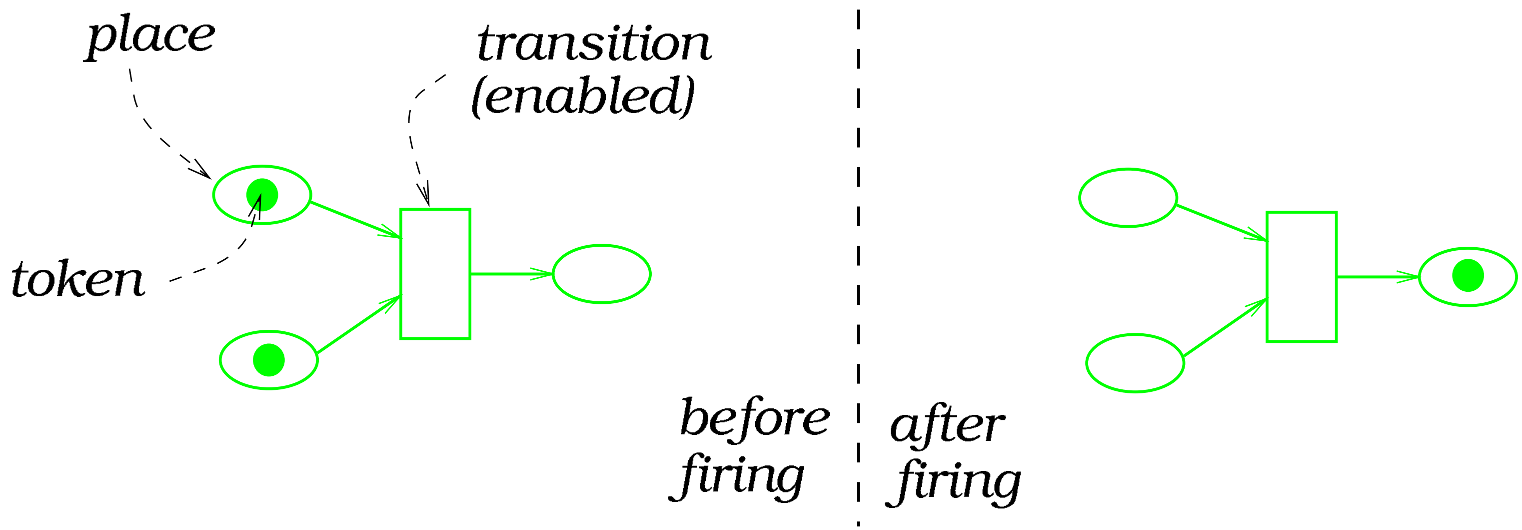
- Computing With Neurons
  - slow
  - simple messages
  - highly parallel
  - no central controller
  - sparse connectivity
- The Human Motor Control System
  - low-level synergies
  - parameterization (e.g. force, direction)
  - high-level coordination
  - concurrency and asynchrony

MODEL ARCHITECTURE

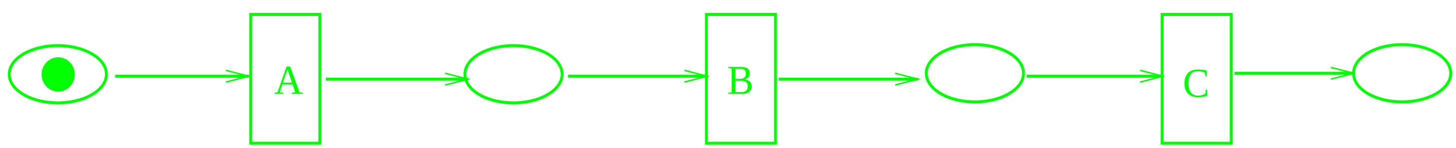


PETRI NETS

BASIC BUILDING BLOCKS

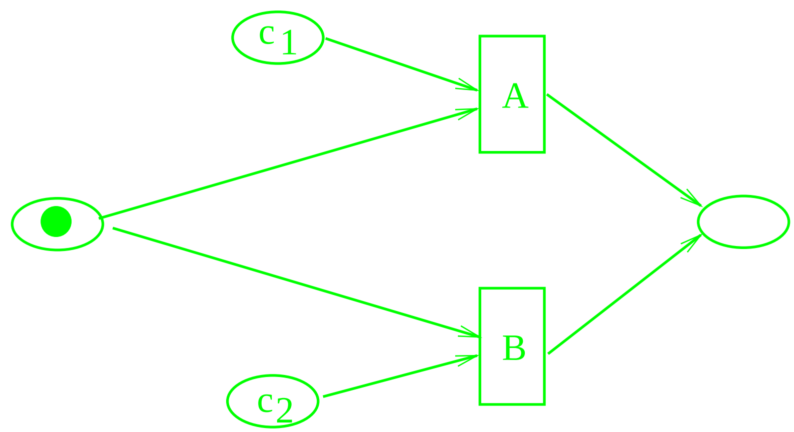


SEQUENTIAL EXECUTION

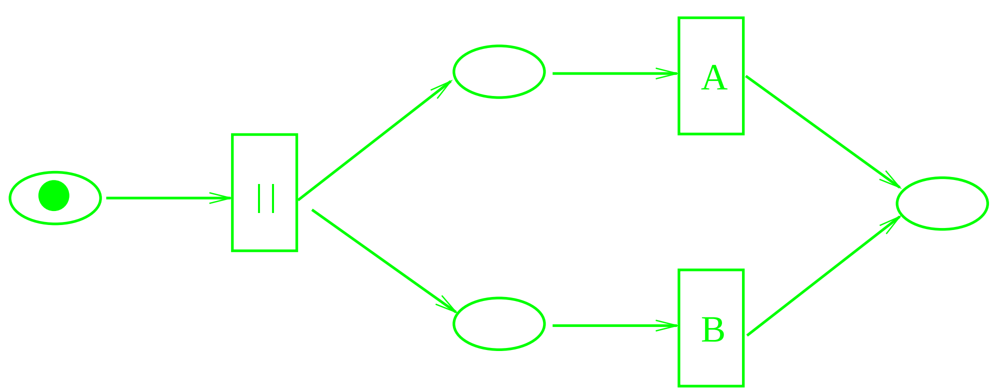


PETRI NETS 2

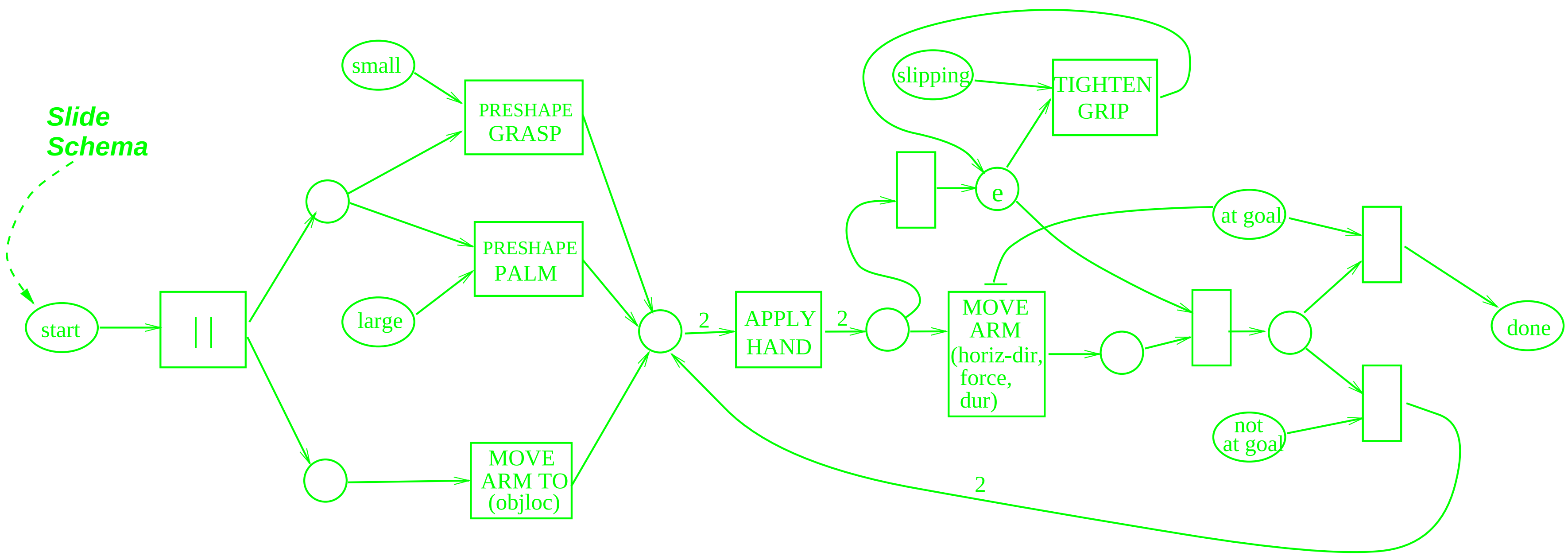
CONDITIONAL EXECUTION



CONCURRENT EXECUTION

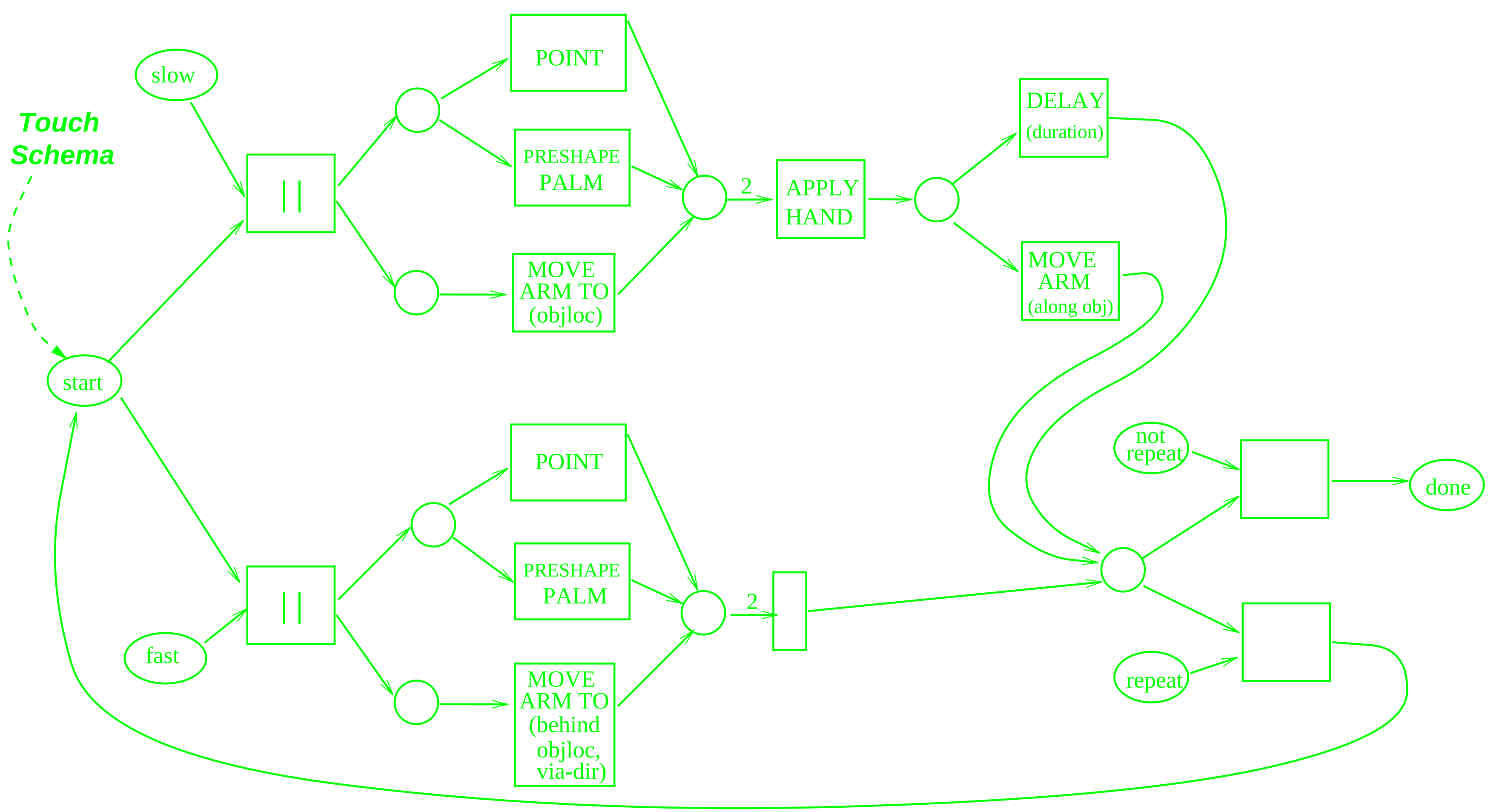


SLIDE X-SCHEMA

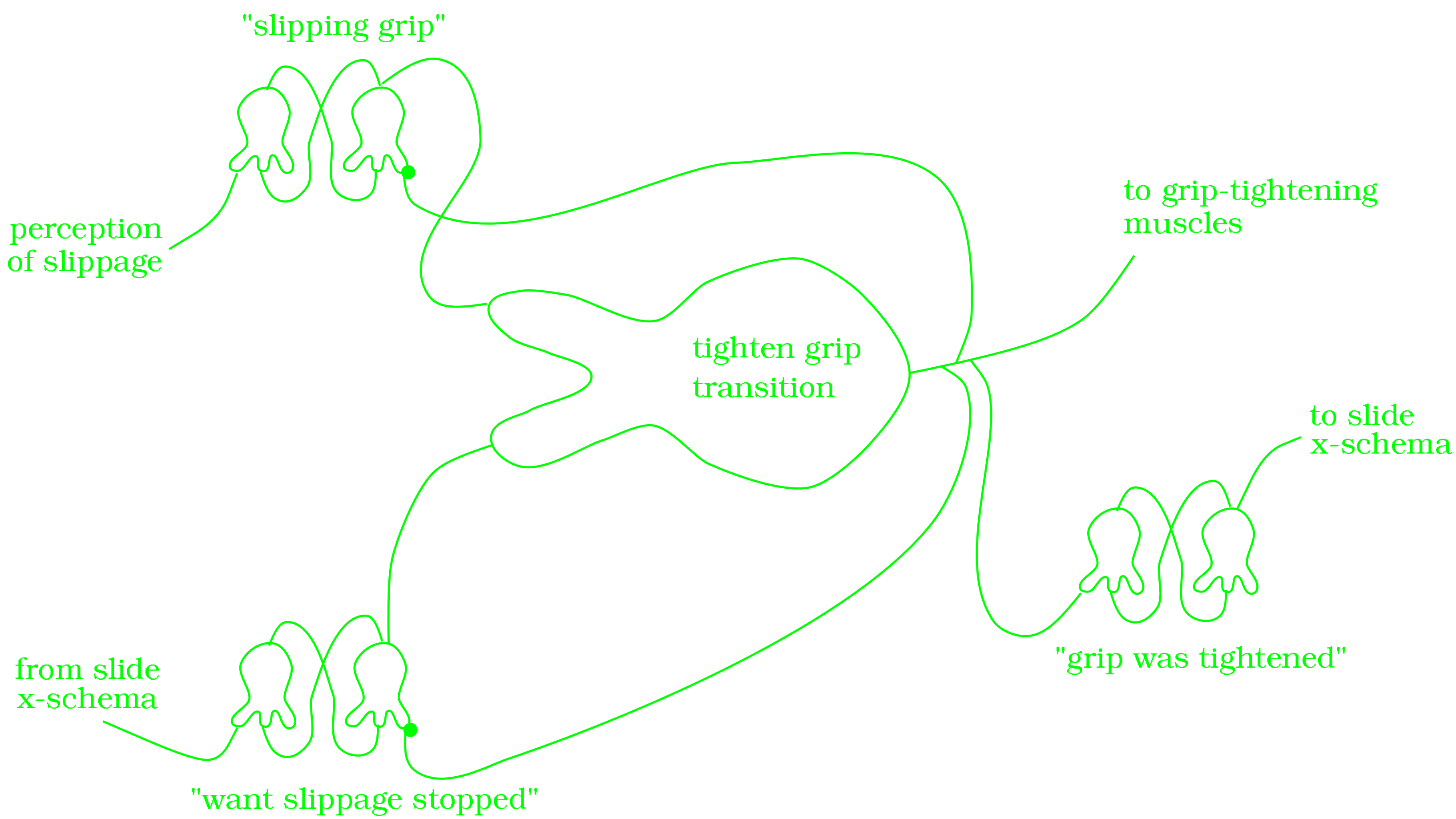


- Other x-schemas: DEPRESS, LIFT, ROTATE, TOUCH

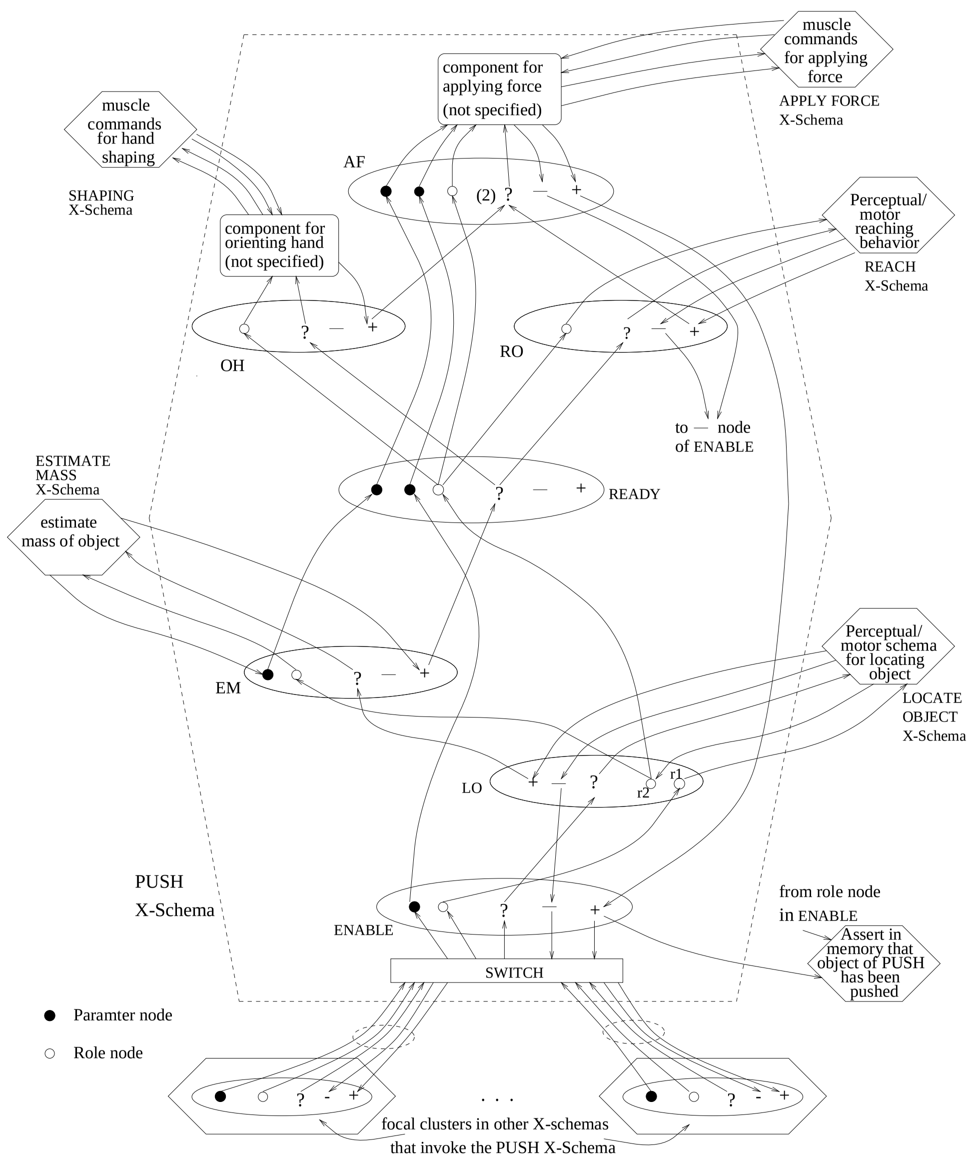
TOUCH X-SCHEMA



CONNECTIONIST ACCOUNT: X-SCHEMAS



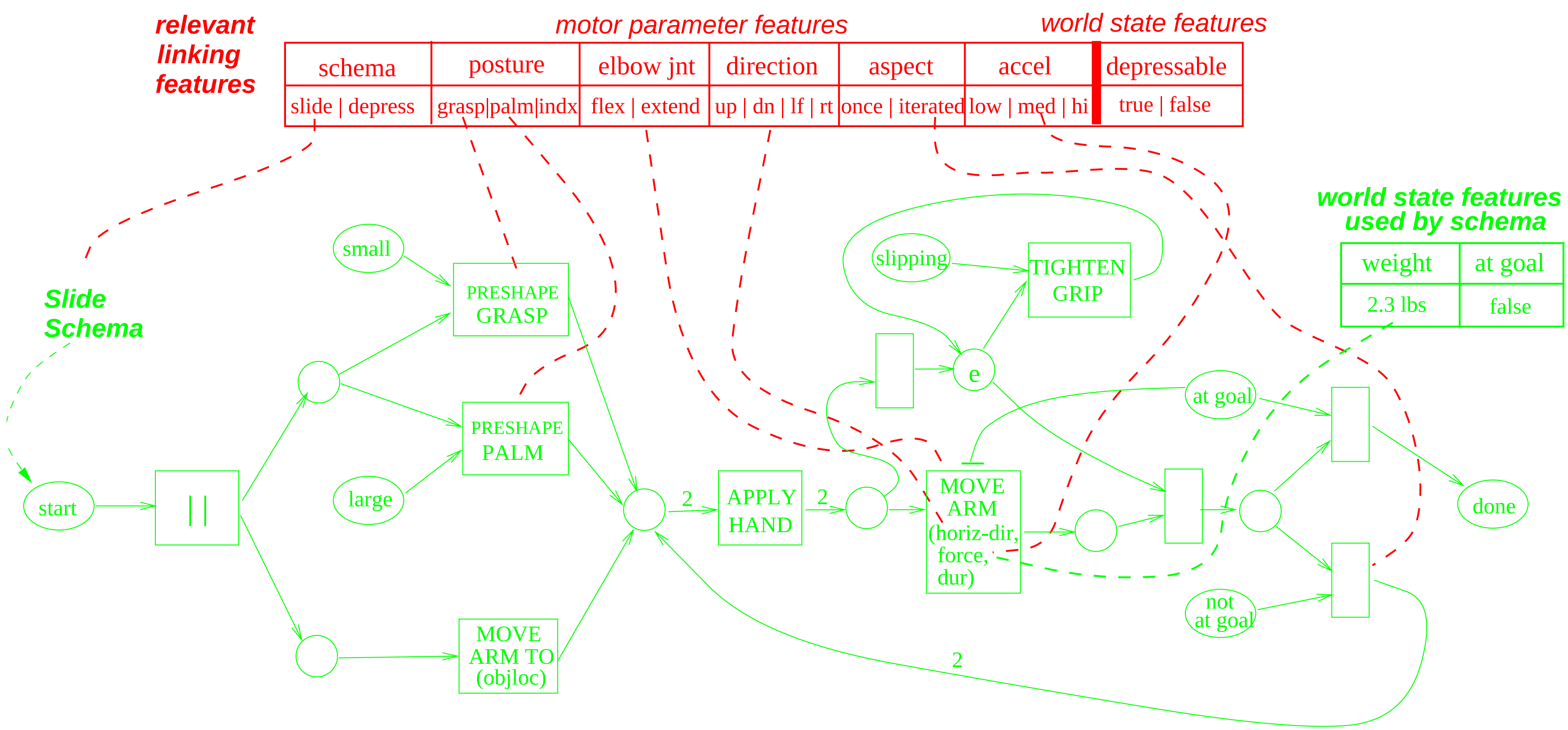
TEMPORAL BINDING OF PARAMETER VALUES



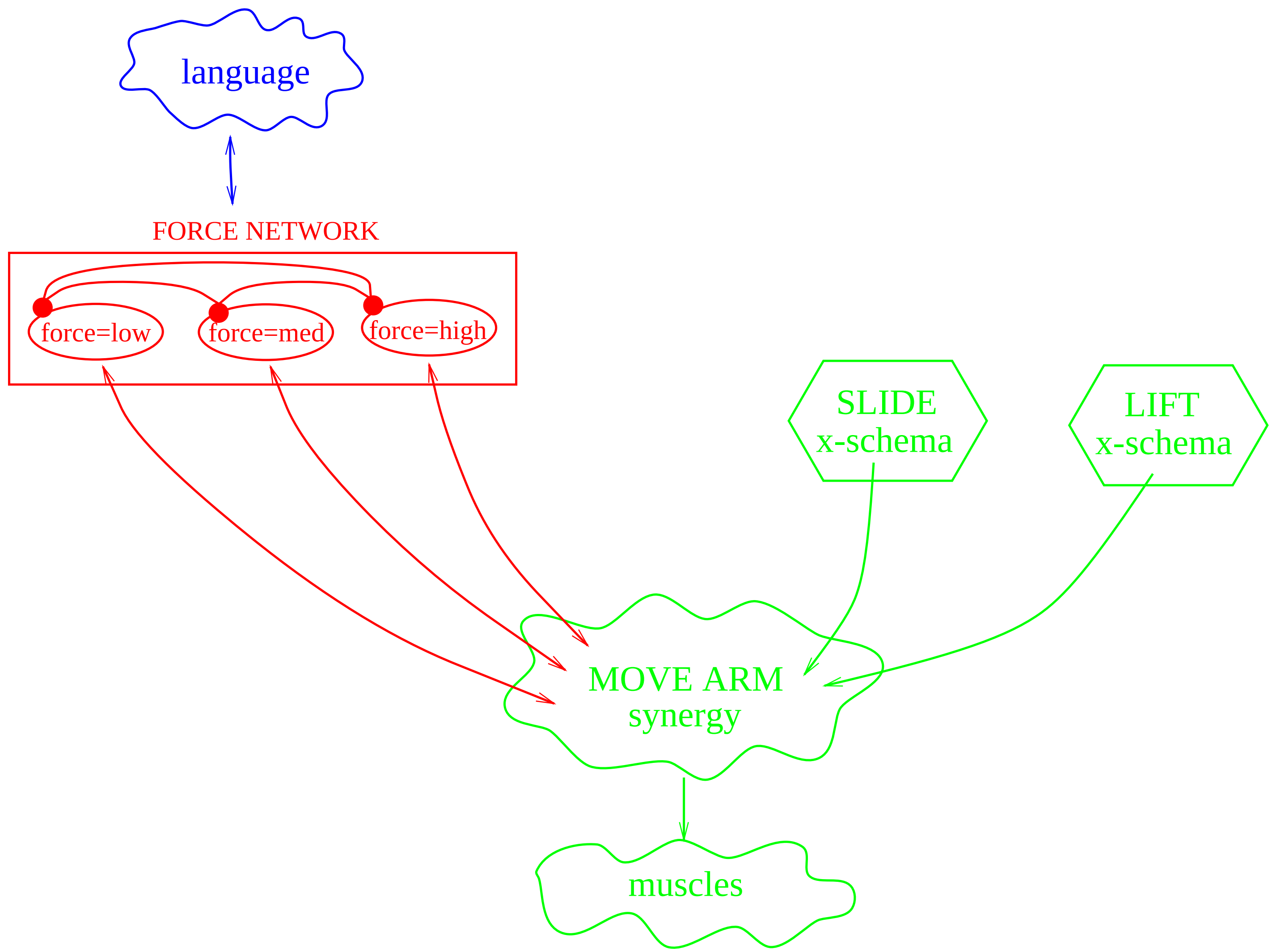
LINKING FEATURES

- Motivation: Extractable and linguistically adequate
- Features include:
    - Schema (which x-schema executes)
    - Hand Posture (grasp, palm, index finger, etc.)
    - Direction (toward, away, up, down, left, right)
    - Elbow Joint Motion (flex, extend, fixed)
    - Force (low, med, high)
    - Aspect (whether x-schema repeats)
    - Object Size (small, med , large)
    - Depressability (a sample object property)
- Extracted from:
    - Synergy parameters
    - Control flow and choice of synergies
    - Perceived world state

CONNECTING LINKING FEATURES TO X-SCHEMAS



CONNECTIONIST ACCOUNT: LINKING-FEATURES



TWO SENSES OF *PUSH*

PUSH: 2 senses

sense 1

schema		posture		direction	
slide	100%	palm	60%	away	50%
touch	0%	grasp	10%	toward	5%
		index	30%	up	15%
				down	30%

commonness 0.2

sense 2

schema		posture		force	
slide	0%	palm	85%	low	10%
touch	100%	grasp	5%	med	30%
		index	10%	high	60%

commonness 0.1

DETAILED VIEW OF MODEL

push

schema	posture	elbow jnt	aspect	depressable
slide 1.0	palm 0.7	extend 0.9	once 0.8	false 0.9

schema	posture	accel	aspect	depressable
depress 1.0	indx 0.9	low 0.7	once 0.6	true 1.0

shove

schema	posture	elbow jnt	accel
slide 1.0	palm 0.9	extend 0.9	high 0.9

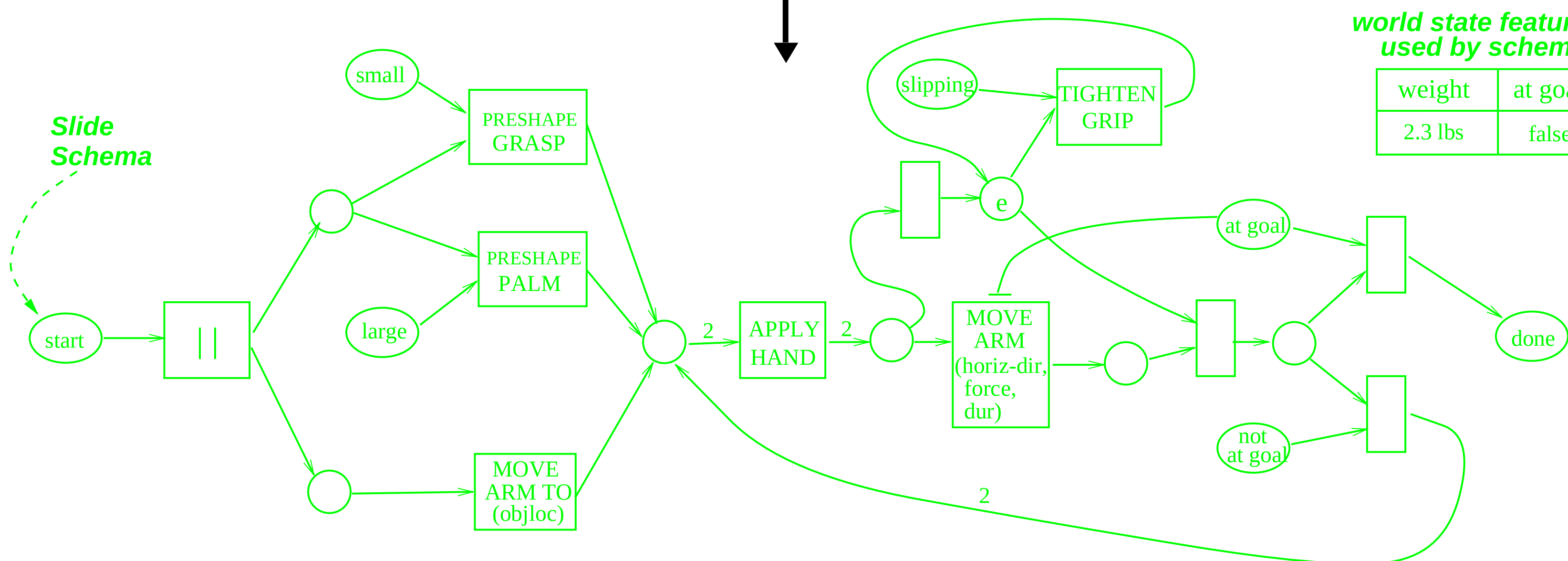
relevant linking features

motor parameter features — *world state features*

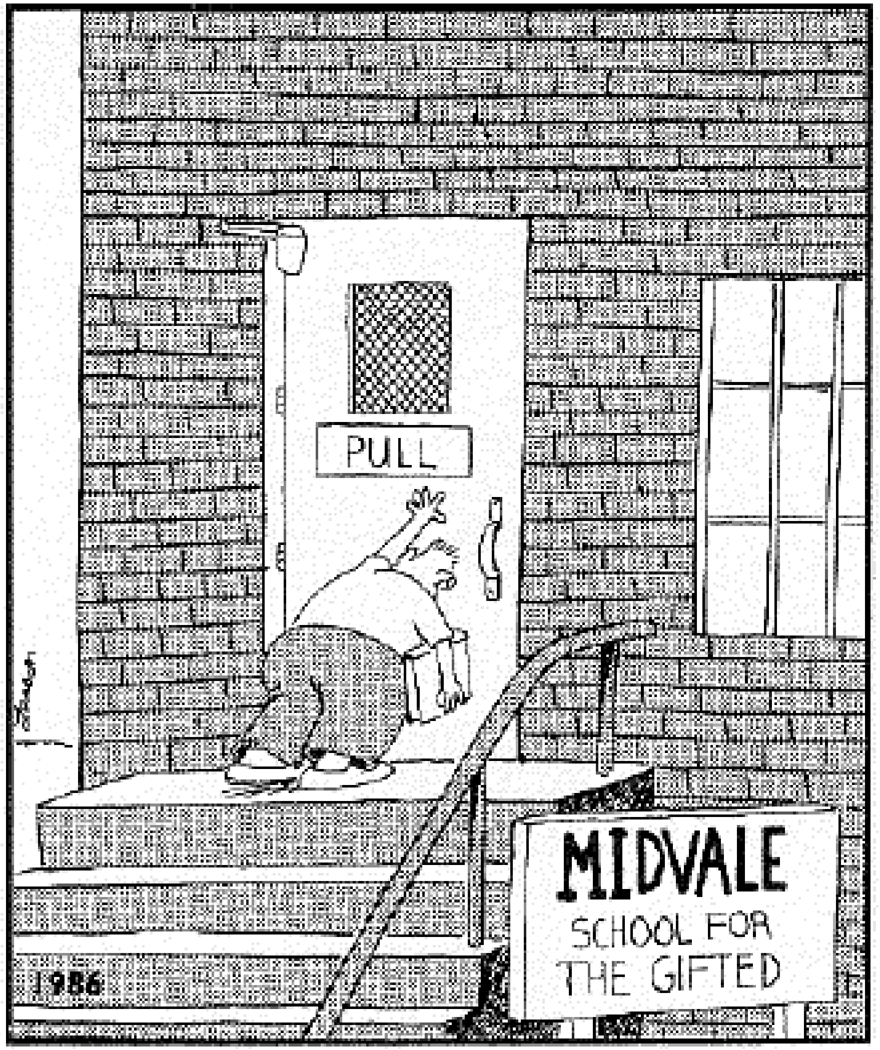
schema	posture	elbow jnt	direction	aspect	accel	depressable
slide | depress	grasp|palm|indx	flex | extend	up | dn | lf | rt	once | iterated	low | med | hi	true | false

world state features used by schema

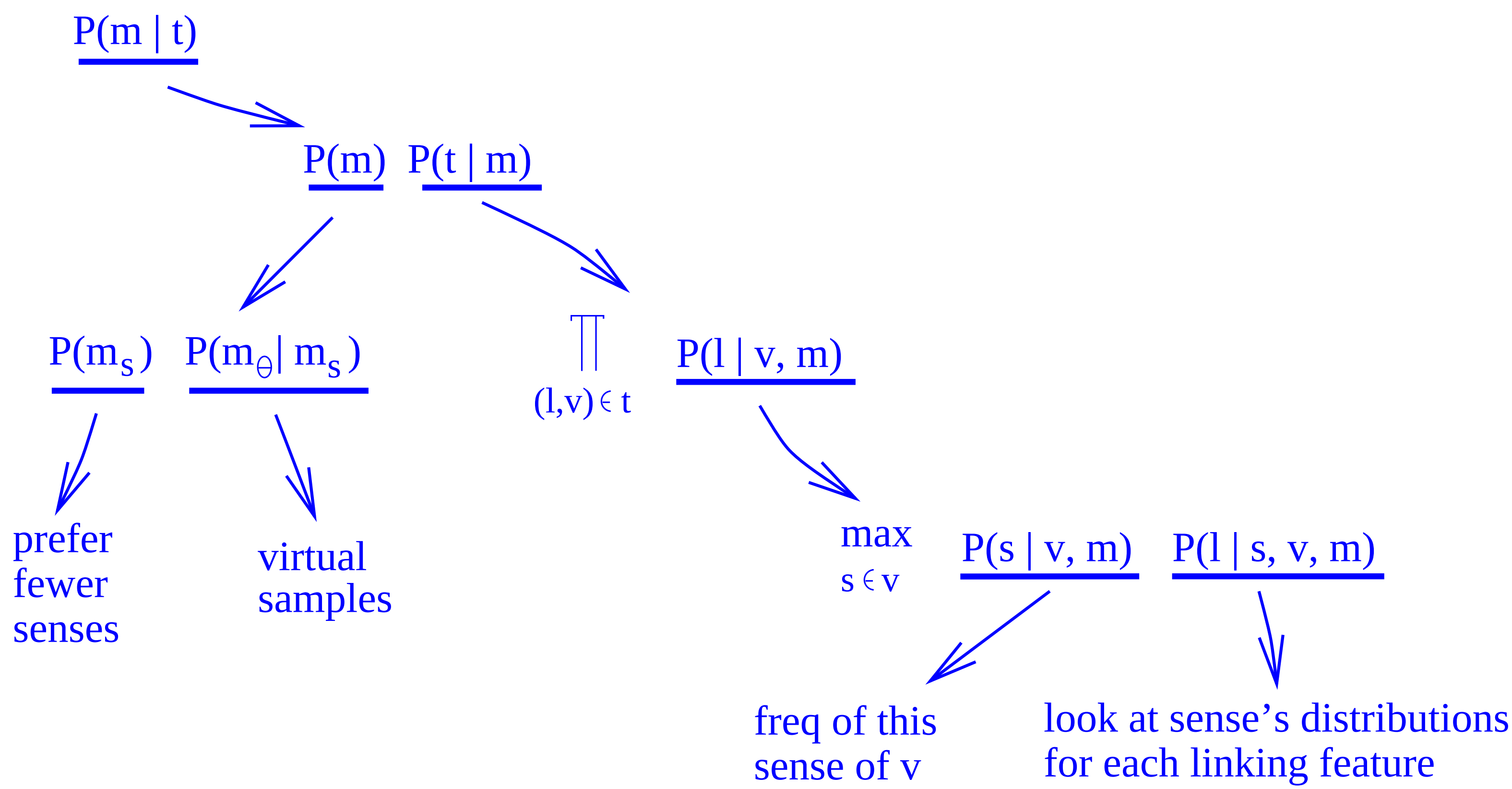
weight	at goal
2.3 lbs	false



HOW TO LEARN VERBS?



RATING A LEXICON MODEL BY ITS POSTERIOR PROBABILITY



Key:
m = model of lexicon
t = training set
v = verb
s = sense of verb
l = linking f-struct

LEARNING WORD SENSES: BAYESIAN MODEL MERGING

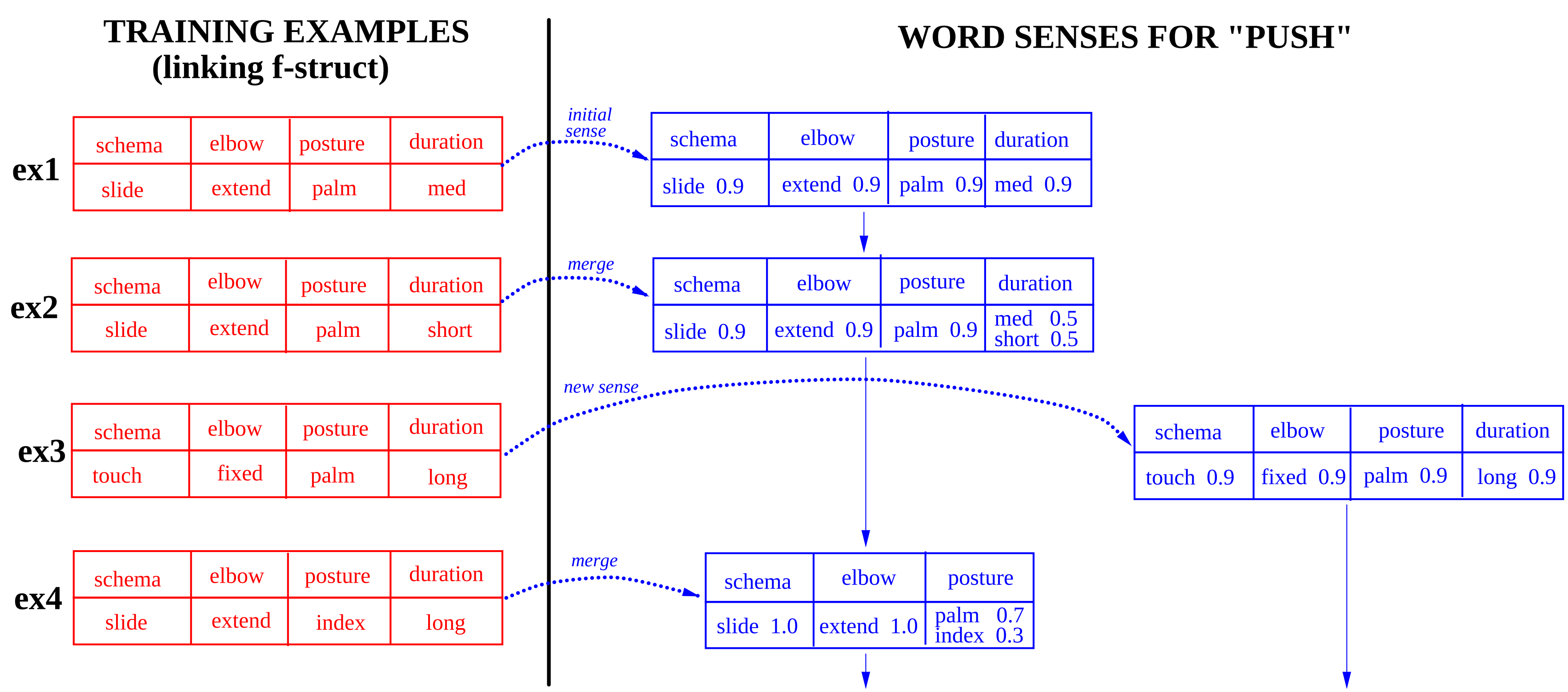
- Basic algorithm:

    1. Create a new word sense for each training ex., generalizing each feature slightly.
    2. While there exist good candidate merge pairs:
        a) Choose best merge pair $\langle sense_1, sense_2 \rangle$.
        b) Replace with a new merged $sense_{12}$.

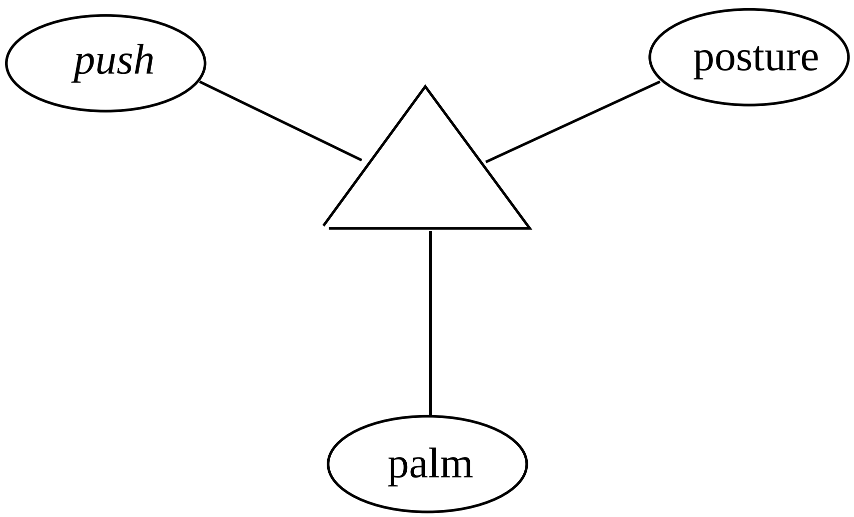
- Merging involves combining probability distributions to form (usually) more general word senses.

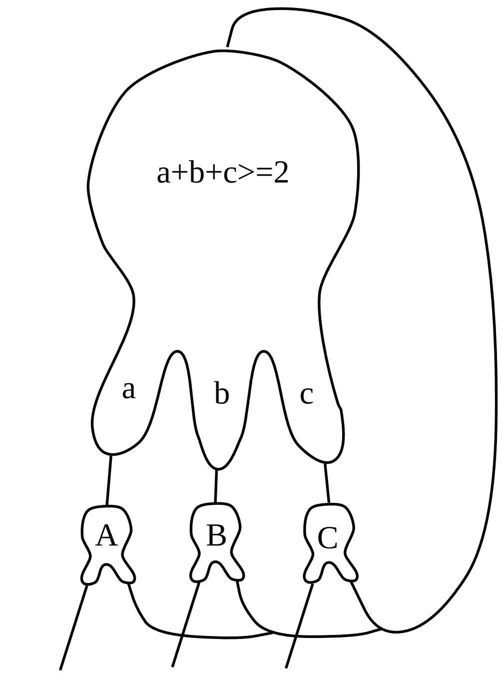
- Merge choice and stopping criterion use Bayes' Rule to trade off fit to data *vs.* desire for "simpler" models.

A LEARNING ILLUSTRATION

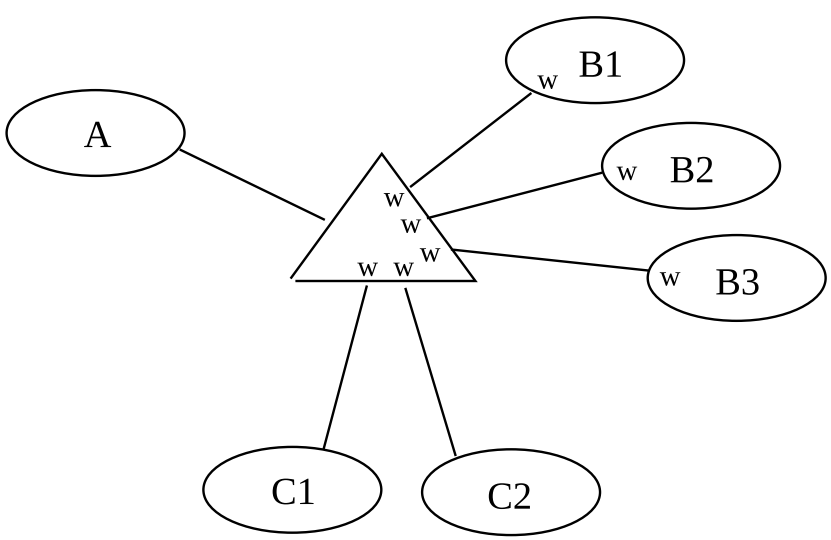


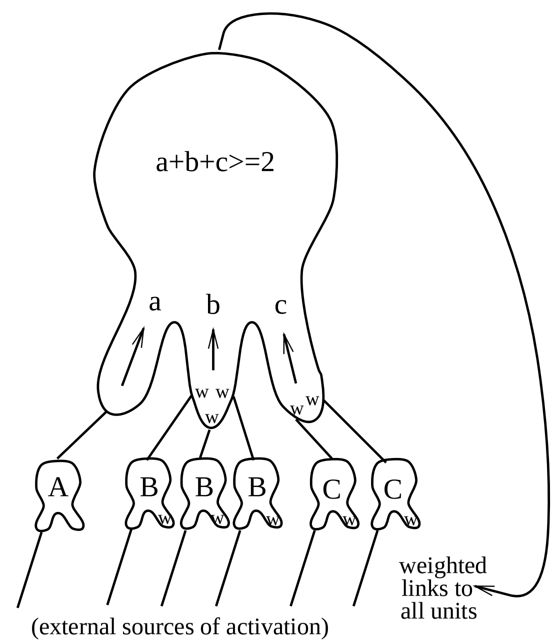
TRIANGLE NODES



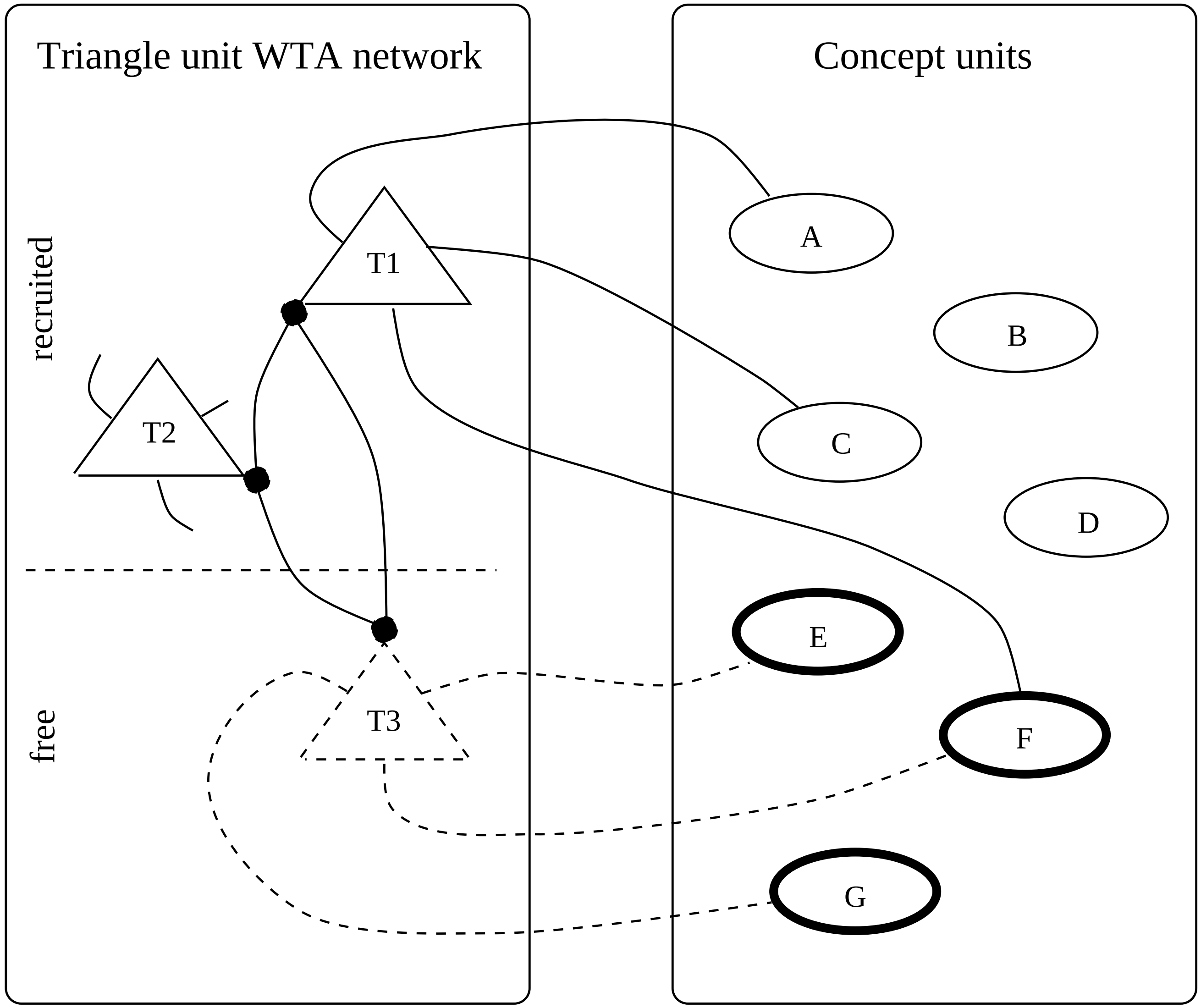


COMPLEX TRIANGLE NODES

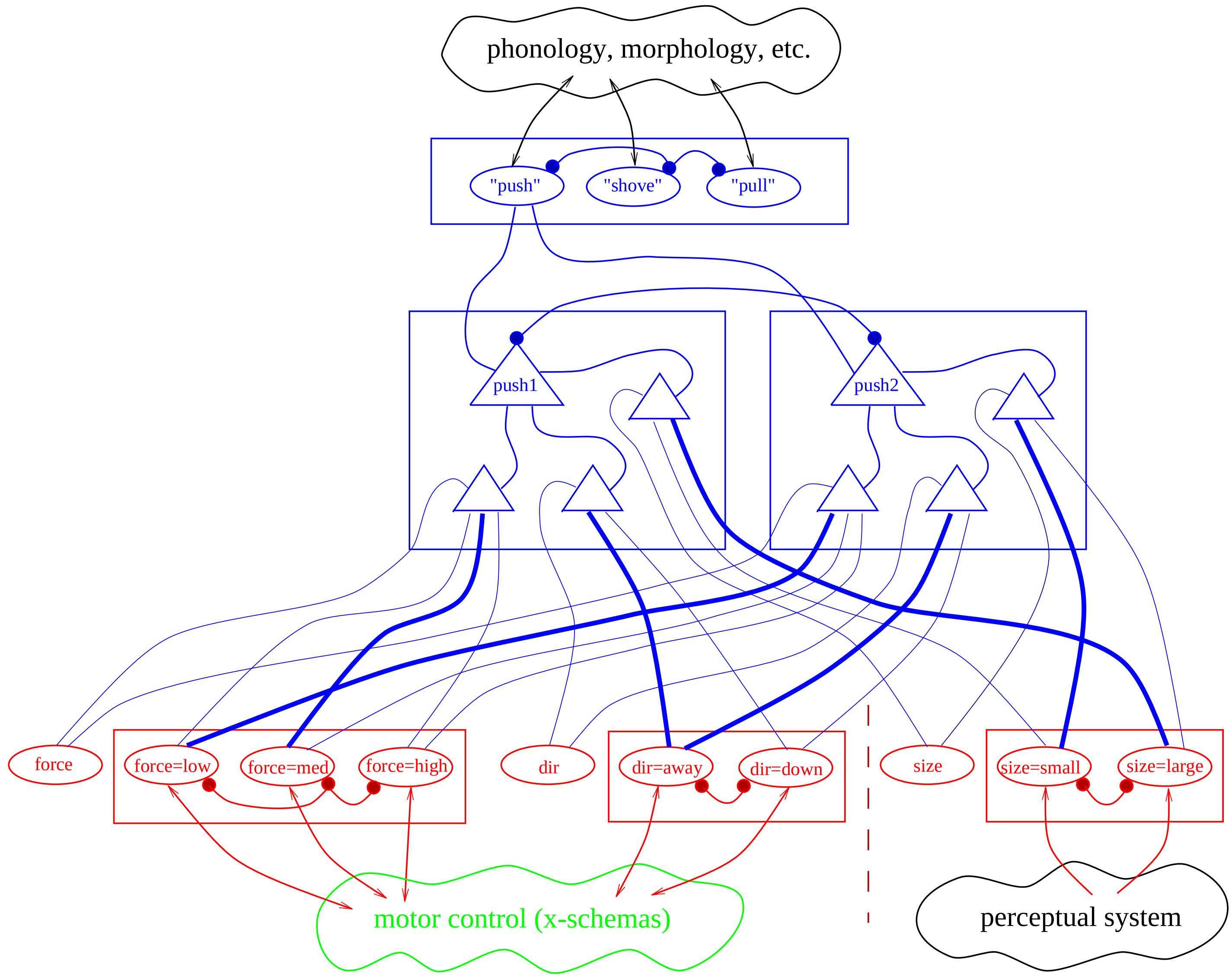




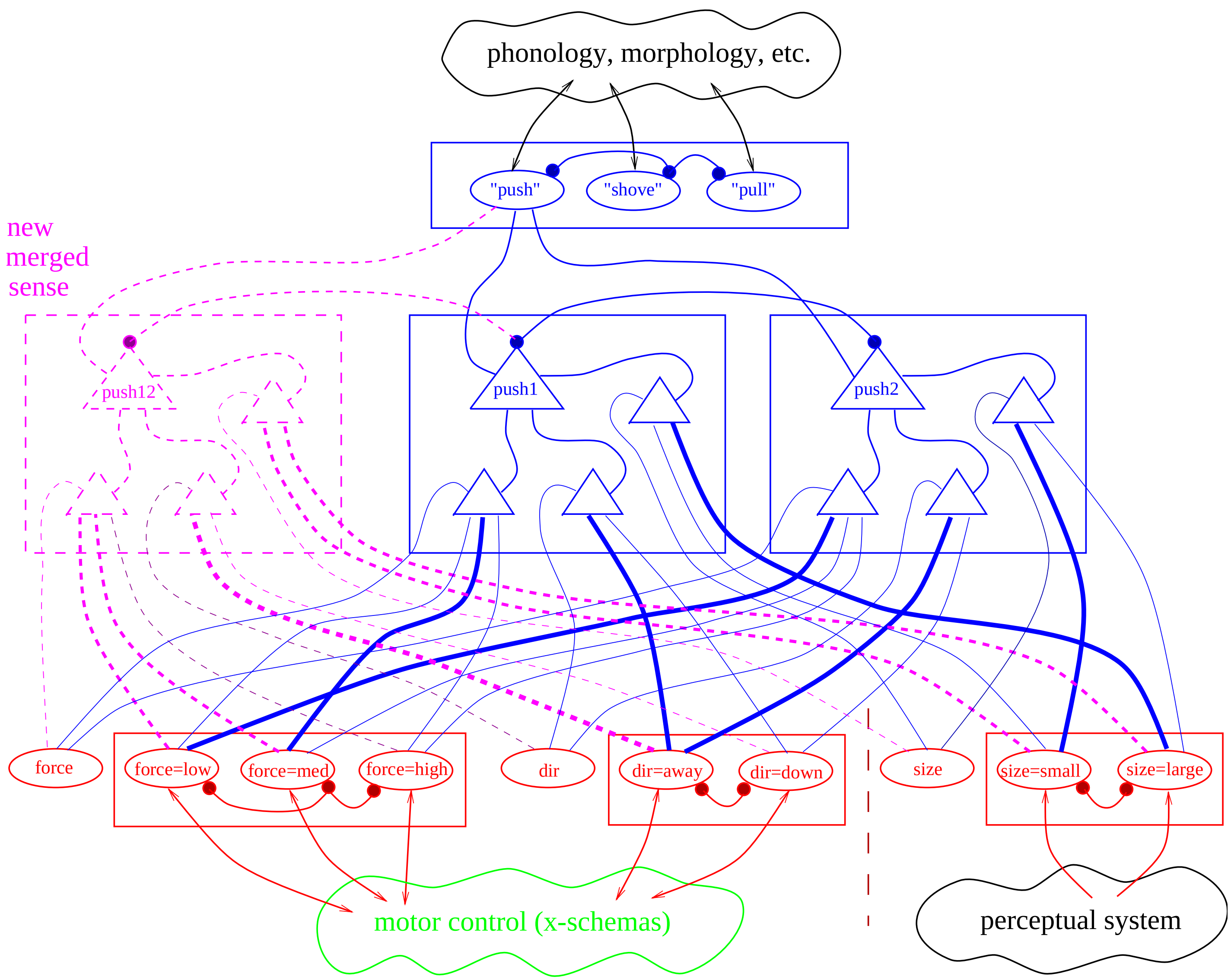
RECRUITMENT LEARNING



CONNECTIONIST ACCOUNT: VERB REPRESENTATION



CONNECTIONIST ACCOUNT: MERGING WORD SENSES



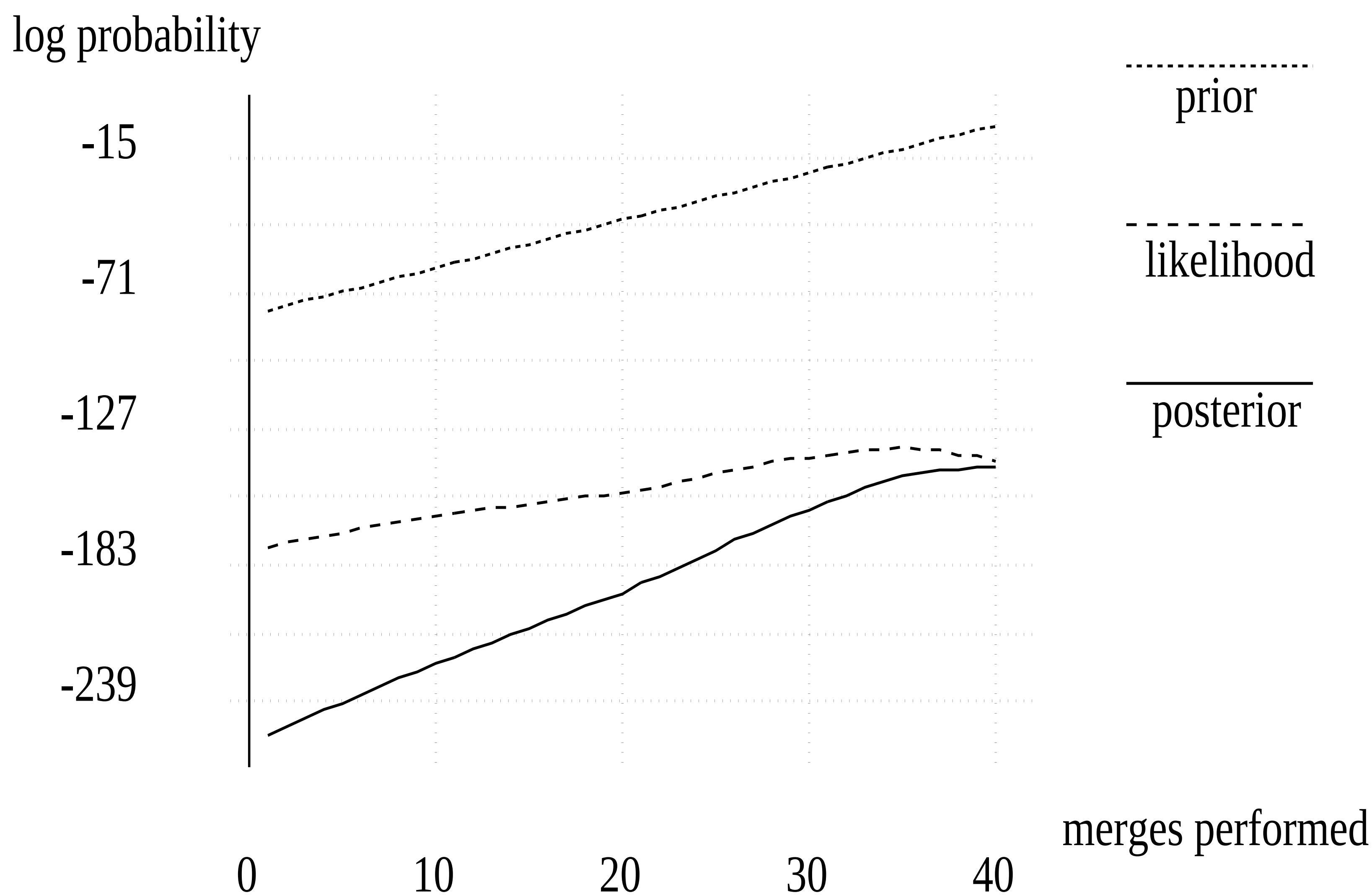
TRAINING RESULTS: ENGLISH

- 165 training examples; verb labels are:

push	pull	shove	yank
slide	press	touch	feel
slap	hit	tap	poke
lift	pick-up	heave	hold
turn	roll		

- Finds optimal number of word senses (21)
- 32 test examples: 78% recognition rate
  81% obeying rate
- Obeying *push*: chooses SLIDE or DEPRESS action as appropriate given the object type

TRAJECTORY OF LEARNING



LEARNED SENSES OF *PUSH*

```
  push74 (12 ex) {
    size {SMALL=0.785 large=0.214}
    elongated {true=0.071 FALSE=0.928}
    depressible {TRUE=0.928 false=0.071}
    contact {true=0.071 FALSE=0.928}
*   schema {slide=0.004 lift=0.004 rotate=0.004
              DEPRESS=0.983 touch=0.004}
    posture {grasp=0.055 wrap=0.055 pinch=0.055 palm=0.055
              platform=0.055 INDEX=0.722}
    elbow {flex=0.333 extend=0.333 fixed=0.333}
    force {low=0.466 med=0.2 high=0.333}
    accel {zero=0.062 low=0.312 MED=0.5 high=0.125}
    dir {away=0.166 toward=0.166 up=0.166 down=0.166
         left=0.166 right=0.166}
    aspect {once=0.5 iterated=0.5}
    dur {SHORT=0.6 med=0.266 long=0.133}
  }

  push79 (8 ex) {
    size {small=0.4 large=0.6}
    elongated {true=0.3 FALSE=0.7}
    depressible {true=0.2 FALSE=0.8}
    contact {TRUE=0.7 false=0.3}
*   schema {slide=0.369 lift=0.006 rotate=0.006
              depress=0.127 touch=0.490}
    posture {grasp=0.142 wrap=0.071 pinch=0.071 PALM=0.5
              platform=0.071 index=0.142}
    elbow {flex=0.1 extend=0.5 fixed=0.4}
    force {low=0.181 MED=0.545 high=0.272}
    accel {zero=0.166 low=0.333 med=0.333 high=0.166}
    dir {AWAY=0.384 toward=0.076 up=0.076 down=0.076
         left=0.230 right=0.153}
    aspect {ONCE=0.9 iterated=0.1}
    dur {short=0.363 med=0.090 long=0.545}
  }

  push78 (21 ex) {
    size {small=0.565 large=0.434}
    elongated {TRUE=0.608 false=0.391}
    depressible {true=0.043 FALSE=0.956}
    contact {true=0.086 FALSE=0.913}
*   schema {SLIDE=0.990 lift=0.002 rotate=0.002
              depress=0.002 touch=0.002}
    posture {grasp=0.481 wrap=0.037 pinch=0.037 palm=0.370
              platform=0.037 index=0.037}
    elbow {flex=0.041 EXTEND=0.916 fixed=0.041}
    force {low=0.333 med=0.416 high=0.25}
    accel {zero=0.04 low=0.32 MED=0.48 high=0.16}
    dir {AWAY=0.518 toward=0.037 up=0.037 down=0.037
         left=0.222 right=0.148}
    aspect {ONCE=0.739 iterated=0.260}
    dur {short=0.333 med=0.333 long=0.333}
  }
```

RECOGNITION PERFORMANCE

```
Beginning recognition test...
  Scenario sc6: desired=push, output=push
  Scenario sc12: desired=push, output=push
  Scenario sc18: desired=pickup, output=pickup
  Scenario sc24: desired=feel, output=feel
  Scenario sc30: desired=shove, output=shove
  Scenario sc36: desired=heave, output=lift      *ERROR*
  Scenario sc42: desired=hold, output=hold
  Scenario sc48: desired=slide, output=push    *ERROR*
  Scenario sc60: desired=slap, output=slap
  Scenario sc66: desired=pull, output=pull
  Scenario sc72: desired=turn, output=turn
  Scenario sc78: desired=pull, output=pull
  Scenario sc84: desired=push, output=push
  Scenario sc90: desired=press, output=press
  Scenario sc96: desired=turn, output=turn
  Scenario sc102: desired=pickup, output=pickup
  Scenario sc108: desired=pull, output=pull
  Scenario sc114: desired=lift, output=lift
  Scenario sc120: desired=hold, output=hold
  Scenario sc126: desired=poke, output=poke
  Scenario sc132: desired=pull, output=pull
  Scenario sc138: desired=pull, output=pull
  Scenario sc144: desired=push, output=push
  Scenario sc150: desired=turn, output=turn
  Scenario sc156: desired=tap, output=touch    *ERROR*
  Scenario sc162: desired=slide, output=slide
  Scenario sc168: desired=yank, output=pull    *ERROR*
  Scenario sc174: desired=press, output=push *ERROR*
  Scenario sc180: desired=push, output=push
  Scenario sc186: desired=heave, output=lift    *ERROR*
  Scenario sc192: desired=press, output=push *ERROR*
  Scenario sc198: desired=push, output=push
  Correctly labelled 25 of 32 test scenarios (78%).
Recognition test done.
```

- Overall success rate is 78%
- Most errors are “close calls,” due to frequency effects (common words overwhelm specific words)

COMMAND-OBEYING PERFORMANCE

- If the command is *push*:

  and the initial world state is:

  {size=small elongated=false depressible=true contact=false}

  then these linking features are set:

  {schema=depress posture=index accel=med dur=short}

  but if the initial world state is:

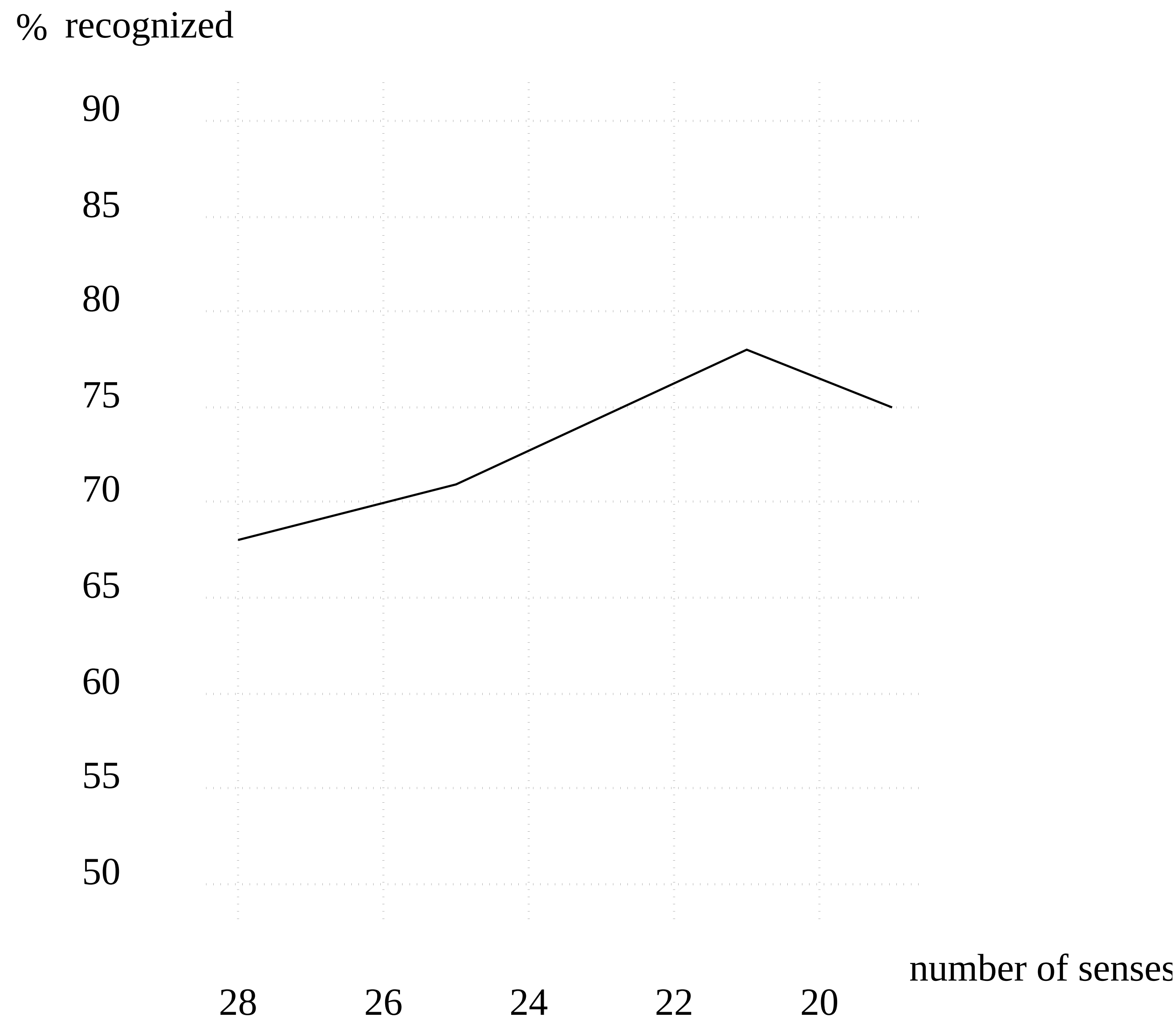
  {size=large elongated=false depressible=false contact=false}

  then these linking features are set:

  {schema=slide elbow=extend accel=med dir=away aspect=once}

- Overall command-obeying performance (judged by subsequent labelling of the action) is 81%.

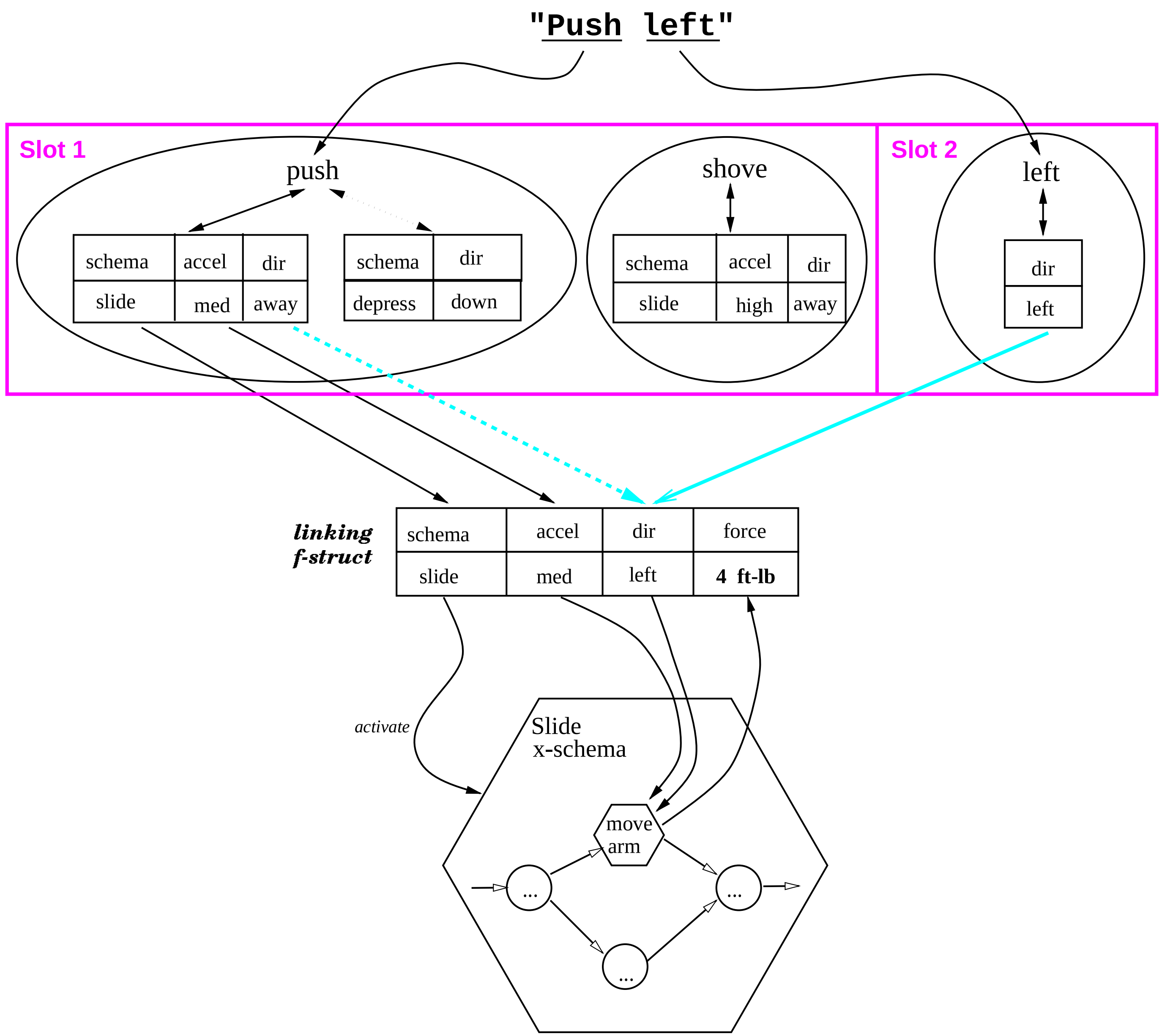
RECOGNITION RATE VS. NUMBER OF WORD SENSES



TRAINING RESULTS: FARSI

- Trained on 4 verbs
- Identical algorithm parameters
- Achieved comparable recognition rates and obeying rates
- *Hol-daadan* and *feshaar-daadan* distinguish 2 senses of English push (move away vs. apply pressure
- Learns 2 senses of *feshaar-daadan*: one for steady pressure, one for pressing a button

MULTI-SLOT ARCHITECTURE



MULTI-SLOT TRAINING RESULTS: RUSSIAN

- Actions labelled with 12 root verbs, 6 prefixes, and 2 suffixes
- Slower training; requires on-line version of learning
- Acceptable learning of root verbs; some spurious associations are captured
- Suffixes capture perfectiveness with two senses, coding for the **duration** and **aspect** features
- "Explaining away" of features hurts recognition performance

CONCLUSIONS

- Action-verb semantics are grounded in motor control
    - high-level control is most relevant
    - universal constraints render learning tractable
    - features link action to reasoning
- Action-based criterion for distinguishing word senses
    - Bayesian formalism provides basis for learning
- Reversible mappings needed for obeying commands
    - suggests hardwiring of features is necessary
- Recruitment learning leads to a connectionist account
- X-schema mechanism has more general application (e.g. aspect, metaphor) and provides a scientific language for linguistic inquiry

NEURAL THEORY OF LANGUAGE PROJECT AT ICSI/BERKELEY