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

Verbs

3 senses

Obeying

Labelling

Linking feature structure

Execution Guidance

Feature Extraction

Motor Actions (X–Schemas)
PETRI NETS

BASIC BUILDING BLOCKS

place

transition (enabled)

before firing

after firing

token

SEQUENTIAL EXECUTION
PETRI NETS 2

CONDITIONAL EXECUTION

CONCURRENT EXECUTION
SLIDE X-SCHEMA

- Other x-schemas: DEPRESS, LIFT, ROTATE, TOUCH
CONNECTIONIST ACCOUNT:
X-SCHEMAS

"slipping grip"

perception of slippage

tighten grip transition

to grip-tightening muscles

to slide x-schema

"grip was tightened"

"want slippage stopped"

from slide x-schema
TEMPORAL BINDING OF PARAMETER VALUES

- Muscle commands for hand shaping
- SHAPING X-Schema
- Component for orienting hand (not specified)
- ESTIMATE MASS X-Schema
- Estimate mass of object
- SHAPING X-Schema
- Component for applying force (not specified)
- APPLY FORCE X-Schema
- Muscle commands for applying force
- REACH X-Schema
- Perceptual motor schema for locating object
- LOCATE OBJECT X-Schema
- From role node in ENABLE
- Assert in memory that object of PUSH has been pushed
- To node of ENABLE
- Focal clusters in other X-schemas that invoke the PUSH X-Schema
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

<table>
<thead>
<tr>
<th>motor parameter features</th>
<th>world state features</th>
</tr>
</thead>
<tbody>
<tr>
<td>slide</td>
<td>depress</td>
</tr>
<tr>
<td>true</td>
<td>false</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>world state features used by schema</th>
</tr>
</thead>
<tbody>
<tr>
<td>weight</td>
</tr>
<tr>
<td>2.3 lbs</td>
</tr>
</tbody>
</table>
CONNECTIONIST ACCOUNT: LINKING-FEATURES

language

FORCE NETWORK

force=low force=med force=high

SLIDE x-schema

LIFT x-schema

MOVE ARM synergy

muscles
# TWO SENSES OF PUSH

**PUSH: 2 senses**

<table>
<thead>
<tr>
<th>sense 1</th>
<th>sense 2</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>schema</strong></td>
<td><strong>schema</strong></td>
</tr>
<tr>
<td>slide 100%</td>
<td>slide 0%</td>
</tr>
<tr>
<td>touch 0%</td>
<td>touch 100%</td>
</tr>
<tr>
<td><strong>posture</strong></td>
<td><strong>posture</strong></td>
</tr>
<tr>
<td>palm 60%</td>
<td>palm 85%</td>
</tr>
<tr>
<td>grasp 10%</td>
<td>grasp 5%</td>
</tr>
<tr>
<td>index 30%</td>
<td>index 10%</td>
</tr>
<tr>
<td><strong>direction</strong></td>
<td><strong>force</strong></td>
</tr>
<tr>
<td>away 50%</td>
<td></td>
</tr>
<tr>
<td>toward 5%</td>
<td></td>
</tr>
<tr>
<td>up 15%</td>
<td></td>
</tr>
<tr>
<td>down 30%</td>
<td></td>
</tr>
</tbody>
</table>

**commonness 0.2**

**commonness 0.1**
DETAILED VIEW OF MODEL

**relevant linking features**

### Push

<table>
<thead>
<tr>
<th>Schema</th>
<th>Posture</th>
<th>Elbow Jnt</th>
<th>Aspect</th>
<th>Depressable</th>
</tr>
</thead>
<tbody>
<tr>
<td>slide 1.0</td>
<td>palm 0.7</td>
<td>extend 0.9</td>
<td>once 0.8</td>
<td>false 0.9</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Schema</th>
<th>Posture</th>
<th>Accel</th>
<th>Aspect</th>
<th>Depressable</th>
</tr>
</thead>
<tbody>
<tr>
<td>depress 1.0</td>
<td>indx 0.9</td>
<td>low 0.7</td>
<td>once 0.6</td>
<td>true 1.0</td>
</tr>
</tbody>
</table>

**motor parameter features**

**world state features**

### Shove

<table>
<thead>
<tr>
<th>Schema</th>
<th>Posture</th>
<th>Elbow Jnt</th>
<th>Accel</th>
</tr>
</thead>
<tbody>
<tr>
<td>slide 1.0</td>
<td>palm 0.9</td>
<td>extend 0.9</td>
<td>high 0.9</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Schema</th>
<th>Posture</th>
<th>Elbow Jnt</th>
<th>Aspect</th>
<th>Depressable</th>
</tr>
</thead>
<tbody>
<tr>
<td>depress 1.0</td>
<td>indx 0.9</td>
<td>low 0.7</td>
<td>once 0.6</td>
<td>true 1.0</td>
</tr>
</tbody>
</table>

world state features used by schema

- **weight**: 2.3 lbs
- **at goal**: false

---

**Slide Schema**

- **Start**
- **Small**
- **Preshape Grasp**
- **Preshape Palm**
- **Move Arm To (objloc)**
- **Apply Hand**
- **Move Arm (horiz-dir, force, dur)**
- **Slipping**
- **Tighten Grip**
- **At goal**
- **Not at goal**
- **Done**
HOW TO LEARN VERBS?
RATING A LEXICON MODEL BY ITS POSTERIOR PROBABILITY

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

P(m | t) → P(m) P(t | m)

P(m_s) → P(m_θ | m_s) → virtual samples
prefer fewer senses

P(l | v, m) = \max_{(l,v) \in t} P(l | v, m)

P(s | v, m) → P(l | s, v, m)
freq of this sense of v
look at sense’s distributions for each linking feature
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

TRAINING EXAMPLES
(linking f-struct)

<table>
<thead>
<tr>
<th>ex1</th>
<th>schema</th>
<th>elbow</th>
<th>posture</th>
<th>duration</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>slide</td>
<td>extend</td>
<td>palm</td>
<td>med</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ex2</th>
<th>schema</th>
<th>elbow</th>
<th>posture</th>
<th>duration</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>slide</td>
<td>extend</td>
<td>palm</td>
<td>short</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ex3</th>
<th>schema</th>
<th>elbow</th>
<th>posture</th>
<th>duration</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>touch</td>
<td>fixed</td>
<td>palm</td>
<td>long</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ex4</th>
<th>schema</th>
<th>elbow</th>
<th>posture</th>
<th>duration</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>slide</td>
<td>extend</td>
<td>index</td>
<td>long</td>
</tr>
</tbody>
</table>

WORD SENSES FOR "PUSH"

<table>
<thead>
<tr>
<th></th>
<th>schema</th>
<th>elbow</th>
<th>posture</th>
<th>duration</th>
</tr>
</thead>
<tbody>
<tr>
<td>ex1</td>
<td>slide 0.9</td>
<td>extend 0.9</td>
<td>palm 0.9</td>
<td>med 0.9</td>
</tr>
</tbody>
</table>

merge

<table>
<thead>
<tr>
<th></th>
<th>schema</th>
<th>elbow</th>
<th>posture</th>
<th>duration</th>
</tr>
</thead>
<tbody>
<tr>
<td>ex2</td>
<td>slide 0.9</td>
<td>extend 0.9</td>
<td>palm 0.9</td>
<td>med 0.5 short 0.5</td>
</tr>
</tbody>
</table>

merge

<table>
<thead>
<tr>
<th></th>
<th>schema</th>
<th>elbow</th>
<th>posture</th>
<th>duration</th>
</tr>
</thead>
<tbody>
<tr>
<td>ex3</td>
<td>touch 0.9</td>
<td>fixed 0.9</td>
<td>palm 0.9</td>
<td>long 0.9</td>
</tr>
</tbody>
</table>

merge

<table>
<thead>
<tr>
<th></th>
<th>schema</th>
<th>elbow</th>
<th>posture</th>
<th>duration</th>
</tr>
</thead>
<tbody>
<tr>
<td>ex4</td>
<td>slide 1.0</td>
<td>extend 1.0</td>
<td>palm 0.7</td>
<td>index 0.3</td>
</tr>
</tbody>
</table>
TRIANGLE NODES

- push
- posture
- palm

a + b + c ≥ 2

A, B, C
(external sources of activation)
A \[ a+b+c \geq 2 \]

(external sources of activation)

weighted links to all units
RECRUITEMENT LEARNING

Triangle unit WTA network

Concept units

recruited

free

T1

T2

T3

A

B

C

D

E

F

G
CONNECTIONIST ACCOUNT: VERB REPRESENTATION

- perceptual system
- motor control (x-schemas)
- phonology, morphology, etc.
- "push" "shove" "pull"
- force levels: force=low, force=med, force=high
- direction: dir=away, dir=down
- size: size=small, size=large

Diagram:

- Connections between concepts and their relationships.
- Visual representation of verb representation within a connectionist framework.
CONNECTIONIST ACCOUNT: MERGING WORD SENSES

motor control (x-schemas)

perceptual system

phonology, morphology, etc.

"push" "shove" "pull"

force force=low force=med force=high
dir dir=away dir=down
size size=small size=large

new merged sense

push1 push2
**TRAINING RESULTS: ENGLISH**

- 165 training examples; verb labels are:

<table>
<thead>
<tr>
<th>push</th>
<th>pull</th>
<th>shove</th>
<th>yank</th>
</tr>
</thead>
<tbody>
<tr>
<td>slide</td>
<td>press</td>
<td>touch</td>
<td>feel</td>
</tr>
<tr>
<td>slap</td>
<td>hit</td>
<td>tap</td>
<td>poke</td>
</tr>
<tr>
<td>lift</td>
<td>pick-up</td>
<td>heave</td>
<td>hold</td>
</tr>
<tr>
<td>turn</td>
<td>roll</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- 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

log probability

prior
likelihood
posterior

merges performed
LEARNED SENSES OF PUSH

**push74 (12 ex)**

```plaintext
{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}
 post {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}}
```

**push78 (21 ex)**

```plaintext
{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}
 post {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}}
```

**push79 (8 ex)**

```plaintext
{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}
 post {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}}
```
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

% recognized

90 85 80 75 70 65 60 55 50

number of senses

28 26 24 22 20
- 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
"Push left"

**Schema Forcedir**

**Slot 1**
- **Push**
  - Schema: push
  - Accel: schema
  - Dir: accel
  - Slide: slide
  - Med: med
  - Away: away

**Slot 2**
- **Shove**
  - Schema: shove
  - Accel: schema
  - Dir: accel
  - Slide: slide
  - High: high
  - Away: away

**Linking b-struct**
- **Slide**
  - Schema: slide
  - Accel: med
  - Dir: left
  - Force: 4 ft-lb

**Activate x-schema**
- **Move arm**
  - Slide...
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