Automatic Identification of Figurative Language

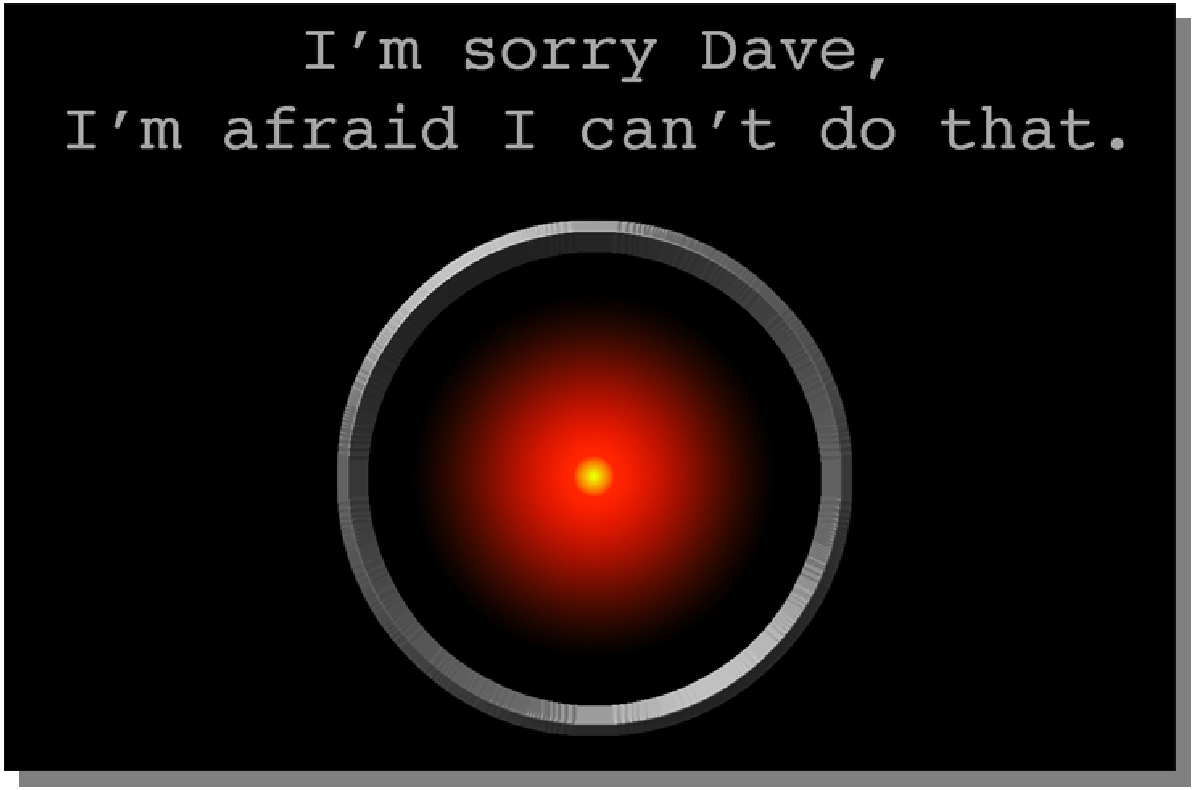
Presented by: Saša Milić

Special thanks to:
My supervisor Suzanne Stevenson
and my mentor Afsaneh Fazly

June 19, 2013

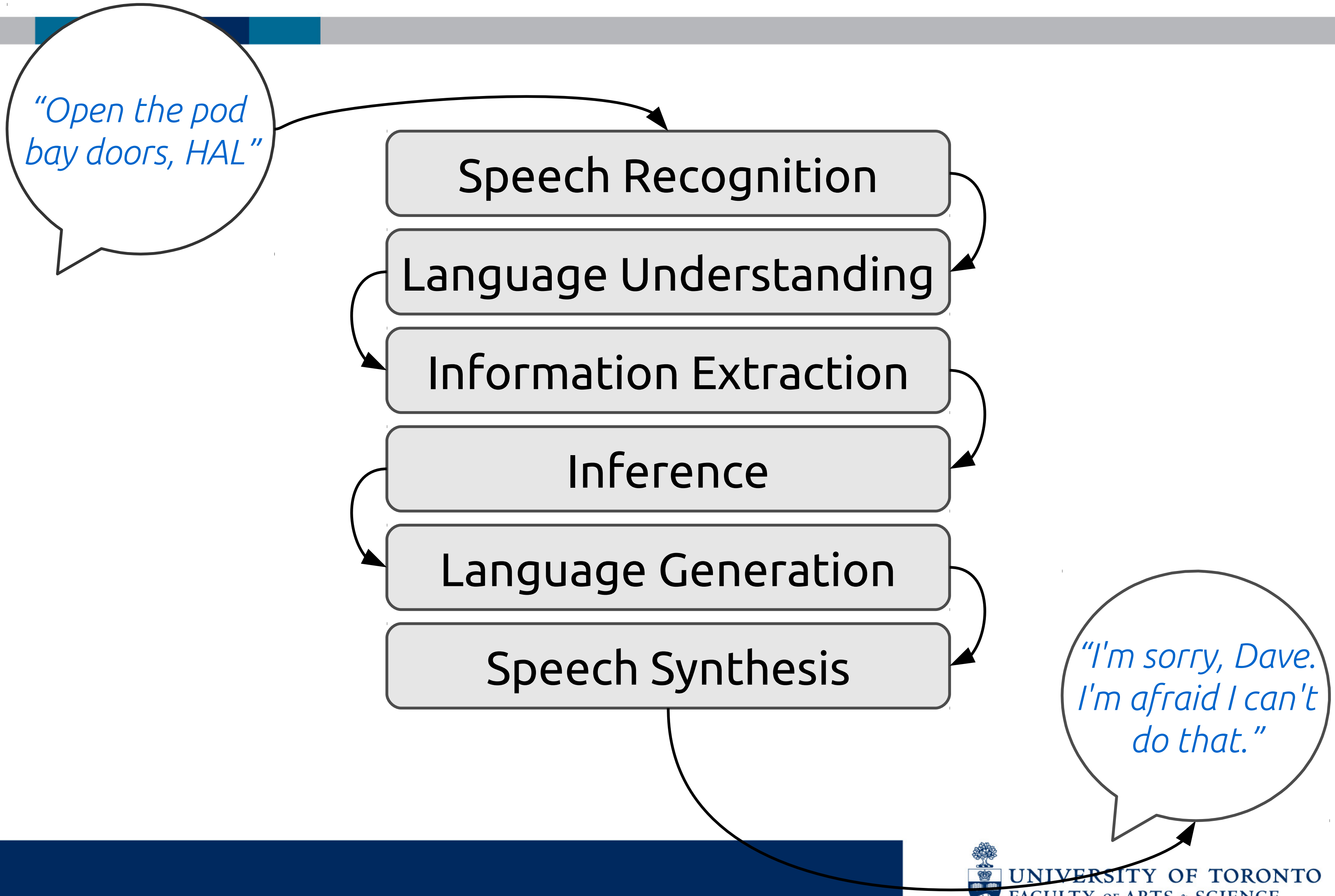
UNIVERSITY OF TORONTO
FACULTY OF ARTS & SCIENCE

A Conversant Computer?



(Source: http://www.604republic.com/gocms/)

The Minimal Requirements



"Open the pod bay doors, HAL"

Speech Recognition

Language Understanding

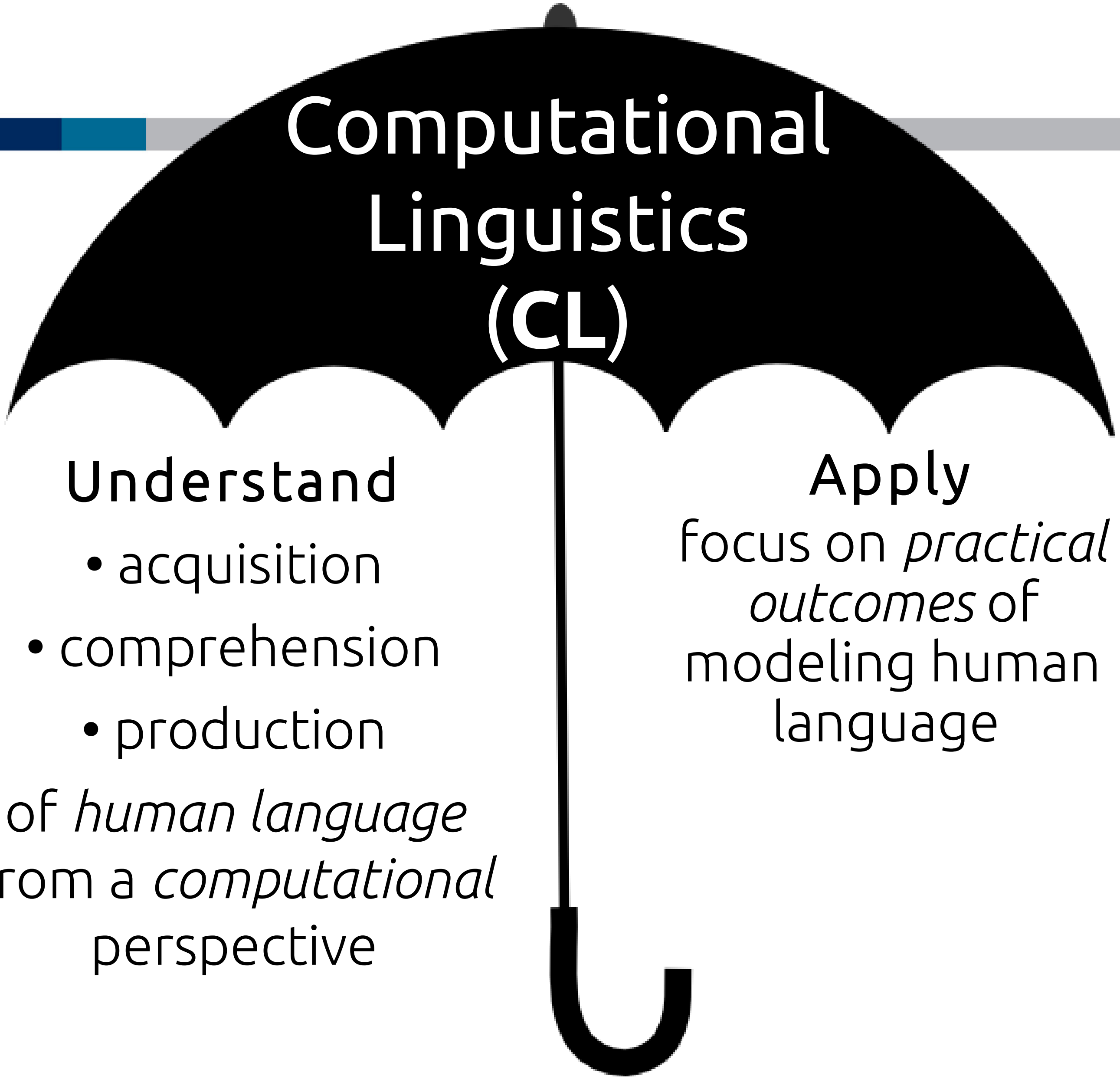
Information Extraction

Inference

Language Generation

Speech Synthesis

"I'm sorry, Dave. I'm afraid I can't do that."



Computational Linguistics (CL)

Understand

- acquisition
- comprehension
- production

of *human language* from a *computational* perspective

Apply

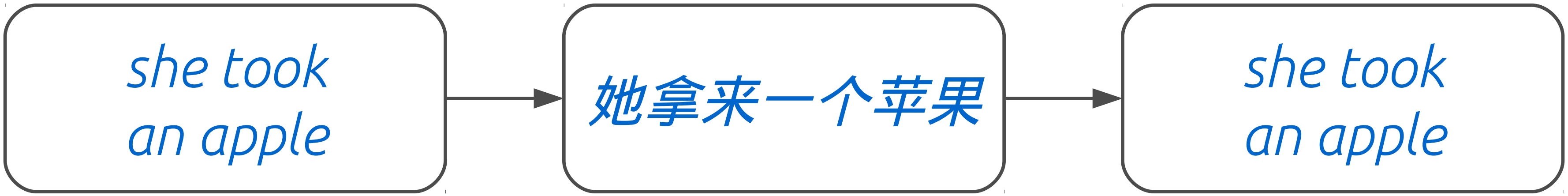
focus on *practical outcomes* of modeling human language

Applications of CL

- Grammar and style checking
- Apple's *Siri*
- Search Engine
- Machine translation

Google Translate : An Informal Experiment

- Translating a literal phrase:

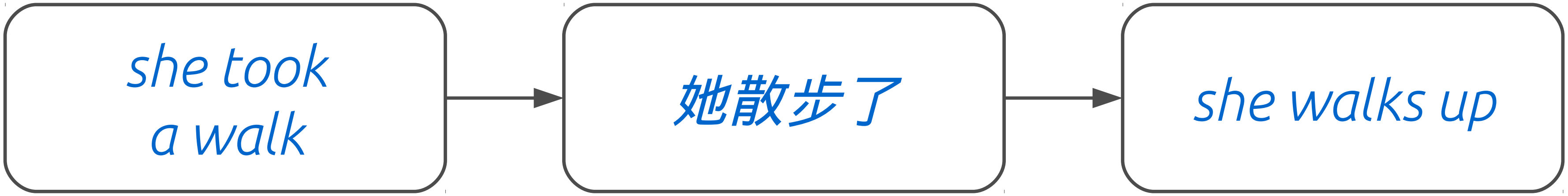


she took an apple

她拿来一个苹果

she took an apple

- Translating a multiword expression:



she took a walk

她散步了

she walks up

Difficulty with Multiword Expressions

- Multiword expression:
  - two or more words that together form a *single unit* of meaning
    - *“frying pan”*
    - *“keep an eye out for”*
    - *“shoot the breeze”*
- overall meaning ≠ sum of the meaning of the components

Light Verb Construction (LVC)

- A multiword expression (in our case, verb + noun) where the noun determines the primary meaning of the whole

LVC	"give a sigh"	"make a decision"	"take a walk"
Literal	"give a present"	"make a cake"	"take an apple"

- *Again:*
  - – overall meaning ≠ sum of the meaning of the components
- *However:*
  - – the component meanings still contribute something to the overall meaning

Identifying LVCs

- Which of the following is a light verb construction?
  - *He gave a donation.*
  - *It took place over there.*
  - *He gave her an advantage.*

- Motivates the question: can we do better than a simple binary classification?

A More Appropriate Measure

- Binary decision-making vs graded decision-making
  - *“Is this an LVC?”* vs *“How acceptable is this as an LVC?”*

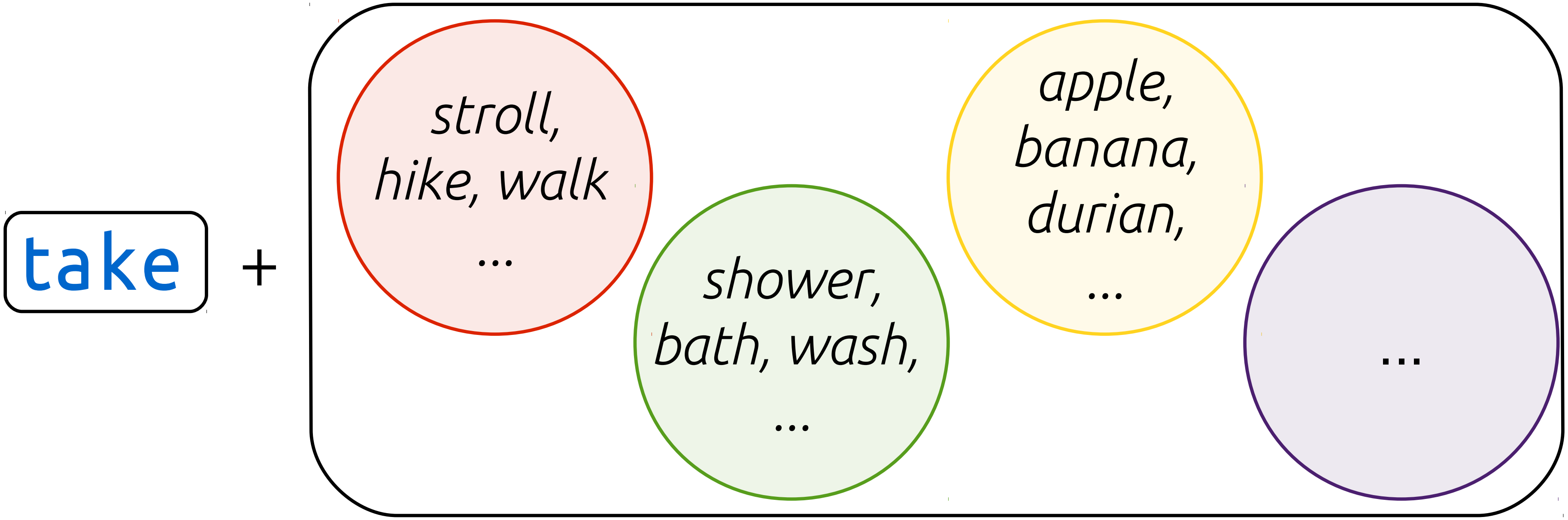
- More formally:
  - What is the probability that some verb + noun combination forms an LVC?

- New measure: **Acceptability**

Measuring Acceptability

- Linguistic studies suggest that a measure of LVC acceptability should incorporate both frequency and semantic similarity.
- Hypothesis:
  - a *novel* LV + noun is considered more acceptable if the noun is similar to a noun in a high-frequency LVC
- Example:
  - How acceptable is "*take a saunter*"?

“take a saunter”



take +

stroll,
hike, walk
…

shower,
bath, wash,
…

apple,
banana,
durian,
…

…

C(take) = { 🔴, 🟢, 🟡, 🟣, … }

C(v): set of semantic classes of nouns that can occur with verb v

"take a saunter"

P(saunter belongs to *stroll, hike, walk, …* *)* = *?*

*P(*saunter| ○ *)* = high

P(n|c): probability that noun n belongs to class c

“take a saunter”

$$P(\text{take} + \text{stroll, hike, walk, …} = \text{LVC}) = ?$$

$$P_{LVC}(\bigcirc \mid \text{take}) = \text{high}$$

$P_{LVC}(c|v)$: probability that class c forms acceptable LVCs with v

Measuring Acceptability

- Acceptability:
  - A *probabilistic* measure

- Components
  - $C(v)$
  - $P(n|c)$
  - $P_{LVC}(c|v)$

Estimating Probabilities

- We can't know the true probabilities. So we estimate.
- In order to estimate $P_{LVC}(c|v)$ we need to know:
  - $P_{LVC}(n|v)$
    - for all n in class c
  - Estimate **directly**
    - *Why can't we do this for novel LVCs?*
  - Estimate **indirectly**

Estimating Probabilities

- We use a machine learning algorithm to estimate this *directly* for frequent combinations :
  - $P_{LVC}(n\,|\,v)$

- Using ~25 features drawing on linguistic properties of LVCs
  - Examples:
    - frequencies
    - association
    - syntactic behavior

Some Features of LVCs

- We expect the noun and the verb in an LVC to have strong associativity

- We expect LVCs to have a preference for indefinite determiners (*"a"*, *"an"*, …)
  - consider:
    - *"make a speech"* **vs** *"make the speech"*
  - Which one occurs more often?
    - ~16 million vs ~2 million Google hits

Evaluation

- Obtain human ratings (on some scale) of LVC acceptability

- Goals:
  - to introduce a more appropriate (*linguistically-motivated*) measure for identifying LVCs
  - to be able to predict LVC acceptability of novel expressions