Source: xkcd

CS 671 NATURAL LANGUAGE PROCESSING

amitabha mukerjee
iit kanpur

Learning Objectives

- Nature of Language
- Levels of Computational Models
  - Sound units (Phonemes / Syllables)
  - Words (Lexical Units)
  - Syntax (Morphology / Grammar)
  - Meaning (Semantics)
- Rules vs Learning
- Applications

The magic of language

The magic of language

- You can't hold two watermelons in one hand
  - Iranian proverb

The magic of language

- Language is about conveying meaning
- Language is one-dimensional – Meaning is multi-dimensional

- Challenges
  Sounds along one-dimension express multi-dimensional aspects of reality
  - Same sounds map to different meanings [**Polysemy**]
  - Same meanings map to different sounds [**Synonymy**]

Language as Representation

Myths about language

- **grammar** is about whether language is correct or incorrect

  It's me.

  Ganesh is at home?

  There are many small-small holes in this dress.

Myths about language

- **grammar** is about whether language is correct or incorrect

  It's me (accusative) → *"It's I"*

  Ganesh is at home? → *Is Ganesh at home?*

  There are many small-small holes in this dress.

- But how do we decide what is right?
- In Linguistics, grammar is determined based on language use.
  - descriptive, not prescriptive

Myths about language

- **grammar** is about the correct and incorrectness of language.

  Ganesh is at home? → *Is Ganesh at home?*

  It's me (accusative) → *"It's I"*

  There are many small-small holes in this dress.

- words are separated by spaces.
- how many sounds are there in English? 26

Myths about language

- **grammar** is about the correct and incorrectness of language.

  Ganesh is at home? → *Is Ganesh at home?*

  It's me (accusative) → *"It's I"*

  There are many small-small holes in this dress.

- words are separated by spaces.
- alphabets are the sounds of language

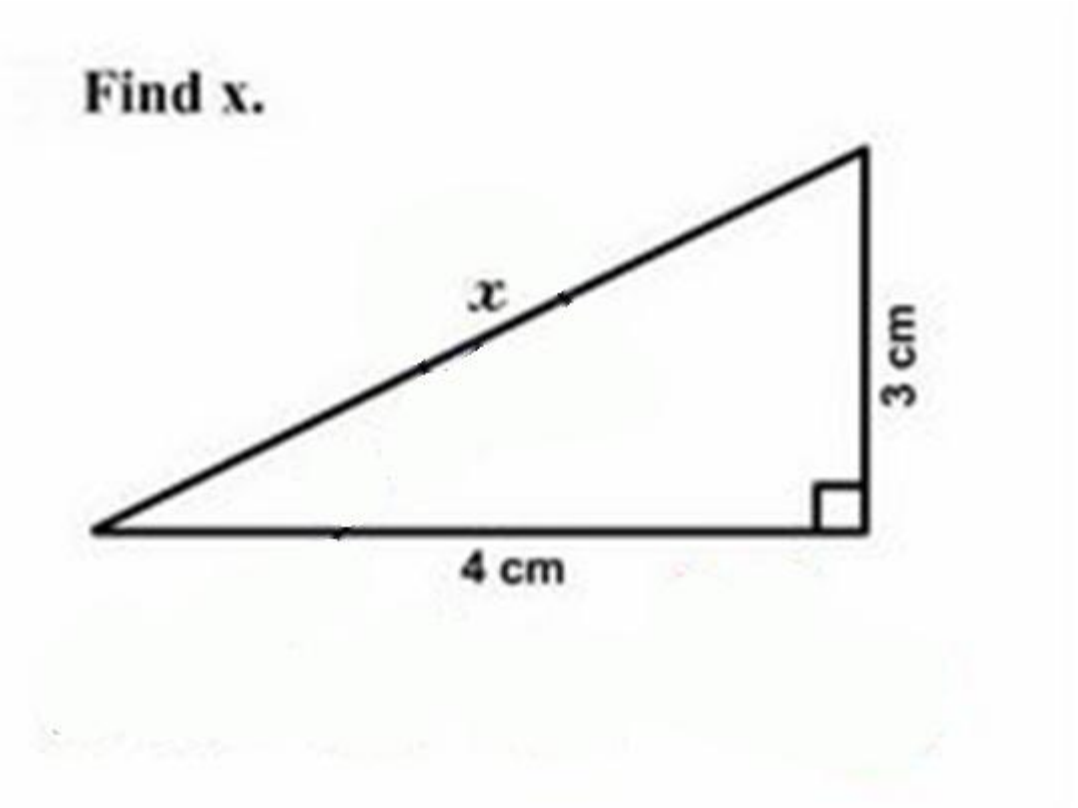
Levels of Grammar

- **Morphology** : how words are formed from smaller bits

  (*unopened* = *un* + *open* + *ed*)
- **Syntax**: how words are combined into sentences

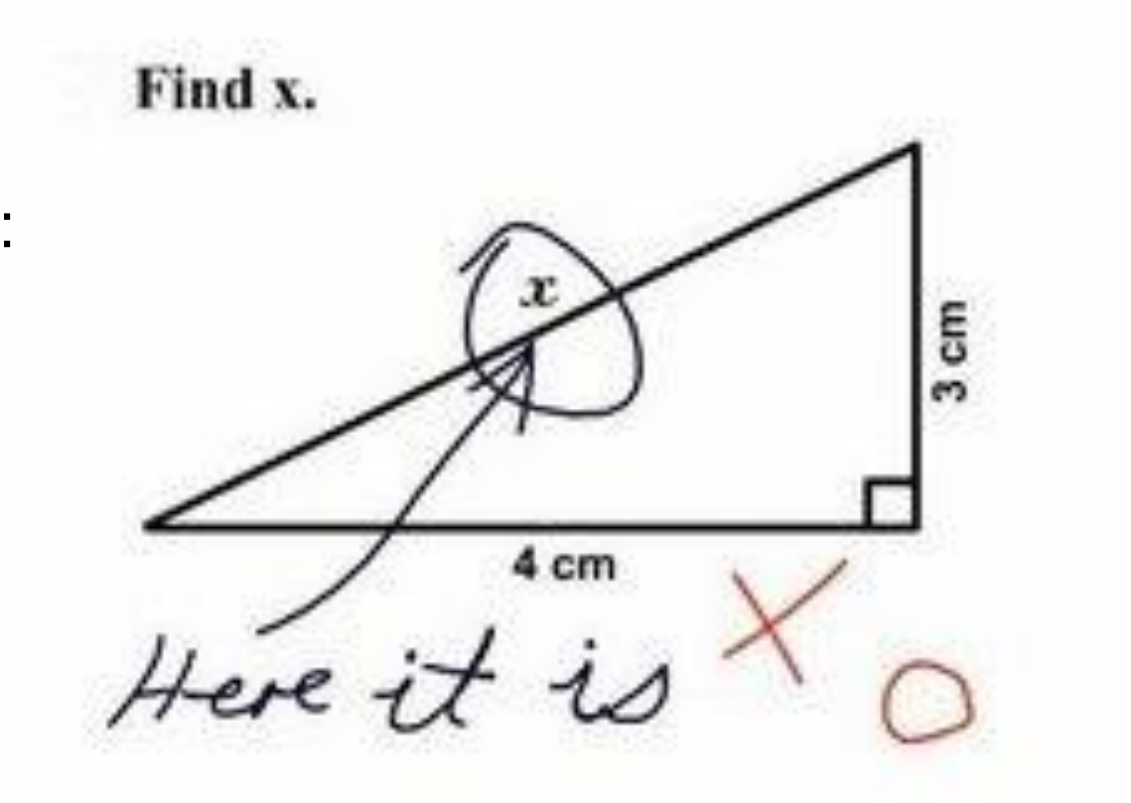
- Other levels of analysis:
  - Phonology : what sounds change the meaning
  - Lexicon : the inventory of *arbitrary* (?) words
  - Semantics : what language means directly
  - Pragmatics : what one infers from an utterance

Pragmatics: Meaning in Context



Pragmatics: Direct vs Indirect meaning

Traditional thinking:
Semantics
Direct meaning
Pragmatics
Indirect meaning



Pragmatics: Meaning in Context

Traditional levels of analysis:

- **Semantics**: composition from lexical meaning of words – “find” = detect, locate. [*direct meaning*]
- **Pragmatics**: social / contextual meaning ; [indirect meaning]

Psycholinguists:

Retrieval of pragmatic meaning is often faster

Language Structure: Levels

boys like girls

Language Structure: Levels

- **Phonology**
- **Lexicon**
- **Syntax [Morphology]**
- **Discourse**

- **Semantics / Compositionality**
- **Pragmatics / Discourse**

Language Structure: Levels

- **Phonology :** sounds of speech
  phoneme /b/ /oy/ /z/
- **Lexicon :** set of meaning-bearing units, **lexemes**
- **Syntax :** composing lexemes **composition**
  - **Word =** base + affixes / suffixes
  - **Phrase**: [[[boys] like] girls]
- **Discourse :** **17**Boy likes girl. They meet.

NLP: Goals

Language → NLP → Decision

(NL Understanding)

NLP (MT)

Language 1 → Machine → Language 2

Translation

Situation → NLP → Language

(NL Generation)

Language Maps: Levels

- **Semantics** direct meaning
- **Pragmatics** social / implied meaning

NLP: Levels

NLP : deals with text. For languages with space-separation, deal with "orthographic words"

- **Morphology:** structures smaller than words
- **Syntax :** structures larger than words

- **Phonology:** impacts how text is written

Phonology

- Wide diversity in pronunciation and in hearing, yet we comprehend each other

- Phonetics: All possible human speech sounds ***phone***
- Phonology: organization and structure of sounds of a language
  - ***Phoneme*** Minimal pair: *zip* | *sip*
    → */z/ and /s/ are different phonemes in English*

Speech sounds (phonemes)

- Which sounds change a meaning?

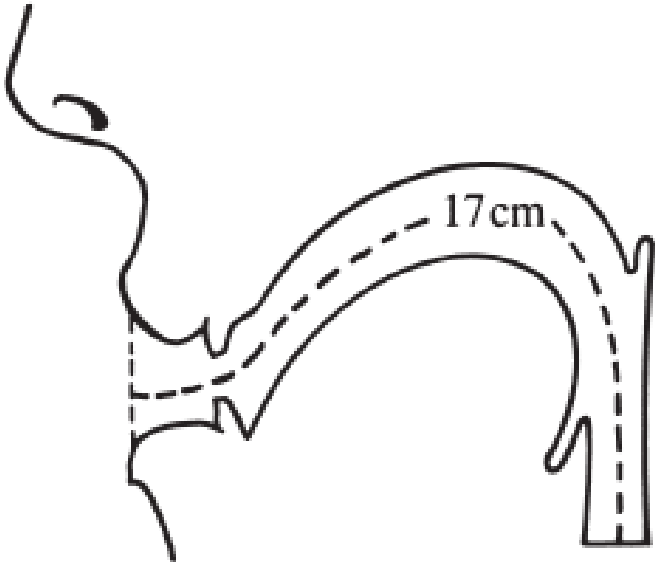
  pin, tin, kin, fin, thin, sin, shin

  dim, din, ding, did, dig, dish

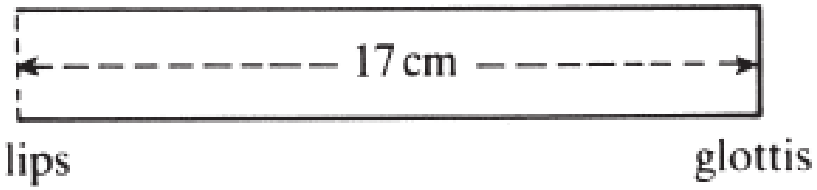
  pin, pen, pan, pun, pain, pine, pawn

- Phonemes at middle of syllable: **vowel**
  start or end: **consonant**

Vocal organs



tube model of vocal tract (for most neutral vowel)

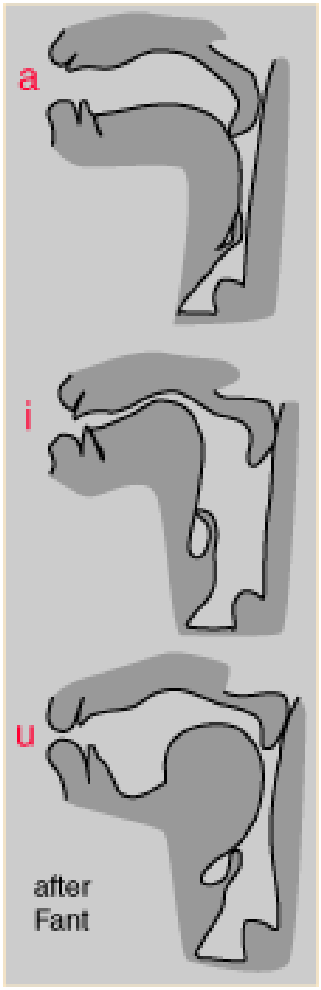


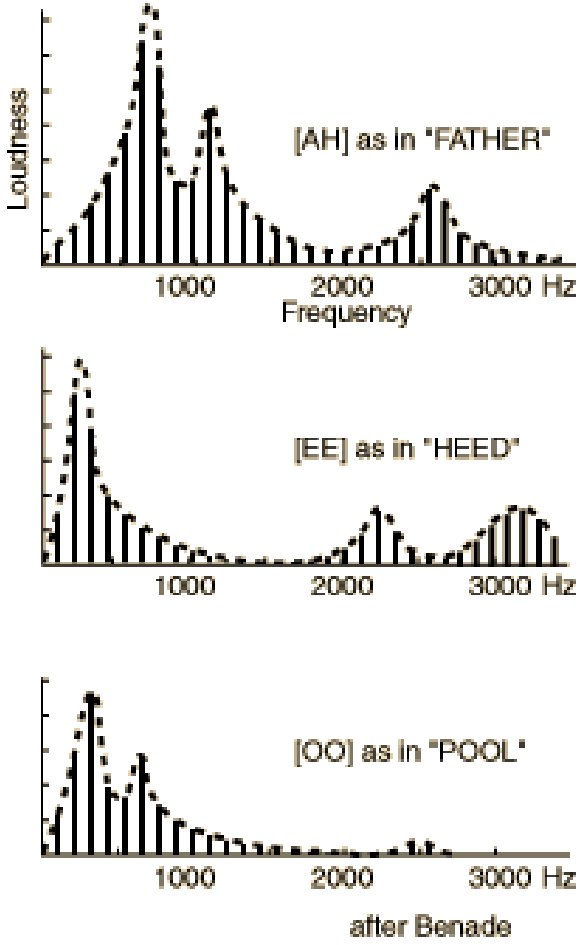
[malmkjaer 02]

Vowels : Formants

formant frequencies: peaks in the harmonic spectrum of vowel sounds

first three: F1, F2, F3





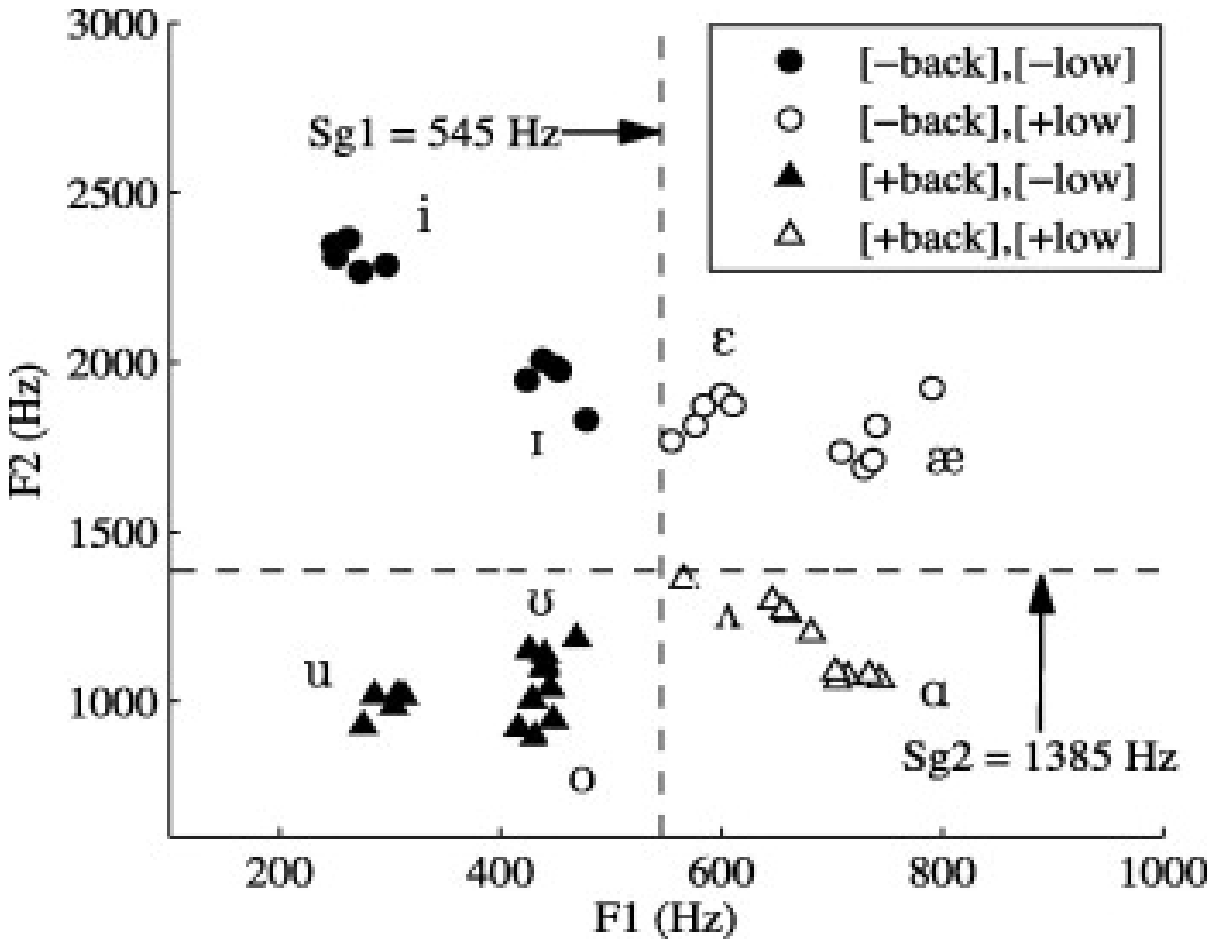
http://hyperphysics.phy-astr.gsu.edu/hbase/music/vowel.html

Vowels : Formants

Vowel space (F1,F2)

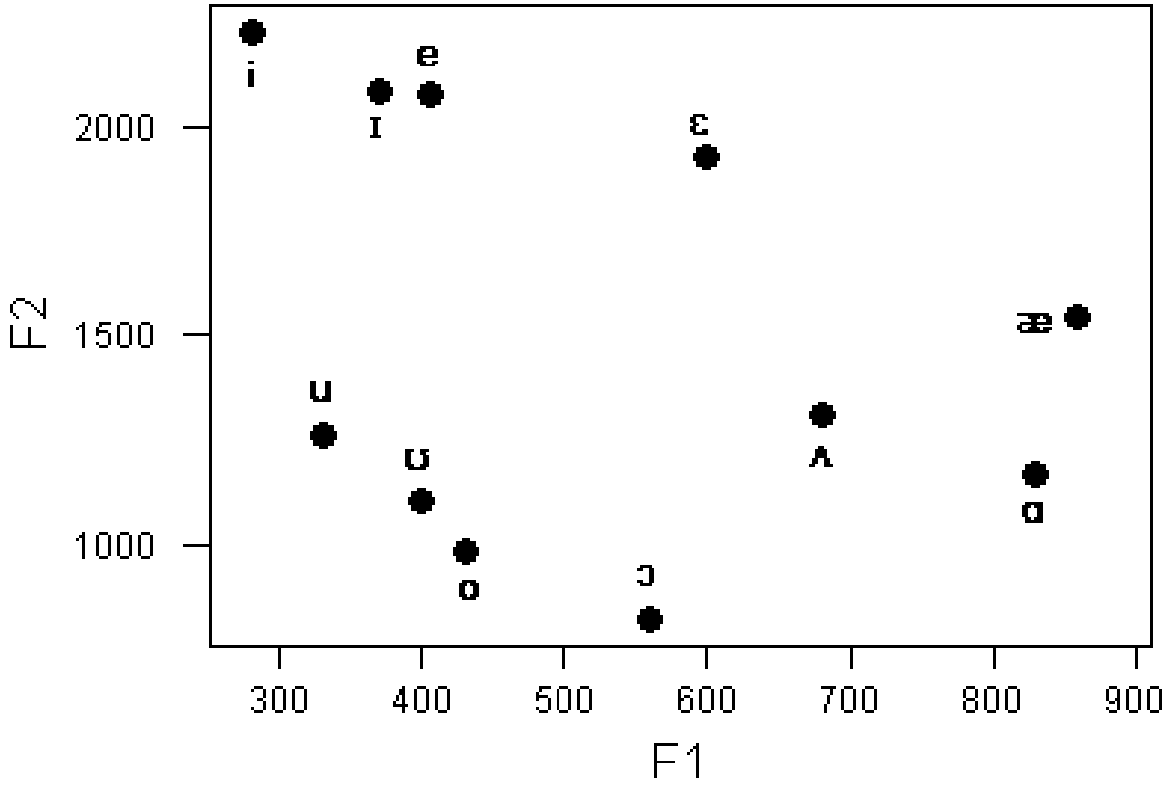
+low: F1 > 545hz;
+back : F2 > 1385hz

for a particular American English speaker (male):



[arsikere etal 11]

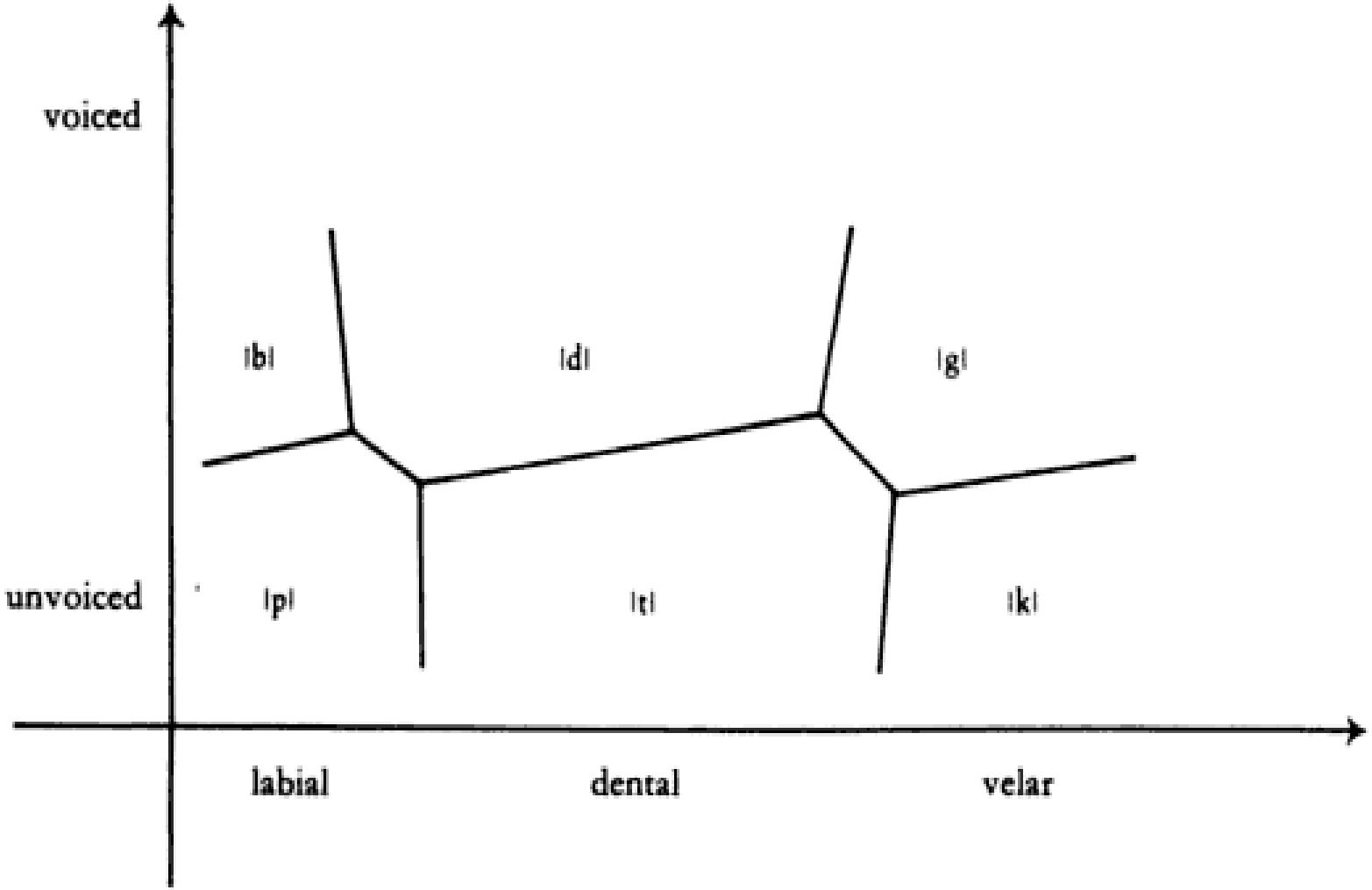
Vowels : Formants



Canadian English

http://hyperphysics.phy-astr.gsu.edu/hbase/music/vowel.html

Partitioning the speech sound space



[petitot 1989], [gardenfors 00]

Writing : Consonants

stop consonants

voiceless		voiced		nasal	
inaspirate	aspirated	in-	aspirated		
क	ख	ग	घ	ङ	[velar]
च	छ	ज	झ	ञ	[palatal]
ट	ठ	ड	ढ	ण	[retroflex]
त	थ	द	ध	न	[dental]
प	फ	ब	भ	म	[labial]

Consonants

stop consonants

voiceless		voiced		nasal	
inaspirate	aspirated	in-	aspirated		
k	kh	g	gh	N	[velar]
c	chh	j[dz]	jh[dzh]	n~	[palatal]
T	Th	D	Dh	N	[retroflex]
t	th	d	dh	n	[dental]
p	ph	b	bh	m	[labial] (bilabial)

Phonetic Notation

boys like girls

/bɔjz/ /lajk/ /gərlz/

Grammar of Phonology

"cats" → "cat" + /s/

"boys" → "boy" + /z/

Language Structures 2
Morphosyntax

Language Structure: Levels

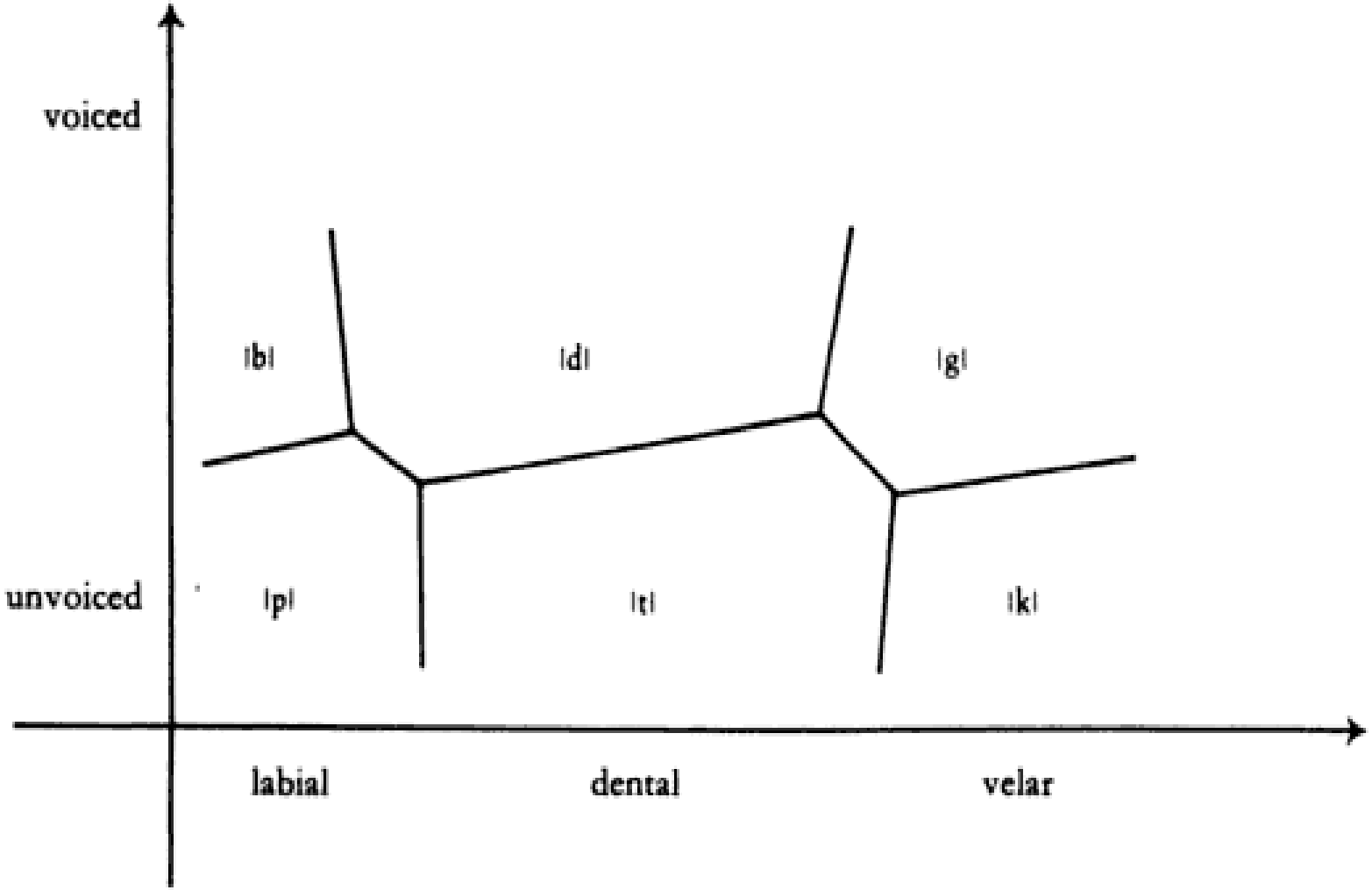
- **Phonology**
- **Lexicon**
- **Syntax [+Morphology]**
- **Discourse**

- **Semantics / Compositionality**
- **Pragmatics / Discourse**

- **Prosody**

- **Orthography / Graphology**

Phonological inventory



[petitot 1989], [gardenfors 00]

Combining sounds : grammar

boys like girls

/bɔjz/ /lajk/ /gərlz/

Lexicon vs Grammar

- Grammar: how larger structures are assembled from smaller ones
- Smallest meaning-bearing structures = unit
- **morpheme :** less likely to appear independently

  -er , -s, -ly, -able
- **lexeme**

  cat, boy, smart, undergraduate student, cook, cooker

Lexicon vs Grammar

- lexicon = mental inventory of units
  = set of all lexemes

- Is “cats” a lexeme?

cook → **cooks** : grammatical (rule-driven, inflection)
→ **cooker** : cook + er (not fully a rule; derivation)

Older thinking : lexicon is separate from grammar
at present : lexicon - grammar is a continuum

Syntax (morphosyntax)

- Regularity in how larger structures are assembled from units or smaller structures

- **morphology**

  cook-er / read-er / *-ercook

- **phrase syntax**

  smart woman / *woman smart

- **sentence syntax**

  boys like girls / girls like boys / *like boys girls