Lecture 1: What is Linguistics?
Ma 191c: Mathematical Models of Generative Linguistics

Matilde Marcolli

Caltech, Spring 2024

• *Linguistics* is the scientific study of language

- What is Language? (langage, lenguaje, ...)
- What is a Language? (lange, lengua,...)

Similar to 'What is Life?' or 'What is an organism?' in biology

• *natural* language
as opposed to artificial (formal, programming, ...) languages

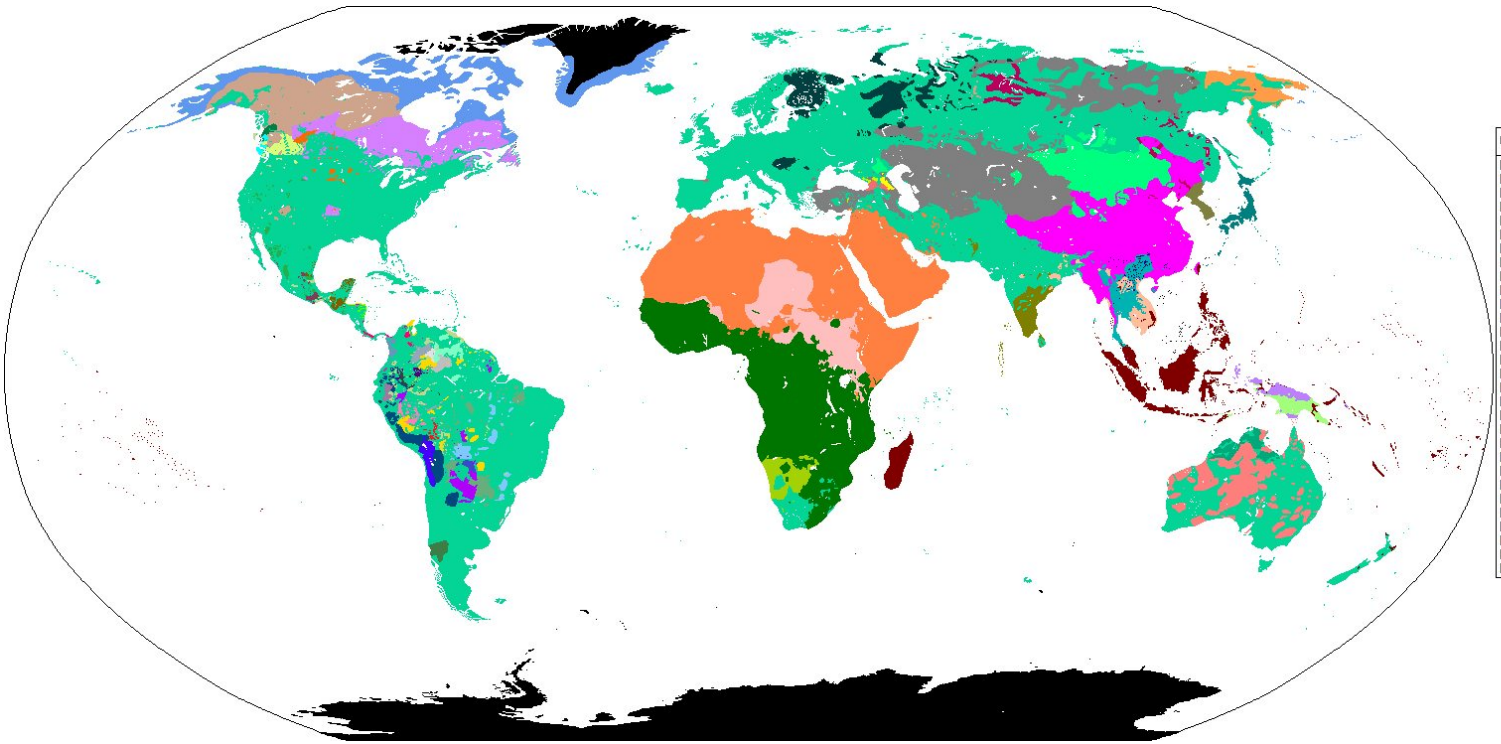
• The point of view we will focus on:
Language is a kind of *Structure*

- It can be approached mathematically and computationally, like many other kinds of structures
- The main purpose of mathematics is the understanding of structures

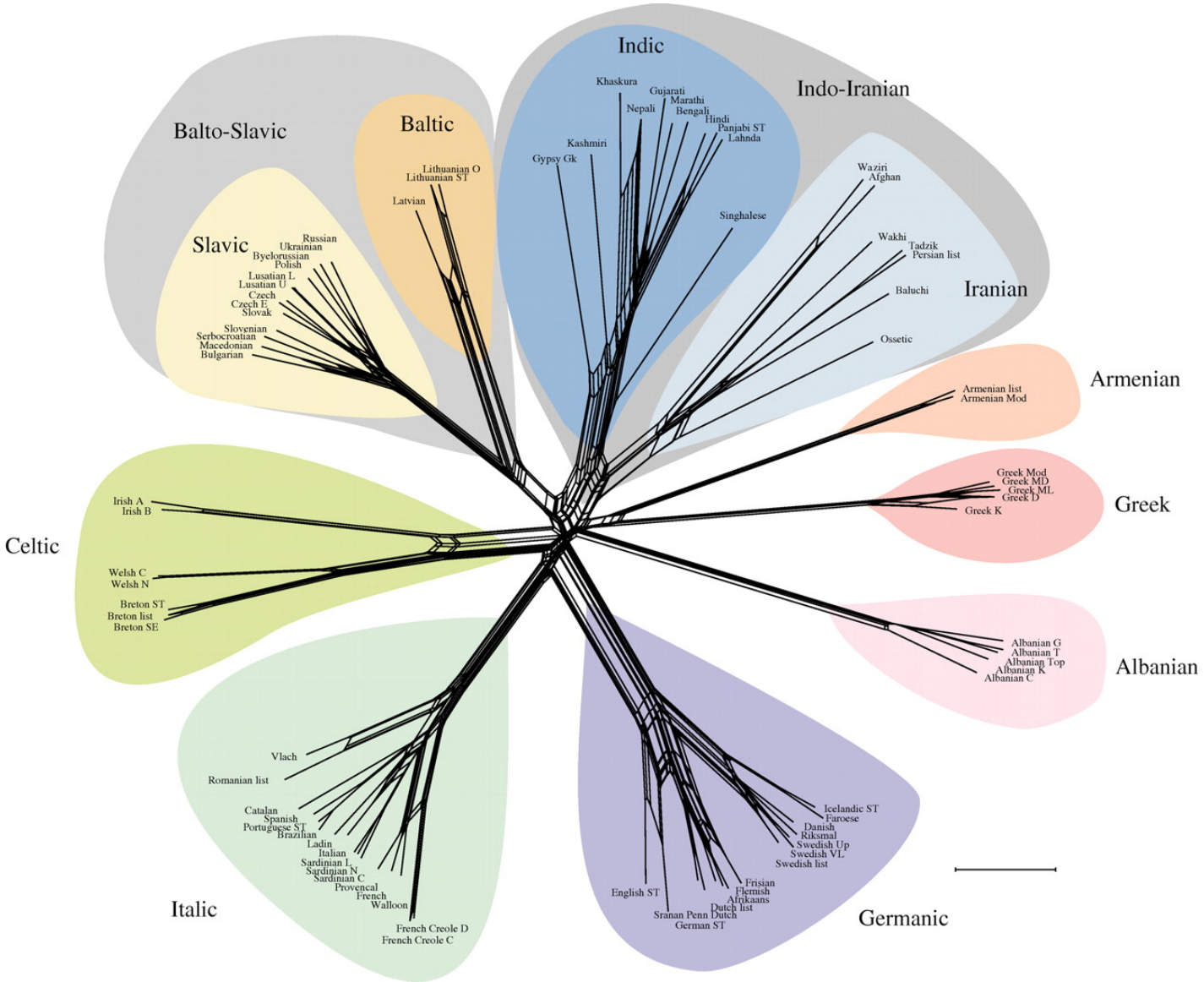
Language Families

- Niger-Congo (1,532)
- Austronesian (1,257)
- Trans New Guinea (477)
- Sino-Tibetan (449)
- Indo-European (439)
- Afro-Asiatic (374)
- Nilo-Saharian (205)
- Oto-Manguean (177)
- Austro-Asiatic (169)
- Tai-Kadai (92)
- Dravidian (85)
- Creole (82)
- Tupian (76)
- Mayan (69)
- Altaic (66)
- Uto-Aztecan (61)

- Arawakan (59)
- Torricelli (56)
- Sepik (55)
- Quechuan (46)
- Na-Dene (46)
- Algic (44)
- Hmong-Mien (38)
- Uralic (37)
- North Caucasian (34)
- Penutian (33)
- Macro-Ge (32)
- Ramu-Lower Sepik (32)
- Carib (31)
- Panoan (28)
- Khoisan (27)
- Salishan (26)
- Tucanoan (25)
- Isolated Languages (75)


Human Language Families
Uninhabited
Afro-Asiatic
Algic
American Indian (several families)
Araucanian
Arawakan
Arawan
Australian (several families)
Austro-Asiatic
Austronesian
Aymaran
Barbacoan
Basque (isolate)
Bora-Witóto
Cariban
Caspian
Chibchan
Choco
Chukotko-Kamchatkan
Dené-Yeniseian
Dravidian
Eskimo-Aleut
Guajiboan
Hmong-Mien
Indo-European
Japanic
Jivaroan
Kartvelian
Koreanic
Khoisan
Macro-Jê
Mataco-Guaicuru
Mayan
Misumalpan
Mongolic
Niger-Congo
Nilo-Saharan
Nivkh (isolate)
Papuan (several families)
Pama-Nyungan
Panoan
Pontic
Oto-Manguean
Quechuan
Saliban
Salishan
Sino-Tibetan
Siouan
Tacanan
Tai-Kadai
Trans-New Guinea
Tsimshianic
Tucanoan
Tungusic
Tupian
Turkic
Uralic
Uto-Aztecan
Wakashan
Yanomaman
Yukaghir
Zamucoan

The Indo-European Language Family

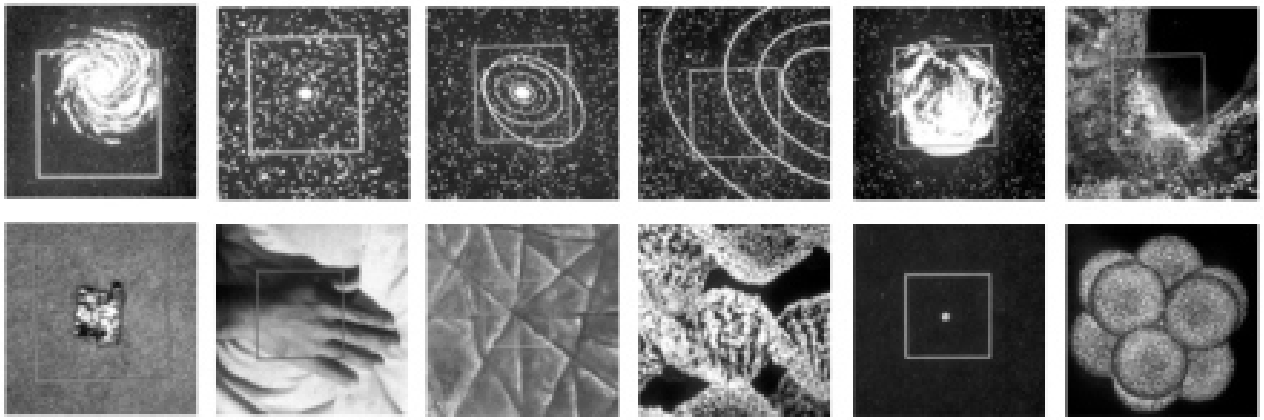


Some questions

- what is the fundamental *generative, computational* principle that allows for the richness of human language and the infinite possibility of sentences?
- what are the *constraints and rules* of this generative process?
- what principles *structure and organize* the diversity of linguistic forms we see across human languages?
- how does this structure formation and structure recognition happen in the human brain?
- how is language acquired?
- how did the faculty of language evolve?
- how do languages change dynamically over time?

Language at different scales

language exists at many different levels of structure just as physics looks very different at different *scales*:



language at different "scales":

- units of sound (phonology)
- words (morphology)
- sentences (syntax)
- global meaning (semantics)

Syntax is the *large scale structure of language*, it is robust and highly structured, it is crucial for the *compositionality* of language, and necessary to the encoding of *complex* meaning

An analogy

Physics looks very different at different *scales*:

- General Relativity and Cosmology ($\geq 10^{10}$ m)
- Classical Physics (~ 1 m)
- Quantum Physics ($\leq 10^{-10}$ m)
- Quantum Gravity (10^{-35} m)

different mathematical models at different levels of structure

Similarly, we view language at different "scales":

- units of sound
- words
- sentences
- global meaning

We expect different mathematical structures at these various levels

Levels of Structure

language exists at many different levels of structure

- Phonology: *sound* structures, building blocks (phonemes) phonetics (physical aspects), autosegmental phonology (tone systems), feature geometry (generative theory), optimality theory (neural networks)
- Morphology: *words* (roots, affixes, stress), building blocks (morphemes), morphological typology (use of morphemes), lexicology, word formation, paradigms (words associated to same lexeme, conjugation, declension), morphosyntax

- Syntax: ($\sigma\upsilon\nu\tau\alpha\xi\iota\varsigma$ = ordering together) *sentences* (principles, rules, and processes by which sentences are formed), generative grammar (i-language, grammaticality), formal languages, transformational grammar, principles and parameters (government and binding), merge and Minimalism
- Semantics: *meaning* (how communication of meaning happens through signifiers like words, sentences), homonymy, synonymy, cognitive linguistics, philosophy of language, truth values (formal logic, propositional calculus), computational semantics, conceptual spaces

Small Scale Structure: Phonetics

• Certain speech-sounds are specific to certain languages (clicks in Khoisan languages)

• Minimum known: 11 phonemes in Rotokas (East Papua)

• Maximum known: 141 phonemes in !Kung (Khoisan group, Namibia)

• Consonants (complete or partial closure of vocal tract): two coordinates

- place: labial, coronal, dorsal, radical, glottal

- manner: nasal, stop, sibilant fricative, non-sibilant fricative, approximant, flap or tap, trill, lateral fricative, lateral approximant, lateral flap

• Vowels (open vocal tract): close/mid/open

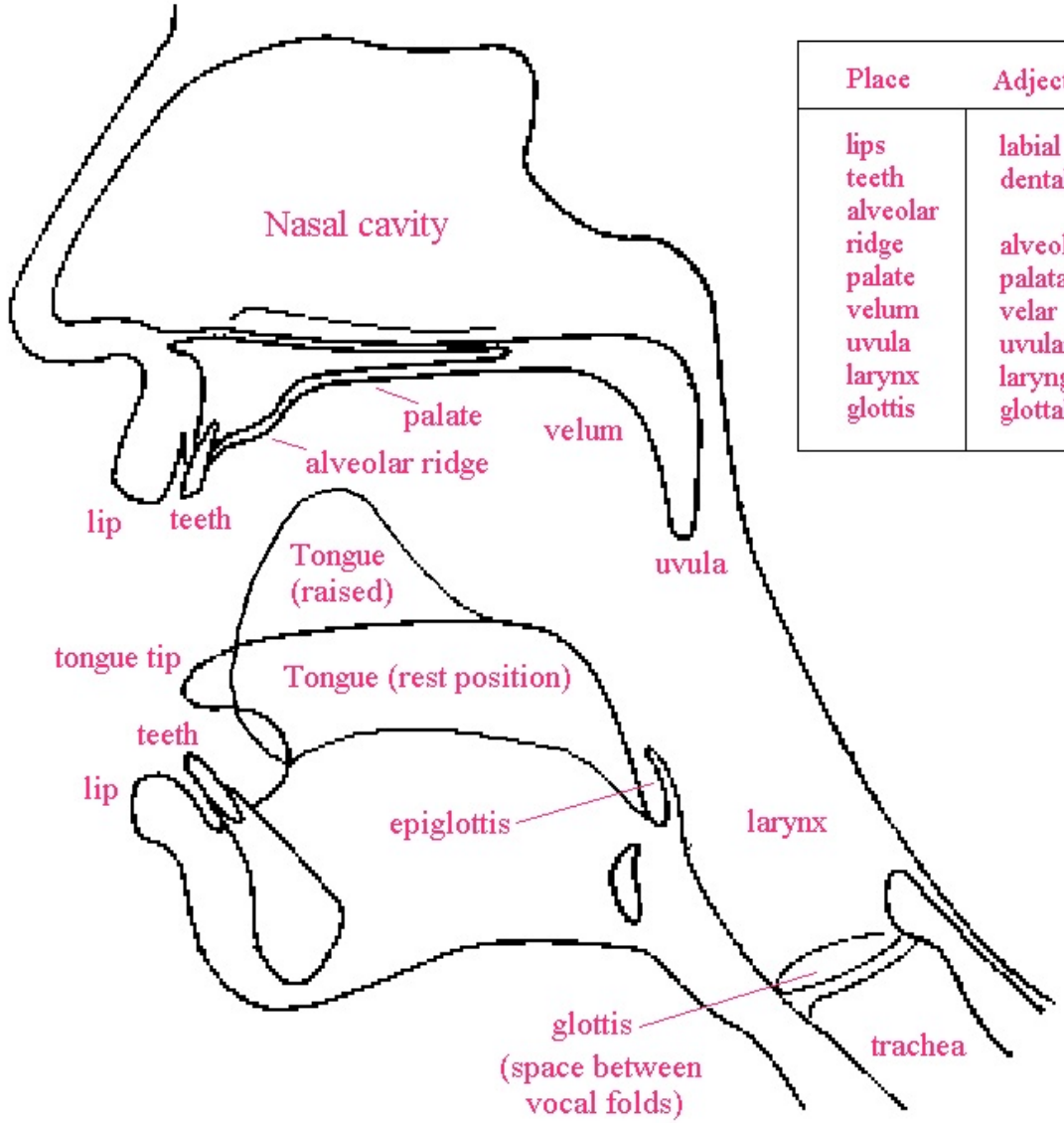
• Other: Semivowels, approximants, diphthongs, fricative vowels

• Production modes: pulmonic egressive (most languages) + other mechanisms (ejectives, clicks, implosives)

• Phonotactics: rules and restrictions on arrangement of sounds within syllables of a language

Articulatory Phonetics

- Phonemes classified according to vocal organs



Place	Adjective
lips	labial
teeth	dental
alveolar ridge	alveolar
palate	palatal
velum	velar
uvula	uvular
larynx	laryngeal
glottis	glottal

International Phonetic Alphabet (IPA)

- universal phonetic alphabet for all world languages

IPA chart of pulmonic consonants:

the international phonetic alphabet (2005)

consonants (pulmonic)	LABIAL		CORONAL				DORSAL				RADICAL		LARYNGEAL
	Bilabial	Labio-dental	Dental	Alveolar	Palato-alveolar	Retroflex	Alveolo-palatal	Palatal	Velar	Uvular	Pharyngeal	Epi-glottal	Glottal
Nasal	m	ɱ		n		ɳ		ɲ	ŋ	ɴ			
Plosive	p b			t d		ʈ ɖ		c ɟ	k g	q ɢ		ʡ	ʔ
Fricative	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ɕ ʑ	ç ʝ	x ɣ	χ ʁ	ħ ʕ	ʜ ʢ	h ɦ
Approximant		ʋ		ɹ		ɻ		j	ɰ				
Tap, flap		ⱱ		ɾ		ɽ							
Trill	ʙ			r						ʀ		ʀ	
Lateral fricative				ɬ ɮ				ʎ̝	ʟ̝				
Lateral approximant				l		ɭ		ʎ	ʟ				
Lateral flap				ɺ		ɭ̆							

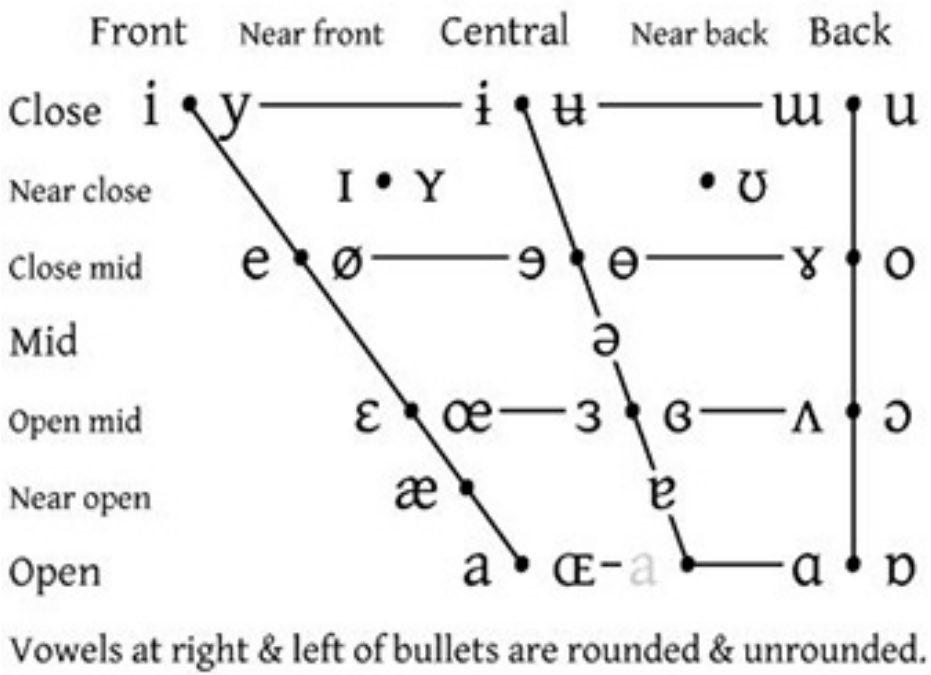
Where symbols appear in pairs, the one to the right represents a modally voiced consonant, except for murmured ɦ. Shaded areas denote articulations judged to be impossible. Light grey letters are unofficial extensions of the IPA.

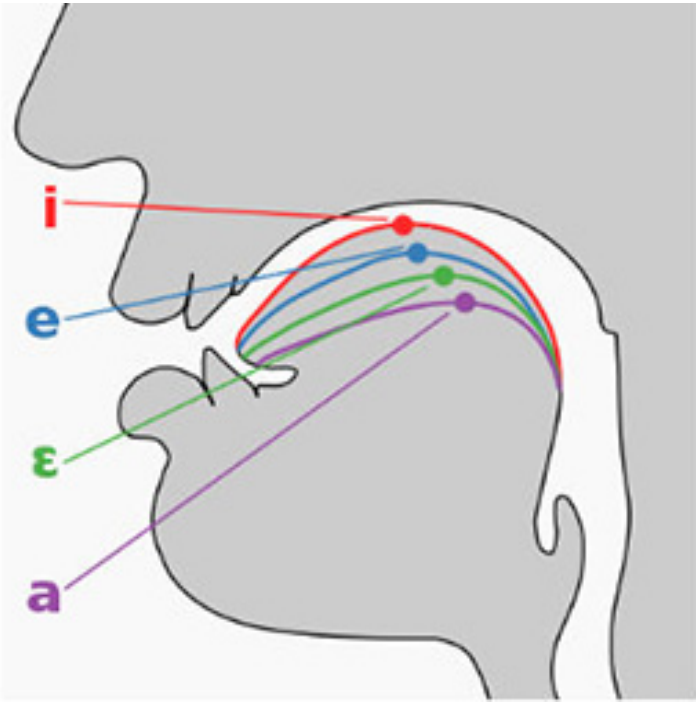
IPA chart of non-pulmonic consonants

CONSONANTS (NON-PULMONIC)

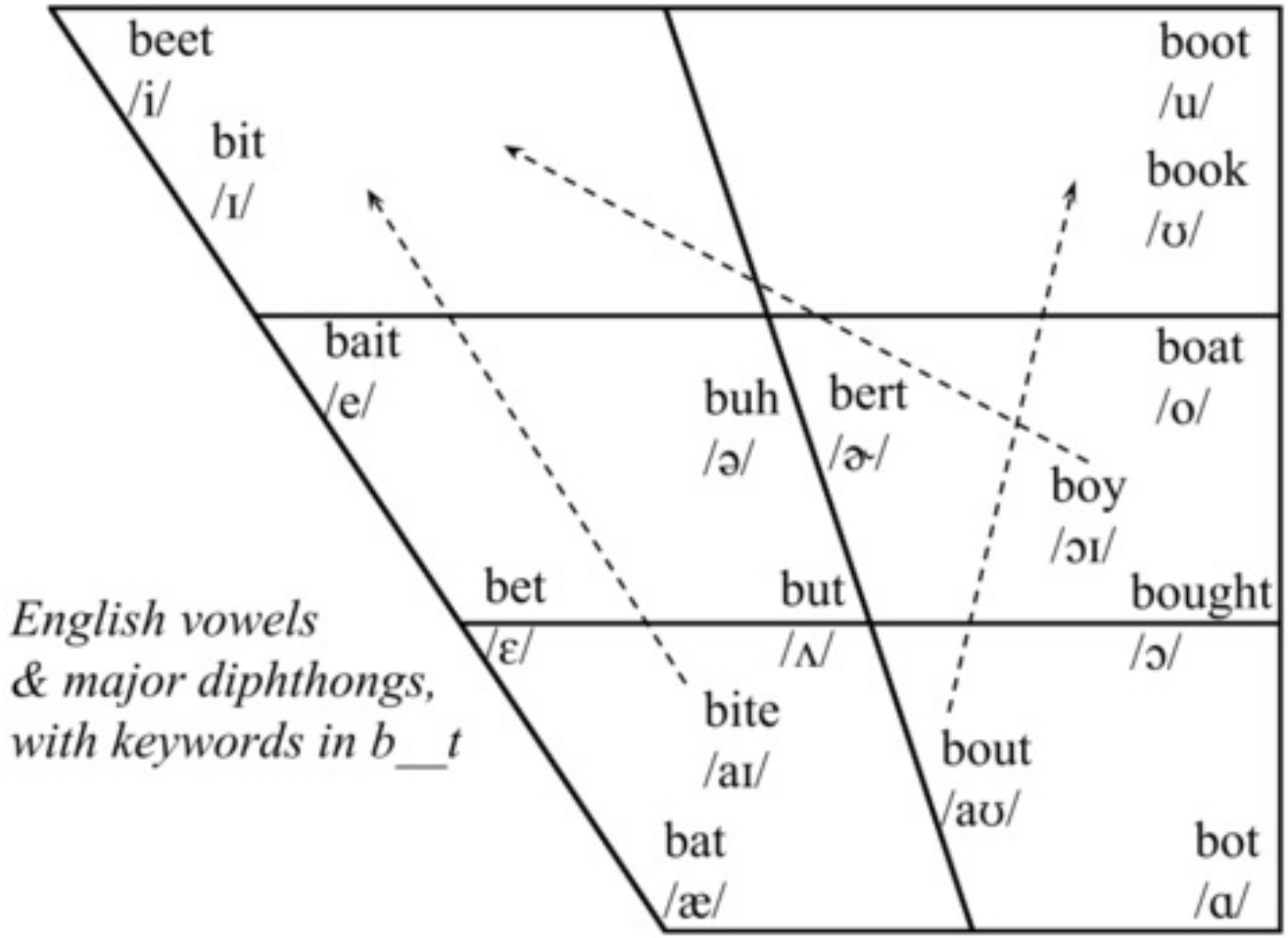
Anterior click releases (require posterior stops)	Voiced implosives	Ejectives
ʘ Bilabial fricated	ɓ Bilabial	ʼ *Examples:*
ǀ Laminal alveolar fricated ("dental")	ɗ Dental or alveolar	pʼ Bilabial
ǃ Apical (post)alveolar abrupt ("retroflex")	ʄ Palatal	tʼ Dental or alveolar
ǂ Laminal postalveolar abrupt ("palatal")	ɠ Velar	kʼ Velar
ǁ Lateral alveolar fricated ("lateral")	ʛ Uvular	sʼ Alveolar fricative

IPA vowel chart





Example: Vowels in English



dashed arrows for diphthongs

Tone systems, tonal languages (tonemes)

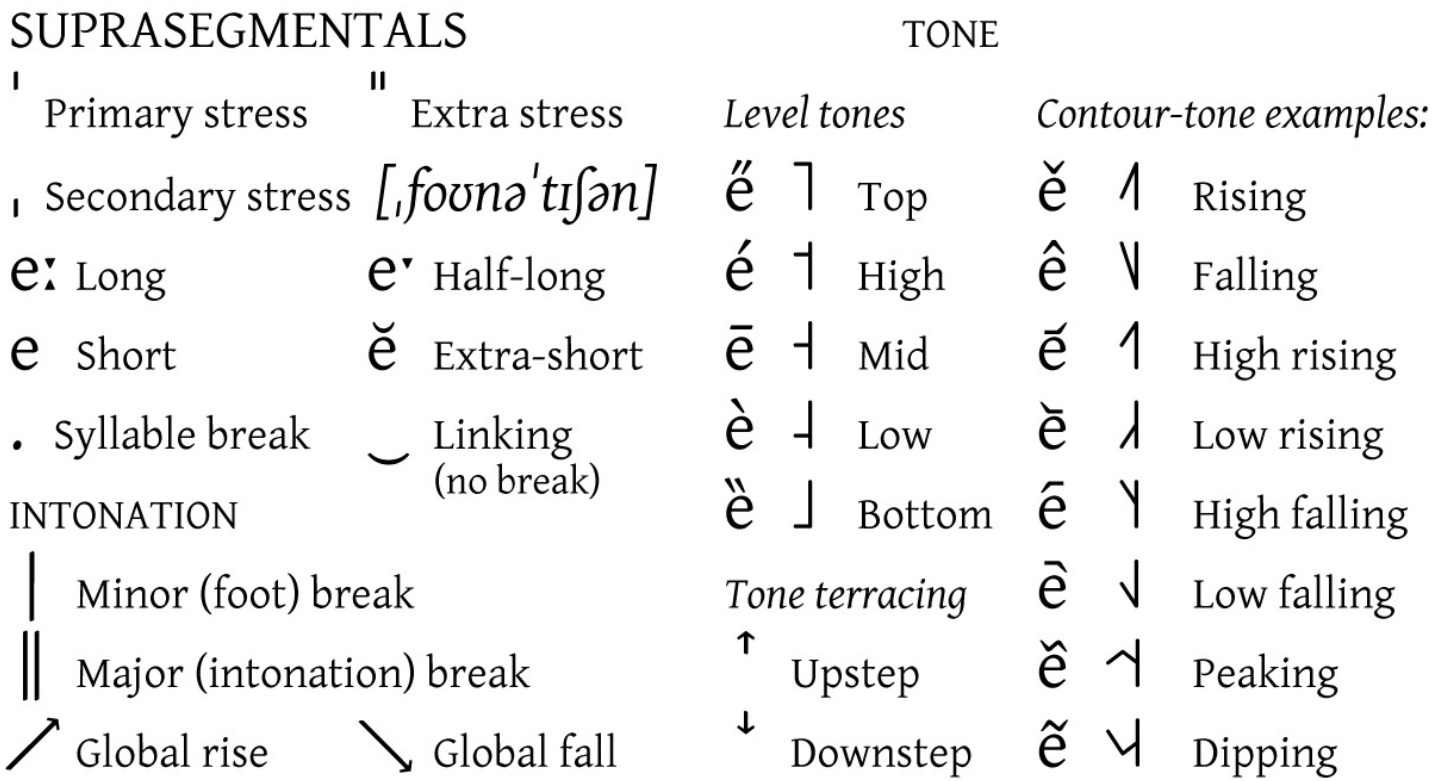
- Mandarin is the most widely spoken tonal language

妈 *mā* = *mom*
麻 *má* = *hemp*
马 *mǎ* = *horse*
骂 *mà* = *scold*
吗 *ma* = *question mark*

妈妈骂马的麻吗?

- A large number (up to 70%) of world languages are tonal
- *Register* tone systems (Niger-Congo languages): tones distinguished by relative pitch
- *Contour* tone systems (Sino-Tibetan languages): tones distinguished by shape;
- combined system (Cantonese) both shape and pitch
- grammatical information may be encoded in the tone

IPA diacritics and suprasegmentals



Other suprasegmental features (long-distance effects)

- Metaphony: general phenomena of assimilation of one type of vowel by another within a word
- Vowel harmony (progressive or regressive)

$$V_a C V_b C V_b C \mapsto V_a C V_a C V_a C$$

vowel of type V_a assimilates vowels of type V_b across intervening consonant segments
(Turkish, Hungarian, Korean)

Examples of Phonological Models of these phonetic structures

Autosegmental Phonology for tone systems, vowel harmony, ...

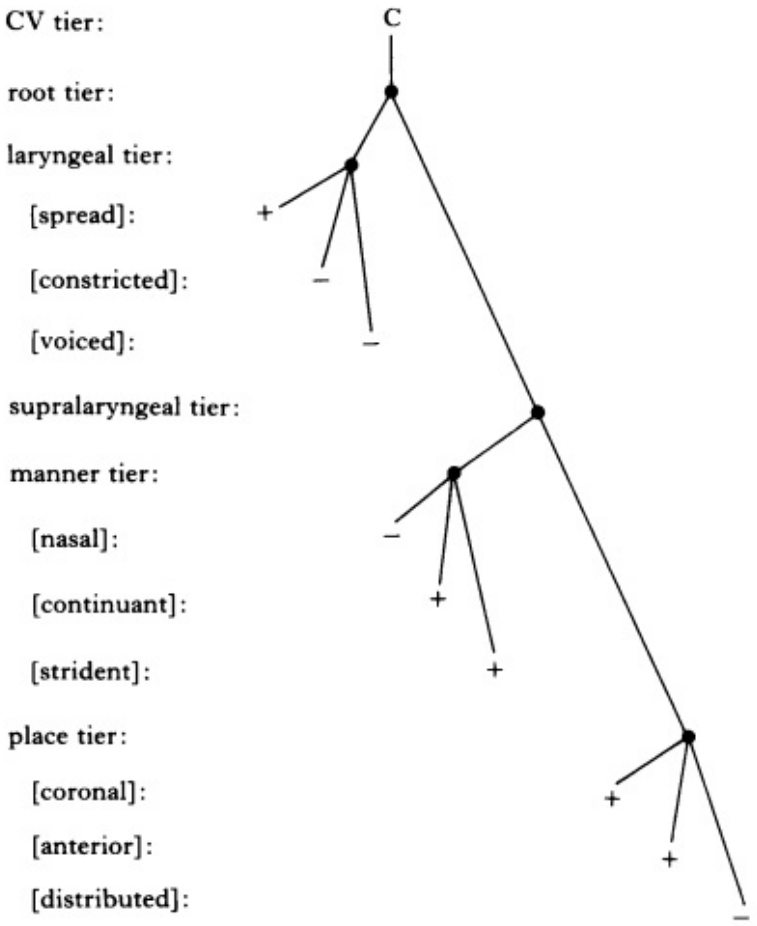
• Describe the segments of the sound structure by multiple tears, each a $+/-$ binary variable describing one feature
• Factoring phonemes into a product of *features*

- Segmental tier: $\pm$ sonorant, $\pm$ continuant, $\pm$ voice, ...
- Timing tier: lenghts of segments (long/short vowels)
- Stress tier: distribution of stress, main stress, ...
- Tone tier: $\pm$ high/low pitch, ...

• John Goldsmith, *Autosegmental and metrical phonology*, Basil Blackwell, 1990.

Feature Geometry

- Autosegmental structures as trees, geometric operations on trees



- G.N. Clements, *The Geometry of Phonological Features*, Phonology Yearbook, Vol.2 (1985) 225–252.

Example: Optimality theory in phonology

- generative phonology aims at predictive models of sound systems: phonological changes as a replication process: *faithfulness constraints* tend to maintain output like input; *markedness constraints* forbid occurrences of certain outputs; *optimality* by smallest number of violations

• Alan Prince and Paul Smolensky, *Optimality Theory: Constraint Interaction in Generative Grammar*, Blackwell Publishers, 2004.

Example: recent work on mathematical structures in phonology theory

- algebraic characterizations of phonological patterns
- mathematical tools: semigroups, formal languages, string-to-string transducers
- strictly local languages (membership decidable by k-width substrings used in describing phonotactic patterns
- how algebraic properties can determine grammatical inference

• Dakotah Lambert, Jeffrey Heinz, *An Algebraic Characterization of Total Input Strictly Local Functions*, Proceedings of the Society for Computation in Linguistics (SCiL) 2023, 25–34

Morphology

- word-forms (sing, sang, sings, singing, . . .)
- lexeme underlying "vocabulary-word", base-form, different word-forms of same lexeme
- morphological rules: two kinds
  - inflection rules (relate different forms of same lexeme): conjugation, declension
  - word formation (combine different lexemes): e.g. dishwasher
- word formation: two kinds
  - derivation: affixing bound-forms (sing-er, slow-ly, ...)
  - compounding: combines complete word forms (dish-washer)

- Some languages extremely rich in compound words, other poor
- Even within same language family huge differences

Curious example: among ancient Indo-European languages, Sanskrit and Ancient Greek are very rich in compound words (Homer's $\rho o \delta o \delta \alpha \kappa \tau \upsilon \lambda o \varsigma$ $H \omega \varsigma$)
but Hittite has no compound word formation at all

- paradigm: the set of all word-forms associated to a given lexeme

Examples:
- conjugation of verbs (tense, aspect, mood);
- declension of nouns (number, gender, case);
- personal pronouns arranged by person, number, gender

Allomorphy

- morpheme: smallest grammatical units, roots and affixes
- allomorphs: different morphemes playing same grammatical role

Example: negation prefixes in English
- *a-*, *an-* (from Greek): anesthesia, anisotropic, acyclic
- *in-*, *im-* (from Latin): impossible, incompressible, invincible
- *un-* (English): unbiased, unaffected, unacceptable

Example: different forms of plural in English
boy $\mapsto$ boys; watch $\mapsto$ watches;
child $\mapsto$ children; woman $\mapsto$ women

Example: strong verbs (sleep/slept)

- phonological allomorphs: regular phonological rules
- suppletive allomorphs: exceptional

Morphological Typology

Grouping languages by morphological structures

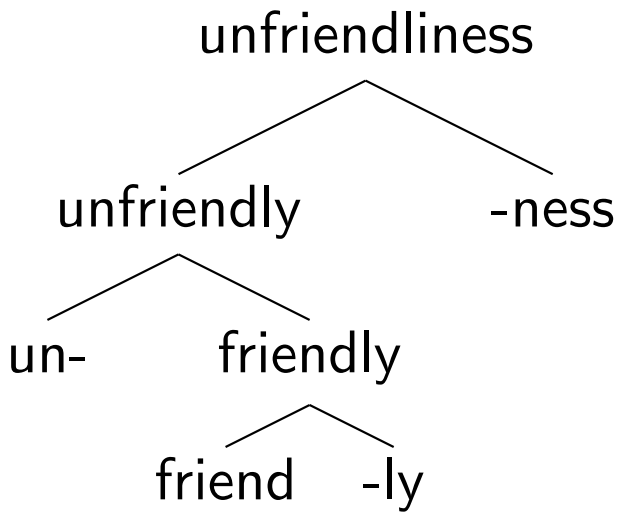
• Analytic: small amount of inflection, replaced by word order and additional word (Mandarin)
• Isolating: few morphemes per word (Vietnamese)
• Synthetic: typically several morphemes can combine in words, high in inflection forms (many Indo-European languages)
Examples:
- German: *Abstimmungsbekanntmachung*
- Russian: Достопримечательность

• Polysynthetic: extremely long compound words with *sentence-words*
(America, Australia, Siberia, Papua New Guinea)
- compositionally polysynthetic, affixally polysynthetic
- incorporating, agglutinating, fusional

Hierarchical structures

- *bracketing* (as in non-associative algebra)

$$\text{unfriendliness} = (\,(\,\text{un-}\,(\,(\,\text{friend}\,)\,\text{-ly}\,)\,)\,\text{-ness}\,)$$



Lexicology (and lexical semantics)

• *diachronic*: changes across time in the use of words and word formation

• *synchronic* (Structuralist): lexical relations (at a given time), syntagmatic lexical relations (culturally determined patterns of association between lexical units)

• various WordNet lexical and semantic databases

• Phraseology: *phrasemes*= multi-word lexical units, includes study of idiomatic expressions (e.g. "it's raining cats and dogs")

• Etymology: origin and history of words, crucial role in historical linguistics: comparative methods, reconstruction of proto-languages

Note: morphology and lexicology are most of what one sees in traditional *grammars* of specific languages, aspects of syntax (rules for the formation of sentences) are also included, but *syntax* is a much richer object of study than what usual introductions to the grammar of individual languages suggest

Syntax the large-scale structure of languages

- the basic units of structure at this level are *sentences*
- rules and principles governing sentence structure (within a language, or across languages)
- origin of scientific syntactic theory: 4th century BCE
अष्टाध्यायी (*Aṣṭādhyāyī*) of पाणिनि (Pāṇini)
- origin of "traditional grammar": 2nd century BCE
Διονύσιος ὁ Θρᾷξ Τέχνη γραμματική
- Dionysius Thrax's Techne was a primarily morphological grammar, little emphasis on syntax, while Pāṇini focused on all aspects (phonology, morphology, syntax): basis of modern syntactic theory

Modern Syntactic Theory:

• i-language versus e-language: internal language (mental) as opposed to external (community based records of language use): focus on i-language as object of study

• grammaticality: judgement on whether a sentence is well formed (grammatical) in a given language, i-language gives people the capacity to decide on grammaticality

• generative grammar: produce a set of rules that correctly predict grammaticality of sentences

• universal grammar: ability to learn grammar is built in the human brain, e.g. properties like distinction between nouns and verbs are universal, ... is universal grammar a falsifiable theory?