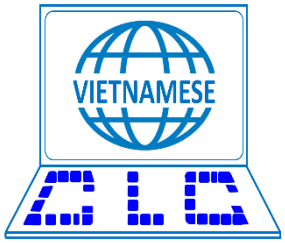
University of Science

Computational Linguistics Center

Introduction to Natural Language Processing

Section 1:
Introduction to Natural Languages



Lecturer: Assoc.Prof. Dr. Dinh Dien

LANGUAGES IN THE WORLD

- Differ: Natural Languages (e.g. Vietnamese, English, French, etc.) vs. Artificial Languages (e.g. C, Pascal,…; Morse; Braille; etc.
- From now: language = natural language.
- How many different (natural) languages are there in Vietnam ?
- ~ 54 (Vietnamese and 53 ethnic languages)
- How many different languages are there in the world ?
- ~ 7015 !
- Distribution of users: very unequally (100M vs. <100)

LANGUAGE POPULATION

Rank	Language Name	Primary Country	Population
1	CHINESE, MANDARIN	China	885,000,000
2	SPANISH	Spain	332,000,000
3	ENGLISH	United Kingdom	322,000,000
4	BENGALI	Bangladesh	189,000,000
5	HINDI	India	182,000,000
6	PORTUGUESE	Portugal	170,000,000
7	RUSSIAN	Russia	170,000,000
8	JAPANESE	Japan	125,000,000
9	GERMAN, STANDARD	Germany	98,000,000
10	CHINESE, WU (Ngô)	China	77,175,000
11	JAVANESE	Indonesia, Java, Bali	75,500,800
12	KOREAN	Korea, South	75,000,000
13	FRENCH	France	72,000,000
14	*VIETNAMESE*	*Vietnam*	*67,662,000*
15	TELUGU	India	66,350,000
16	CHINESE, YUE (Việt)	China	66,000,000

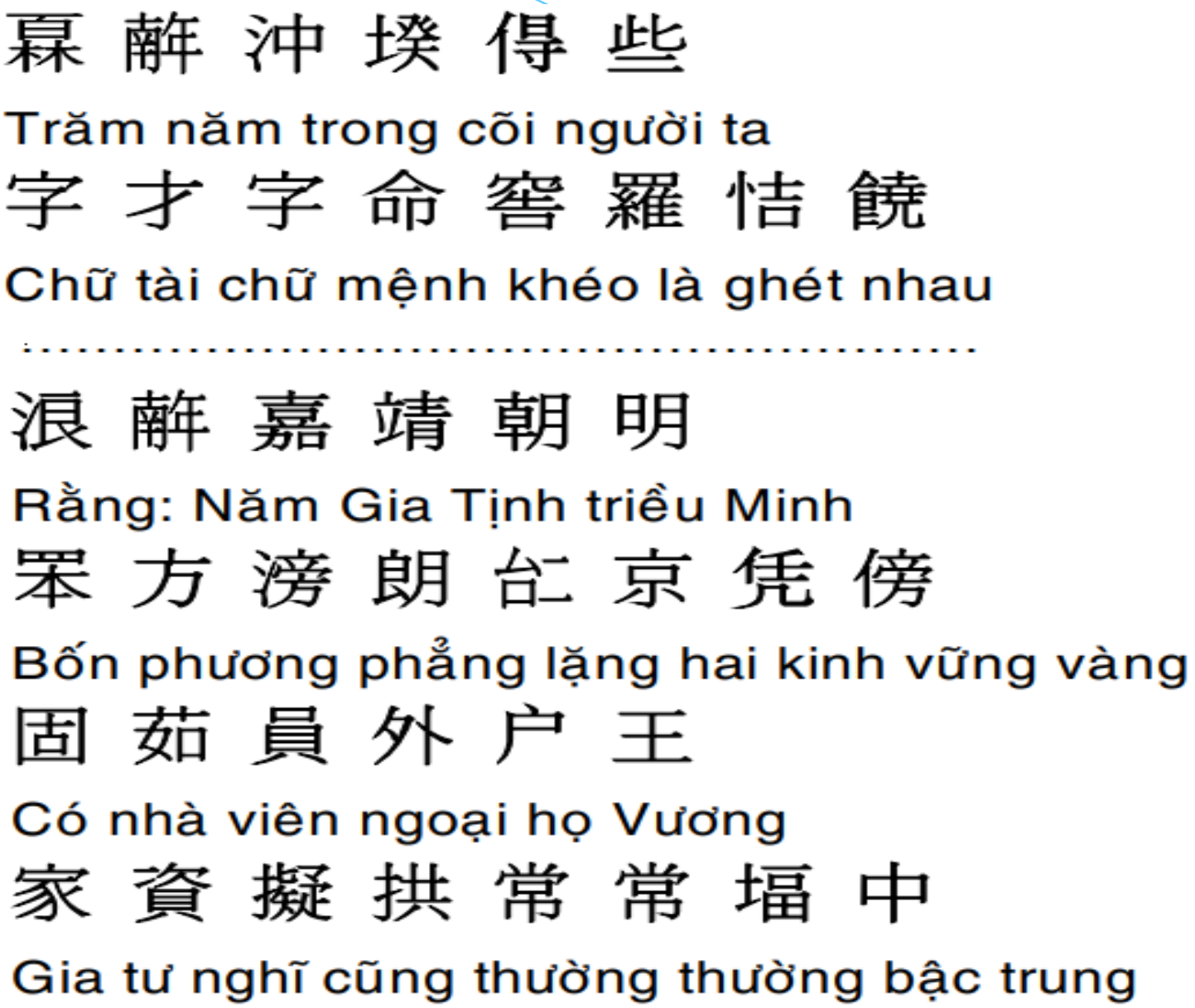
Endangered Languages

- Vanishing Languages:
- One language dies every 14 days.
- By the next century nearly half of the roughly 7,000 languages spoken on Earth will likely disappear,
- as communities abandon native tongues in favor of English, Mandarin, or Spanish.
- What is lost when a language goes silent?
- Cultural treasures, historical lessons, mankind knowledge, etc.

THE ORIGIN OF LANGUAGES

- Who invented English?
- Who invented Vietnamese?
- Differ: voice (natural) vs. writings (manmade)
- Only popular languages have writings.
- Vietnamese writings: who invented ?
- before 10th century: has no (using Chinese writings);
- from 10th-19th century: using Nôm writings (borrow Chinese characters: one for sound, one for meaning),
- e.g.: 爸(*ba/father*), 巴(*ba,3*), 罘(*4*), 靡(*year*), 甪(*5*), …
- Most famous works (in ~ 1000 years) written in Nôm: literature (*The Tale of Kiều*) / history/culture/ Agriculture/ Geography/ Traditional Medicine (*Hải Thượng Lãn Ông*, *Tuệ Tĩnh*)/…

The Tale of Kiều in Nôm

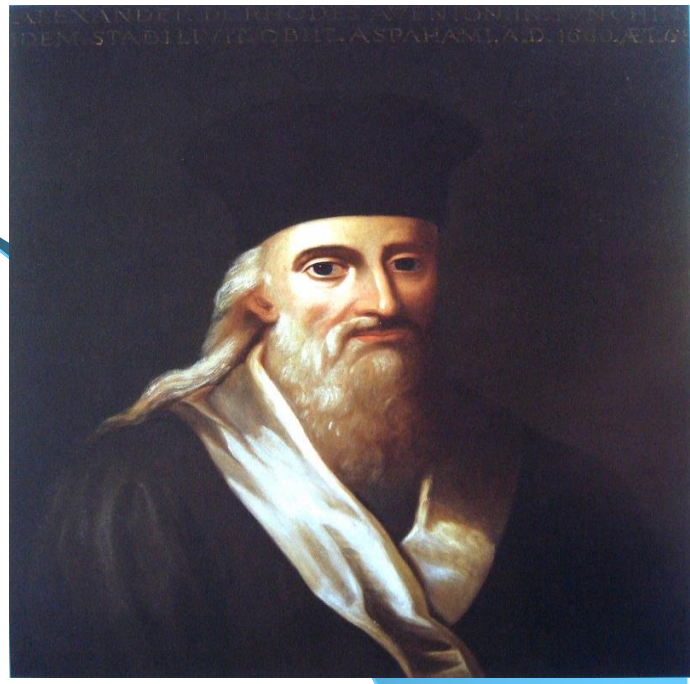


WRITINGS = artificial

- Vietnamese alphabet:
- since 19th century – till now:
- from Latin-letters + diacritics: *ba*, *bốn*, *năm*,
- Latin alphabet: from Greek (Y);
- Greek from Phoenician …

Α Alpha	Β Beta	Γ Gamma	Δ Delta	Ε Epsilon	Ζ Zeta
Η Eta	Θ Theta	Ι Iota	Κ Kappa	Λ Lambda	Μ Mu
Ν Nu	Ξ Xi	Ο Omicron	Π Pi	Ρ Rho	Σ Sigma
Τ Tau	Υ Upsilon	Φ Phi	Χ Chi	Ψ Psi	Ω Omega

α Alpha	β Beta	γ Gamma	δ Delta	ε Epsilon	ζ Zeta
η Eta	θ Theta	ι Iota	κ Kappa	λ Lambda	μ Mu
ν Nu	ξ Xi	ο Omicron	π Pi	ρ Rho	σ Sigma
τ Tau	υ Upsilon	φ Phi	χ Chi	ψ Psi	ω Omega



Alexandre de Rhodes (1591-1660)

WRITINGS = artificial

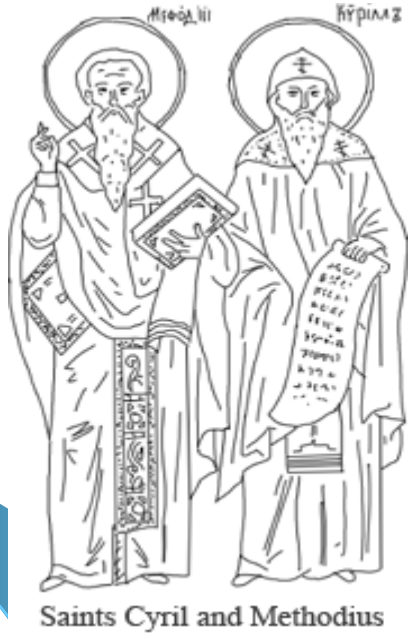
Russian alphabet: from Cyrillic; Greek

Аа Бб Вв Гг Дд Ее Ёё Жж Зз Ии

Йй Кк Лл Мм Нн Оо Пп Рр Сс Тт

Уу Фф Хх Цц Чч Шш Щщ Ъъ Ыы

Ьь Ээ Юю Яя



Saints Cyril and Methodius

1000 AD

Мы учим язык *(We are learning a language)*

Korean alphabet:

우리는 언어를 배우고 있어요

한국 (H-a-n k-u-k) 중국 (Tr-u-ng k-u-k)

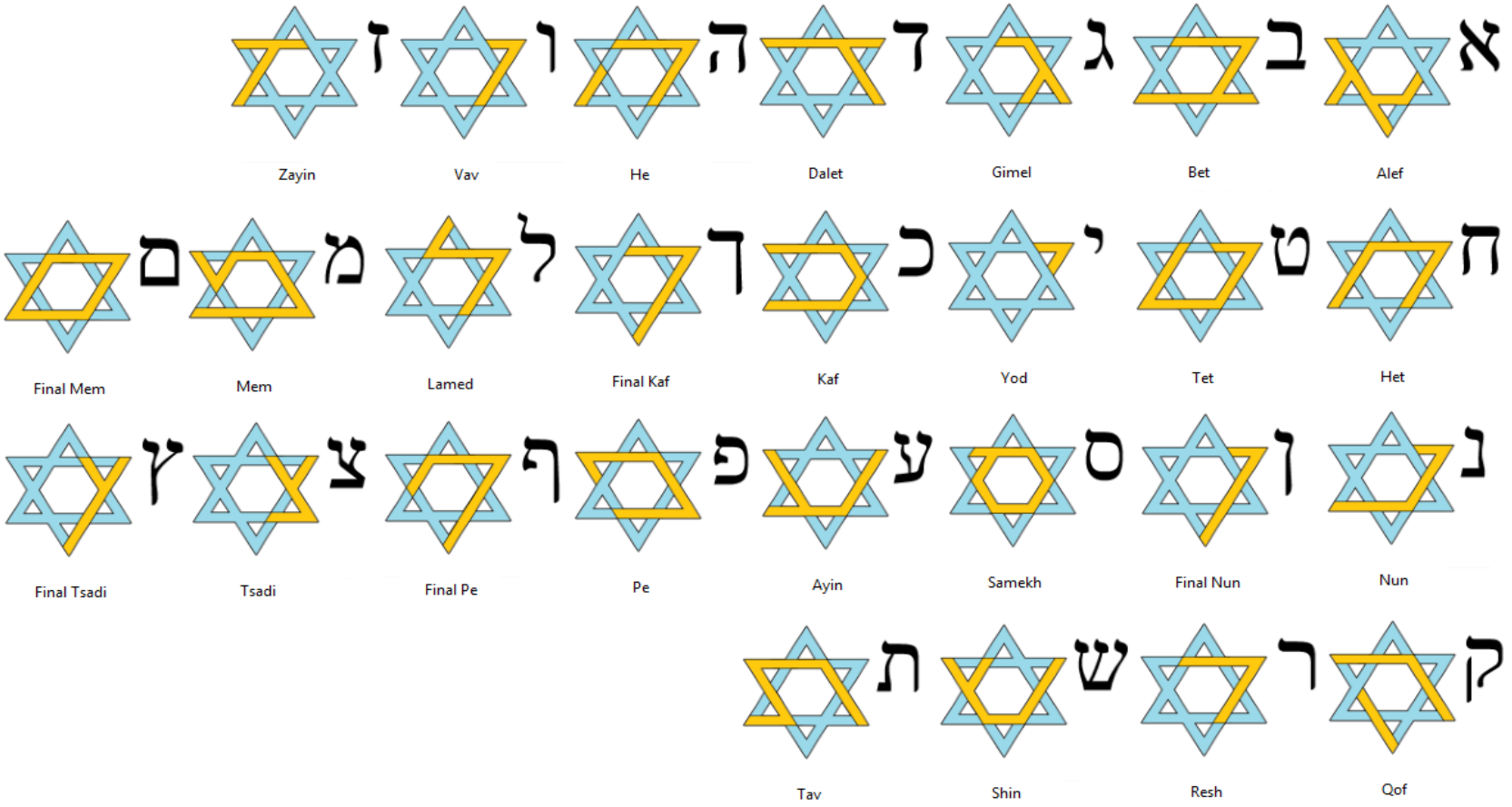
미국 (M-i k-u-k) 삼성 (S-a-m S-o-ng)

King Sejong (1397; 1418-1450; 1443)

Hebrew alphabet:

tấm khiêng/ngôi sao David:

ירושלים (yerushalayim)



THE ORIGIN OF LANGUAGES

- [Genesis]: At the beginning, the whole world had one language and a common speech, settled in the same land named Shinar.
- As the population was growing, they decided to build a tall "reach to the heavens", proud symbol of how great they had made their nation.
- God did not like the pride and arrogance, God caused the people to suddenly speak different languages so they could not communicate and work together to build the tower.
- This caused the people to scatter across the land with different languages as nowadays.
- The tower was named The Tower of Babel because the word Babel means “confusion”.

The Tower of Babel

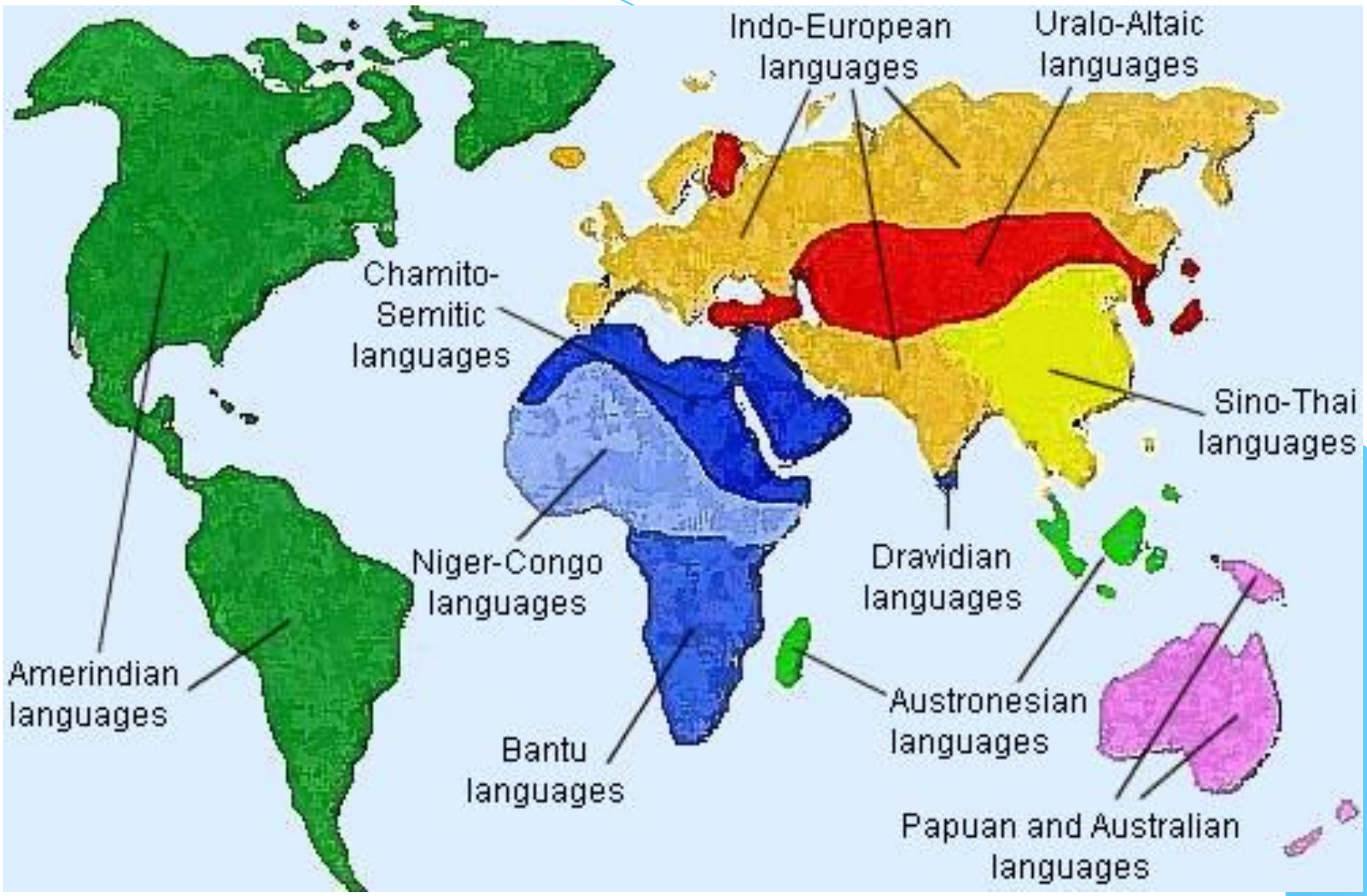
- This is only a legend !

THE CLASSIFICATION OF LANGUAGES

There are 2 main kinds of **classification** of **languages**:
(1) genetic (or genealogical) and (2) typological.
The purpose of genetic **classification** is to group **languages** into families according to their degree of diachronic relatedness
There are 5 genres of languages:

1. Indo-Euro: India, Iran, Bantic, Slave, Roman, Greek, German (German, ***English,*** Dutch,...).
2. Semitic: Semit, Egypt, Kusit, Becbe,...
3. Turkish: Turkish,
4. Sino-Tibetan: Sino (Chinese), Tibeto-Burman,...
5. Austro-Asia: : Nahali, Munda, Nicoba, Mon-Khmer. Mon-Khmer branch: Viet-Muong group; Viet-Muong group: Muong and ***Vietnamese***.

THE CLASSIFICATION OF LANGUAGES: GENRE



THE CLASSIFICATION OF LANGUAGES: GENRE

Indo-Euro: India, Iran, Bantic, Slave, Roman, English, ...

Typological Classification of Languages

Definition

- Languages are described by their *types* rather than by their origins and relationships
- The type under which languages are classified follows morphological classification

Language Types

1. Isolating
2. Agglutinating/agglutinative
3. Inflecting/flectional/fusional
4. Polysynthetic/incorporating

Isolating languages

- One-to-one correspondence between words and morphemes
- One word formations
- Free morphemes are the only forms used
- The “word” (free morpheme) can occur by itself and is not dependent on any other morphemes.

Isolating languages

- It is an unalterable unit whose function in the sentence is not usually marked by some grammatical device (affix, auxiliary) but only by position.
- Since the boundaries of syllables and morphemes *coincide*, these languages are sometimes referred to as monosyllabic.

Isolating Languages

- Examples: Chinese, Vietnamese, Thai, Laos, and many languages of South East Asia
- Ex (Chinese):我看他 *wo kan ta*

  “I see him”; “I am seeing him”

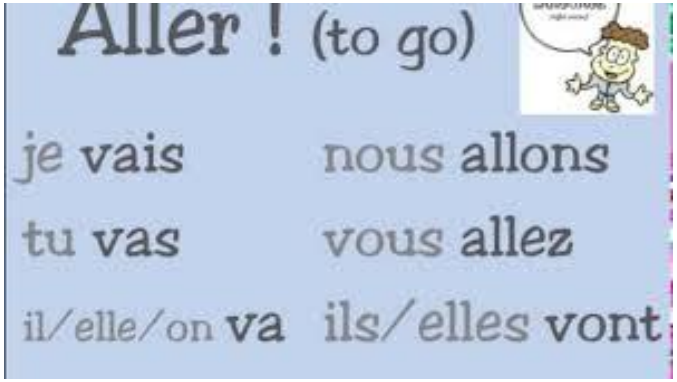
  他看我朋友 *Ta kan wo peng you*

  “He sees my friend”

Flexional/Fusional/Inflecting Languages

- Grammatical devices like *affixes* or internal changes in words to show grammatical relationships
- Free and bound morphemes are united
- Ex. Walk, walk-s, walk-ing, walk-ed
- Internal change: mouse-mice
  goose-geese

- Inflections of French, Russian, Latin:



THE SIX (6) TENSES OF THE LATIN "STATE OF BEING" VERB
sum (esse, fuī, futūrus)

	PRESENT	IMPERFECT	FUTURE	PERFECT	PLUPERFECT	FUTURE PERFECT
Singular						
1st Person	su m *I am*	era m *. . . was*	er o *. . . will be*	fu ī *. . . have been/was*	fuera m *. . . had been*	fuer o *. . . will have been*
2nd Person	e s *you are*	erā s *. . . were*	eri s *. . . will be*	fu istī *. . . have been/were*	fuera s *. . . had been*	fueri s *. . . will have been*
3rd Person	es t *he/she/it is*	era t *. . . was*	eri t *. . . will be*	fu it *. . . has been/was*	fuera t *. . . had been*	fueri t *. . . will have been*
Plural						
1st Person	su mus *we are*	erā mus *. . . were*	eri mus *. . . will be*	fu imus *. . . have been/were*	fuera mus *. . . had been*	fueri mus *. . . will have been*
2nd Person	es tis *you are*	erā tis *. . . were*	eri tis *. . . will be*	fu istis *. . . have been/were*	fuera tis *. . . had been*	fueri tis *. . . will have been*
3rd Person	su nt *they are*	era nt *. . . were*	eri nt *. . . will be*	fu ērunt *. . . have been/were*	fuera nt *. . . had been*	fueri nt *. . . will have been*

	1st person	2nd person	3rd person (masc.)	3rd person (fem.)	3rd person (neut.).
English	*I, Me*	*You*	*He, Him*	*She, Her*	*It*
Nominative Case	Я	Ты	Он	Она	Оно
Accusative Case	Меня	Тебя	Его	Её	Его
Genitive Case	Меня	Тебя	Его	Её	Его
Dative Case	Мне	Тебе	Ему	Ей	Ему
Instrumental Case	Мной	Тобой	Им	Ей	Им
Prepositional Case	Мне	Тебе	Нём	Ней	Нём

Flexional/Fusional/Inflecting Languages (2)

- Several units of meaning are contained within a single world
  - Latin, ***ib*** “I shall go” (base: *i* “go”; *-b(i)-* is the future tense morpheme; *-* *-* ‘is the first person singular
  - Sanskrit ***vad mi*** (*vad-* the base ‘speak’; *(a)mi)* ‘first person singular’

Agglutinating/Agglutinative Languages

- A type of flexional language with the exception that the morphemes attached have a separate existence (= free morpheme)
- Implication: the boundaries between the morphemes are always clear because their shape remains the same

Agglutinating/Agglutinative Languages: Example

- Turkish *adam* ‘man’
  - nominative: adam (sg) adam-lar (pl)
  - accusative: adam-i (sg) adam-lar-i (pl)
  - genitive: adam-in adam-lar-in (pl)
  - dative: adam-a adam-lar-a
  - locative adam-da adam-lar-da
  - ablative adam-dan adam-lar-dan

Agglutinative vs Flexional

Hungarian

- Nom. *su* “water”
- Gen. *su-num*
- Acc. *su-yu*
- Abl. *su-dan*

Latin

- *aqua*
- *aquæ*
- *aquam*
- *aqu*

Japanese *tabesaserarareru*

- *tabe* “eat” (the base)
- *sase* “the causative element (i.e. to cause someone to do something)
- *rare* “the passive form”
- *ru* “the infinitive”

Polysynthetic/Incorporating Languages

- These languages make use of affixation and often incorporate what English would represent with nouns and adverbs.
- The word forms are often very long and morphologically complex
- Languages: Inuktitut (Baffin Island Eskimo), Oneida)

Polysynthetic/Incorporating Languages (2)

- *g-nagla-sl-i-zak-s*
  - ***g*** **“I” (first person)**
  - ***nagla*** **(conveys idea of) “living”**
  - ***sl*** **(causes *nagla* to be noun-like; the combination conveys the idea of “village”)**
  - ***i*** **verbal prefix, indicates that *zak* is to carry a verbal idea**
  - ***zak*** **‘look for’**
  - ***s*** ***‘continued action***

Polysynthetic/Incorporating Languages (3)

- *ngirruunthingapukani*
- I past for some time eat repeatedly

Polysynthetic/Incorporating Languages (4)

- Tavva- -guuq ikpiarju(q) -ku(t)
  Then (suddenly) they say work-bag by
- -Luni- tigualaka -mi -uk takanu-
  while she swept up (loc) by (poss) that one
  (in one motion) there below
- -nga ikijaq- tuq- Luni quja(q)r- mun
  her way out she kayak towards

- "Then (suddenly) she swept up (poss) work-bag that one there below her she way out towards kayak"

Non-exclusivity

- None of these four types are mutually exclusive.
- In English, there is a movement towards a more isolating type of structure.
- Yet, all elements appear in English.

English

- Isolating: *The boy will ask the girl.*
- Inflecting: *The biggest boys will be asking all the girls to the party.*
- Agglutinating: *anti-dis-establish-ment-arian-ism*
- Incorporating: *"whacchamacallit"*

  "This is the *whatchamacallit.*

An aggluinating example: *Antidisestablishmentarianism*

- *establish (9)*
  - to set up, put in place, or institute (originally from the Latin *stare*, to stand)
- *dis-establish (12)*
  - ending the established status of a body, in particular a church, given such status by law, such as the Church of England
- *disestablish-ment (16)*
  - the separation of church and state (specifically in this context it is the political movement of the 1860s in Britain)
- *anti-disestablishment (20)*
  - opposition to disestablishment
- *antidisestablishment-arian (25)*
  - an advocate of opposition to disestablishment
- *Antidisestablishmentarian-ism (28)*
  - the movement or ideology that opposes disestablishment

Typology: morphology: The fusional/inflectional type

- A fusional/inflectional word contains several morphemes which indicate grammatical categories.
  - *Ein kleiner Hamster* "a little hamster" (nominative case)
  - *Der kleine Hamster* "the little hamster" (nominative case)
  - *Ich sah den kleinen Hamster* "I saw the little hamster" (accusative case)
  - *Mit kleinem Hamster* "with little hamster" (dative case).

Word Order Typology: syntax

Ví dụ: I eat rice Tôi ăn cơm

S V O S V O

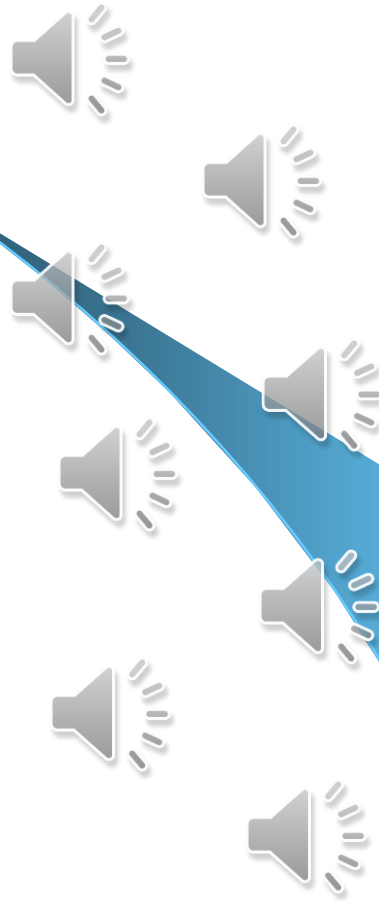
1. SVO: 32.4 - 41.8 %, e.g. **English**, Chinese, French, **Vietnamese**, Thai, Bulgaria, ...
2. SOV: 41 - 51.8 %, e.g. Japanee, Korean, Mongolian, Turki, Eskimo,...
3. VOS: 9 - 18 %, e.g. Cakchiquel (Guatemala), Huave (Mexico),...
4. VSO: 2 - 3 %, e.g. Tagalo, Egypt(old), Hebrew (Bible), Ireland,...
5. OVS: 1 % , e.g. Apalai (Brazil), Barasano (Columbia), Panare (Venezula),..
6. OSV: 1 %, e.g. Apurina, Xavante (Brazil),..

word order

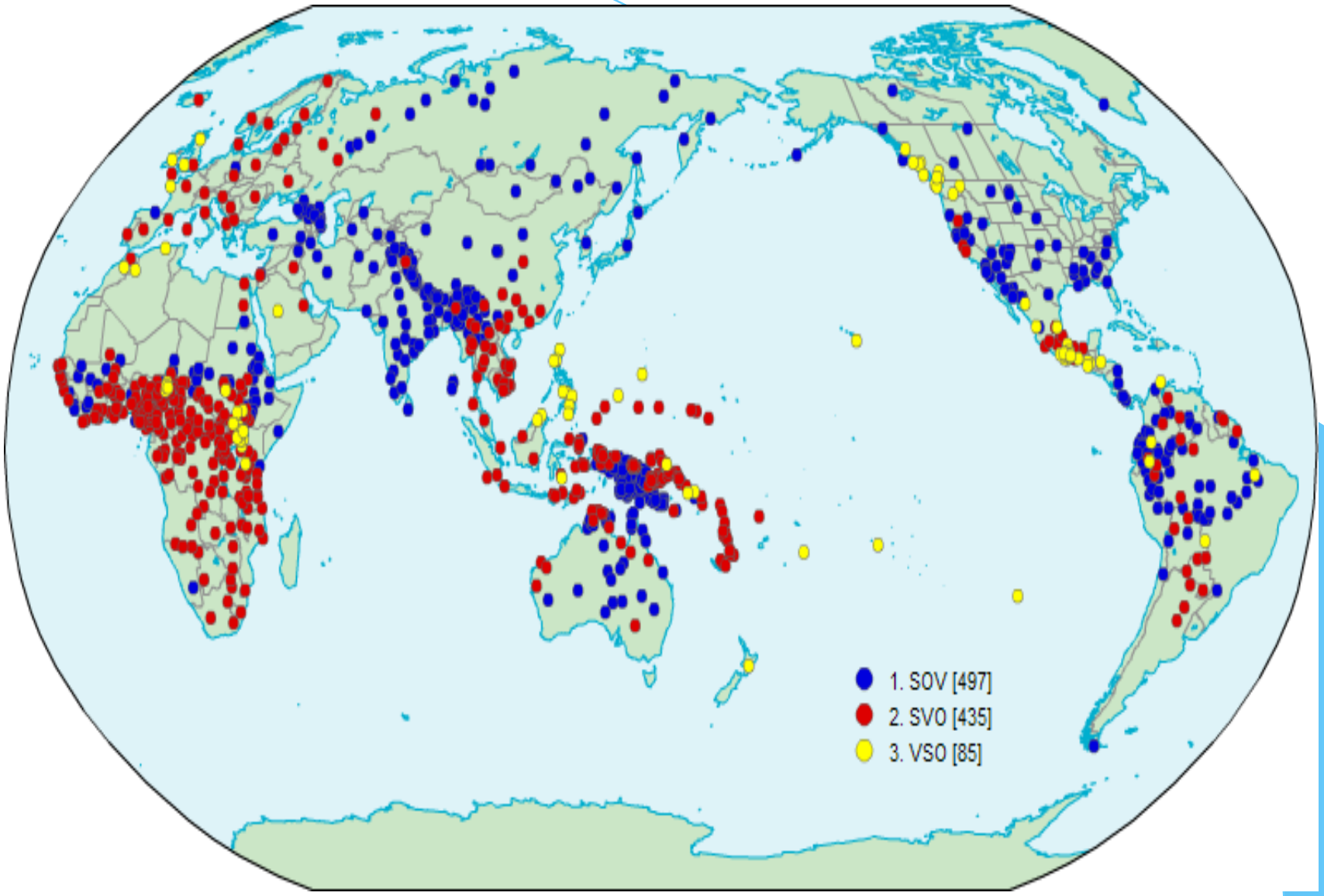
- I read a book
- 나는 책을 읽고;
- 私は本を読みます

Word Order

- We are learning a language.
- Nous apprenons une langue.
- 我们 学习 一门 语言。
- 言語を 習います。
- 우리는 언어를 배우고 있어요.
- Wir lernen eine Sprache.
- Мы учим язык.
- Ni lernas lingvon.



Word Order typology map



Characteristics of Natural Languages

- NL is a social phenomenon, not a natural phenomenon, or personal or biology (not hereditary).
- NL is the most important means of communication between humans.
- NL is the special semiotic system: differ: “signifier” (sound/image) vs. “signified”(concept).
- Ex: in the traffic signal system: “red light” (signifier) => stoppage (signified).
- Ferdinand de Saussure: “NL seems like a chess-board”. The value of each chessman is regulated by the system of the chess-board.
- => The meaning of a word is dependent on the context.

Vietnamese characteristics

- Vietnamse is the isolated language typology.
- Vietnamese words have no inflections. The grammatical meaning is outside the word. Ex: *Tôi nhìn anh ấy* vs. *Anh ấy nhìn tôi (I see him* vs. *He sees me)* .
- Grammatical methods are: *word order* and *function words*. Ex*: Gạo xay vs. Xay gạo ; đang học* vs. *học rồi (learning* vs. *learned).*
- There are a special linguistic unit: morpho-syllable ("hình tiết") whose its phonetic-cover exactly coincides with its syllable (âm tiết), and morpheme (hình vị) aka "tiếng".

CHARACTERISTICS OF VIETNAMESE LANGUAGE

- The word boundary is ambiguous (not delimited by space as flexional typology languages). Ex: “học sinh học sinh học” (pupils learn biology).
- => Morphological analysis becomes difficult.
- Word Segmentation is the pre-requisite for next moduls, e.g: spelling checker, POS tagger, word frequency, …
- There is a special classifier which go accompanied with nouns, e.g. : *cái* bàn, *cuốn* sách, *bức* thư, *con* chó, *con* sông, *vì* sao, ...(same phenomena in Chinese).

CHARACTERISTICS OF VIETNAMESE LANGUAGE

- In the phonetics aspect, Vietnamese is the tone language. Each syllable carries 1 of 6 following tones: no mark (ngang); acute (sắc), breve (huyền), question mark (hỏi), tilde (ngã) and dot below (nặng).
- This is suppra-segmental phoneme (âm vị siêu đoạn tính).
- Reduplicative words: *lấp lánh, lung linh, ..*
- Spoonerism (nói lái): by exchanging the initial consonant and the nucleus and/or the tone-mark between 2 syllables within a word due to their loose links, e.g. *hiện đại -> hại điện, thầy giáo -> tháo giầy,…*

CHARACTERISTICS OF ENGLISH LANGUAGE

- English is the flexional language typology with following characteristics:
- In the running texts, the word will be inflected.
- The grammatical meaning is inside the word.
- E.g.: *I see him* vs. *He sees me.*
- Grammatical methods are: suffix. Ex: *learning* vs. *learned.*
- Word formations are: affix. Ex: anticomputerizational (anti-compute-er-ize-ation-al).
- The morpheme boundary is ambiguous.
- The word boundary is clear (delimited by space or punctuation marks).

ENGLISH – VIETNAMESE COMPARISON

- Due to language/cultural typology (English-VNese comparative/contrastive linguistics)=> many differences.
- E.g.:in phonetics: English (no tone), Vietnamese (tone)
- Word boundary; lexicalization: e.g.: ox – bò đực, anh – elder brother , "carry out" -> "thực hiện";…;
- Part-Of-Speech: "thank you for your attention/N" ("cám ơn các bạn đã lắng nghe/V")
- Word order: "head-initial" vs. "head-final"; "pre-position" vs, "post-position". Ví dụ: "A prety new green dress" vs. "một cái áo dài mới đẹp màu xanh";

⇒ "Didn't we learn this lesson *yesterday*? *Hôm qua*, mình đã không học bài này sao ?"; "They went home *quietly*." "Họ *lặng lẽ* về nhà." or "Họ về nhà *một cách lặng lẽ*" (Adverb positions: Manner-Place-Time).

A little bit of Esperanto

- Artificial Language invented by Zamenhof in 1887
- Unambiguous
- Easiest language => save time to learn other Indo-Euro lang.
- Several areas in Central European uses as a native language.
- Examples:
- mi kaj vi
- Mi amas vin
- Mi kaj amiko
- Mi amas amikon
- Mi amas amikinon
- Mi amas amikinon belan
- Mi ne amas amikinon malbelan
- Amikinon malbelan ne amas mi