What is Linguistics? Part III

Matilde Marcolli

CS101: Mathematical and Computational Linguistics

Winter 2015

Language Change and Evolution

- Languages are dynamical, diachronic evolution
- Historical Linguistics deals with reconstructing the history of language evolution
- identify mechanisms of language change
- construct models of language evolution

- Mechanisms: sound change, borrowing, semantic and lexical change, morphological and syntactic change, grammaticalization
- Methods and models: comparative linguistic reconstruction, phylogenetic tress, wave theory

... What we know from Historical Linguistics

Sound change

- Regularity principle: sound change is regular (Neogrammarian hypothesis)

if a sound change happen in a Language it happens everywhere where a certain rule applies

Example: Latin to Spanish $p \mapsto b$, $t \mapsto d$ and $k \mapsto g$ when in between two vowels (lenition, sonorization): *vita/vida*, *lupa/loba*, *caeca/ciega*

- Unconditioned/conditioned sound change (context independence)

Phonemic changes

- Merger: $(X_1, X_2) \mapsto X_2$ or $(X_1, X_2) \mapsto X_3$

- Example: in Latin American Spanish
$\ell\ell$ and y phonemes merge to y
- Example: in Sanskrit: e and o merge into a
proto-Indo-European *agro*, Latin *ager*, Ancient Greek ἀγρός
becomes Sanskrit अज्र *ajra*, field

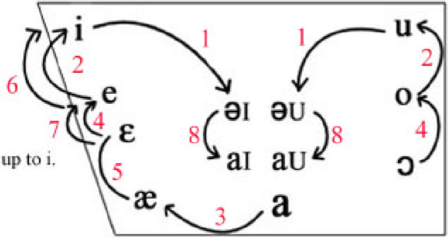
- Merger is irreversible

- Split (Umlaut): responsible for phenomena like *mouse/mice* or *foot/feet*, transition $u \mapsto \overline{y} \mapsto \overline{i}$

- Contact assimilation: Latin to Italian *somnium* $\mapsto$ *sonno*; *noctem* $\mapsto$ *notte*
- Deletions and Insertions: Latin to Spanish *apoteca* $\mapsto$ *bodega*; German *Landsknecht* borrowed in French as *lansquenet* (inserted vowel)
- Other sound changes: rhotacism (s/r); metathesis (transposition of two sounds: *brid*/*bird* Old/Modern English); final devoicing, intervocalic voicing, palatalization (k/č), vowel rising/lowering...

The Great Vowel Shift in English

Step 1: i and u drop and become əɪ and əʊ
Step 2: e and o move up, becoming i and u
Step 3: a moves forward to æ
Step 4: ɛ becomes e, ɔ becomes o
Step 5: æ moves up to ɛ
Step 6: e moves up to i
A new e was created in Step 4; now that e moves up to i.
Step 7: ɛ moves up to e
The new ɛ created in Step 5 now moves up.
Step 8: əɪ and əʊ drop to aɪ and aʊ



massive sound change of long vowels of English: XIV to XVIII century

Borrowing

• a major source of linguistic change is influx of words from other languages (for need of new terminology, for prestige, or mixed use where overlapping populations): loan words

• phonological and morphological remodeling of loan words

• Identifying loan words and direction of borrowing: phonological patterns of the language; history of phonological changes; morphological complexity of word decreases when borrowed; if borrowing across different language families existence of cognates in other languages reveals direction of borrowing

• loanwords may be "fossils" revealing past linguistic changes in the language of origin

• Example of loan word: *money* borrowed in English from French *monnaie*, Latin *moneta*

Analogical Change

- remodeling of words morphology or semantic on similar but unrelated words
- Example: *sorry* from Old English *sārig* = sore; *sorrow* from Old English *sorh* = grief; unrelated but the modern use of sorry has been modeled on sorrow (semantic)
- Example: *speak/spoke/spoken* remodeled based on verbs like *break/broke/broken* from Old English form *sprec/spræc/gesprecen* German: *sprechen/sprach/gesprochen* (morphology)
- Example: *female* had Middle English form *femelle*, changed by analogy to *male* (phonology)

Other evolutionary mechanisms we know from Historical Linguistics

Semantic shifts: narrowing, metaphor, metonymy/synecdoche, ellipsis/displacement, pejoration, amelioration, euphemism (taboo avoidance), hyperbole, litotes (understatement), semantic shifts by contact

Syntactic changes
- reanalysis: when ambiguity is present in possible analysis of a sentence, shift from one parsing to another
(change of "deep structure")
- extension: widens use of a syntactic construction
(change of "surface structure")
Example: use of *reflexive* in Old Spanish and Modern Spanish
Juanito se vistió
Los vinos que se venden

- syntactic borrowing importing a syntactic construction from another language

Example: the Uto-Aztecal language Pipil imported the *más ... que* Spanish construction (*más linda que tú*) used in Pipil as *mas ... ke* (*mas galá:na ke taha*)

Grammaticalization

Example: *will* in English, original meaning *want* (like German *will*); acquires grammatical use as future auxiliary

Example: *going to* from verb of motion acquired grammatical meaning as future/future intention

Note: there are known phenomena of cyclic grammaticalization

Some references

of general introduction to Linguistics and Historical Linguistics

- John Lyons, *Languages and Linguistics. An introduction*, Cambridge University Press, 1981
- Lyle Campbell, *Historical Linguistics. An Introduction*, MIT Press, 2013.

Comparative Method and Reconstruction of Proto-Languages

- To identify if two languages belong in a (sub)family: search for cognate words
- After identifying a set of cognate words, establish sound correspondence between cognate words
- recently done computationally... but, without accompanying etymological information, it generates false positives

Example: English *much* and Spanish *mucho* may appear to be cognate words but they do not come from a common root
Old English *myċel* = large; Latin *multo* = many

- ... but argued the number of false positives is sufficiently small
- reconstruction of proto-sound from sound correspondences within family and directionality of general sound change rules, plus majority rule (among languages in (sub)family)

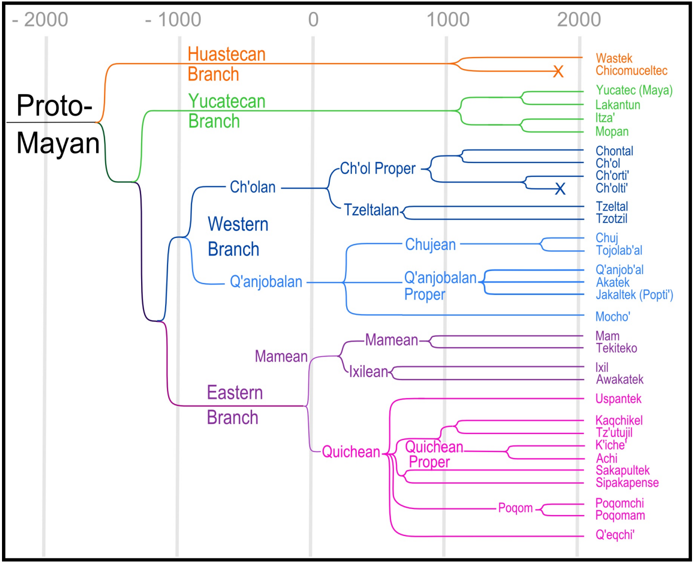
Phylogenetic Linguistics

- Constructing family trees for languages (sometimes possibly graphs with loops)
- Main information about subgrouping: shared innovation

a specific change with respect to other languages in the family that only happens in a certain subset of languages

- Example: among Mayan languages: Huastecan branch characterized by initial *w* becoming voiceless before a vowel and *ts* becoming *t*, *q* becoming *k*, ... Quichean branch by *velar nasal* becoming *velar fricative*, *ć* becoming *č* (prepalatal affricate to palato-alveolar)...

Mayan Language Tree



Computational Methods for Phylogenetic Linguistics

- Peter Foster, Colin Renfrew, *Phylogenetic methods and the prehistory of languages*, McDonald Institute Monographs, 2006
- Several computational methods for constructing phylogenetic trees available from mathematical and computational biology
- Phylogeny Programs
`http://evolution.genetics.washington.edu/phylip/software.html`
- Standardized lexical databases: Swadesh list
(100 words, or 207 words)

- Use Swadesh lists of languages in a given family to look for cognates:

- without additional etymological information (keep false positives)
- with additional etymological information (remove false positives)

- Two further choices about loan words:

- remove loan words
- keep loan words

- Keeping loan words produces graphs that are not trees

- Without loan words it should produce trees, but small loops still appear due to ambiguities (different possible trees matching same data)

... more precisely: coding of lexical data ...

Coding of lexical data

- After compiling lists of cognate words for pairs of languages within a given family
(with/without lexical information and loan words)

- Produce a binary string $S(L_1, L_2) = (s_1, \dots, s_N)$ for each pair of languages L_1, L_2, with entry 0 or 1 at the i-th word of the lexical list of N words if cognates for that meaning exist in the two languages or not (important to pay attention to synonyms)

- lexical Hamming distance between two languages

$$d(L_1, L_2) = \#\{i \in \{1, \dots, N\} \mid s_i = 1\}$$

counts words in the list that do not have cognates in L_1 and L_2

Distance-matrix method of phylogenetic inference

- after producing a measure of "genetic distance"
Hamming metric $d_H(L_a, L_b)$
- hierarchical data clustering: collecting objects in clusters according to their distance
- simplest method of tree construction: neighbor joining

(1) - create a (leaf) vertex for each index a
(ranging over languages in given family)

(2) - given distance matrix $D = (D_{ab})$
distances between each pair $D_{ab} = d_H(L_a, L_b)$
construct a new matrix Q-test

$$Q = (Q_{ab}) \quad \text{with} \quad Q_{ab} = (n-2)D_{ab} - \sum_{k=1}^{n} D_{ak} - \sum_{k=1}^{n} D_{bk}$$

this matrix Q decides first pairs of vertices to join

(3) - identify entries Q_{ab} with lowest values: join each such pair (a, b) of leaf vertices to a newly created vertex v_{ab}

(4) - set distances to new vertex by

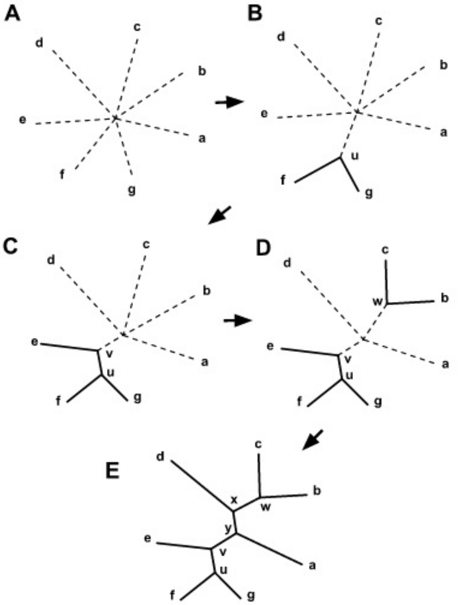
$$d(a, v_{ab}) = \frac{1}{2} D_{ab} + \frac{1}{2(n-2)} \left(\sum_{k=1}^{n} D_{ak} - \sum_{k=1}^{n} D_{bk} \right)$$

$$d(b, v_{ab}) = D_{ab} - d(a, v_{ab})$$

$$d(k, v_{ab}) = \frac{1}{2}(D_{ak} + D_{bk} - D_{ab})$$

(5) - remove a and b and keep v_{ab} and all the remaining vertices and the new distances, compute new Q matrix and repeat until tree is completed

Neighborhood-Joining Method for Phylogenetic Inference



Example of a neighbor-joining lexical linguistic phylogenetic tree

from Delmestri-Cristianini's paper

Variants of the neighbor-joining method

• incorporate better information on the metric on the tree (distance between vertices)

• using a time dependent distance between languages:

$$\dot{D} = -\alpha(1 - D) - \beta D$$

$\alpha =$ effects such as deletion/insertion... increasing difference between words (increasing D) and $\beta =$ effects of analogical change/borrowing decreasing difference (Petroni-Serva paper)

• showed different results on the Austronesian languages with an earlier separation of the Oceanic languages and a two cluster split of Formosan languages

- N. Saitou, M. Nei, *The neighbor-joining method: a new method for reconstructing phylogenetic trees*, Mol Biol Evol. Vol.4 (1987) N. 4, 406-425.
- R. Mihaescu, D. Levy, L. Pachter, *Why neighbor-joining works*, arXiv:cs/0602041v3
- A. Delmestri, N. Cristianini, *Linguistic Phylogenetic Inference by PAM-like Matrices*, Journal of Quantitative Linguistics, Vol.19 (2012) N.2, 95-120.
- F. Petroni, M. Serva, *Language distance and tree reconstruction*, J. Stat. Mech. (2008) P08012

Other methods of Computational Phylogenetics

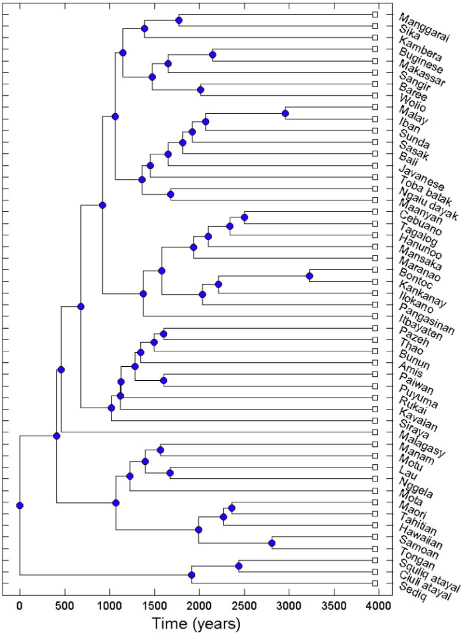
- Neighbor-joining produces unrooted tree
- related method UPGMA (unweighted pair group method with arithmetic mean) gives a rooted tree under equal distance assumption from root to leaves

Use hierarchical clustering by Hamming distance; at each step identify nearest clusters and combine into a higher level cluster: distance between clusters $\mathcal{C}_1$, $\mathcal{C}_2$ is average of distances between objects

$$d(\mathcal{C}_1, \mathcal{C}_2) = \frac{1}{\#\mathcal{C}_1 \cdot \#\mathcal{C}_2} \sum_{x \in \mathcal{C}_1} \sum_{y \in \mathcal{C}_2} d_H(x, y)$$

- drawback: assumes a constant rate of evolution (realistic assumption?)
- R.Sokal, C.Michener, *A statistical method for evaluating systematic relationships*, University of Kansas Science Bulletin 38 (1958) 1409–1438.

UPGMA tree of Austronesian Languages (Petroni-Serva)



Non-uniqueness problem

- often many different trees can match the same data
- Maximum parsimony principle: select the one that requires the minimum number of changes (evolutionary events) to explain the data ...but search is NP-hard
- can increase search efficiency by branch and bound algorithms

organize set of all possible candidate solutions as a rooted tree, with full set at the root and a splitting procedure that separates out subregions of the "solution space" (branches); computing upper and lower bounds for function one wants to minimize over some regions; discard regions where minimum cannot be found (pruning)

Maximum likelihood

• assign probabilities to various possible phylogenetic trees and discard improbable ones

• require evolution at different nodes and along different branches statistically independent

• assign probabilities to particular types of changes (related to maximum parsimony: larger number of changes decreases probability of tree)

• optimization search over all tree topologies: computationally hard

• B.Chor, T.Tuller, *Finding the Maximum Likelihood Tree is Hard*, JACM, 2005.

Bayesian inference

- assume a prior probability distribution for all the possible trees
- this can accommodate models of evolutionary changes as some kind of stochastic process
- Bayesian rule for posterior probability: probability of hypothesis H given observed data D

$$\mathbb{P}(H|D) = \frac{\mathbb{P}(D|H)}{\mathbb{P}(D)} \cdot \mathbb{P}(H)$$

first factor, how well hypothesis H matches data D; second factor, how unlikely hypothesis H in the prior probability

• evaluating posterior probabilities again hard for large set of data: use random sampling method to generate a sample of trees, frequencies distribution of these will approximate posterior probabilities

• typical method used: Markov Chain Monte Carlo (MCMC) approach

• a choice of a set of moves on trees (eg swapping descendant subtrees, cyclically permuting leaves,...) use these moves for a random walk through the space of possible trees

• converge to a stationary distribution which gives maximum posterior probability tree

• drawbacks: dependence on prior probability, and on choice of set of moves

... syntactic instead of lexical phylogenetic trees?

• instead of coding lexical data based on cognate words, use binary variables of syntactic parameters

• Hamming distance between binary string of parameter values

• shown recently that one gets an accurate reconstruction of the phylogenetic tree of Indo-European languages from syntactic parameters only

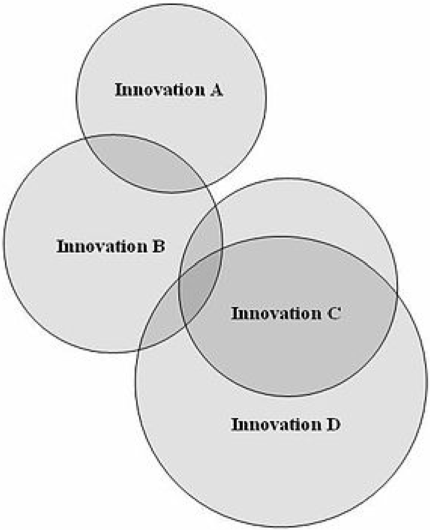
• G. Longobardi, C. Guardiano, G. Silvestri, A. Boattini, A. Ceolin, *Towards a syntactic phylogeny of modern Indo-European languages*, Journal of Historical Linguistics 3 (2013) N.1, 122–152.
• G. Longobardi, C. Guardiano, *Evidence for syntax as a signal of historical relatedness*, Lingua 119 (2009) 1679–1706.

• also recently results obtained using phonetic phylogenetic trees

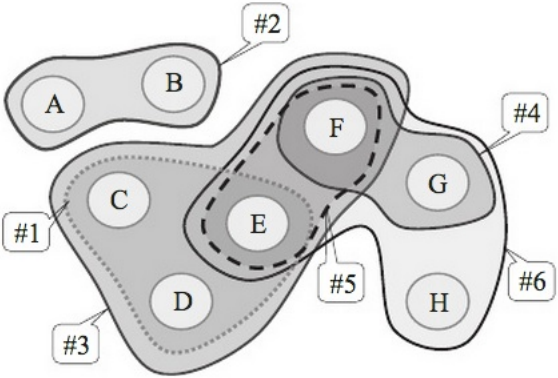
Wave Theory of Languages an alternative to Phylogeny

- Phylogenetic trees rely on the assumptions that languages are discrete entities (nodes in a tree)
- Evidence to the contrary: dialects
- Languages are a dialect continuum
- in this continuum medium innovations and changes spread like waves in a pond moving outward in time from where they originate
- Model developed originally by German linguist Johannes Schmidt (end of XIX century) in opposition to the Neogrammarian school, more recently considered a good model for Oceanic languages
- W.Labov, *Transmission and Diffusion*, Language, Vol.83 (2007) N.2, 344–387.
- J.Lynch, M.Ross; T.Crowley, *The Oceanic Languages*, Curzon, 2002.

Wave Model of Languages

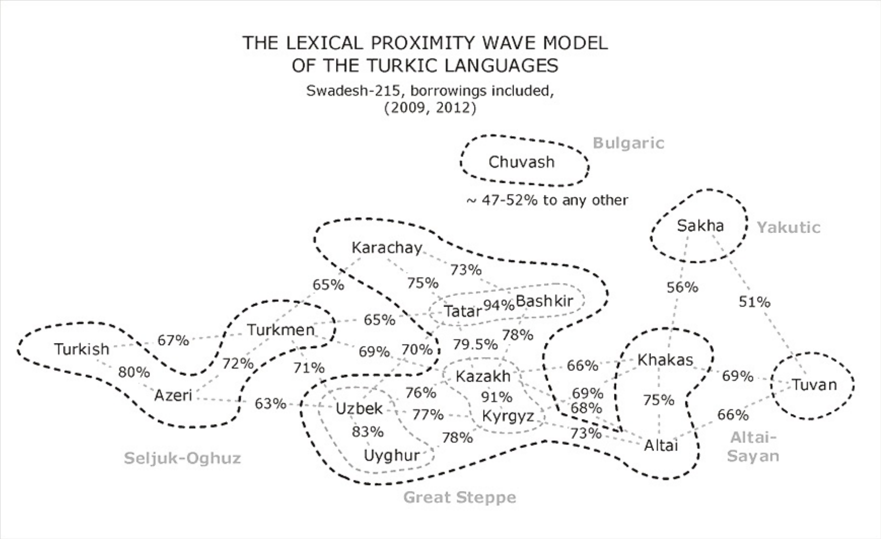


Intersecting Wave Isoglosses



- A. François, *Trees, Waves and Linkages. Models of language diversification*, in "The Routledge Handbook of Historical Linguistics"

Wave Model: Turkic Languages



From Pāṇini to de Saussure and Chomsky
Historical Origins of Modern Linguistics

- by mid XIX century European linguists proficient with Sanskrit
- Franz Bopp studied Sanskrit/Greek comparative linguistics; first European who seriously studied Pāṇini's text
- end of XIX century, early XX century: Ferdinand de Saussure, professor of Sanskrit, devised modern structural linguistics influenced by his reading of Pāṇini
- early XX century: Leonard Bloomfield, who started the American school of Structuralism, studied Sanskrit with Jacob Wackernagel in Göttingen, and refers to Pāṇini as a major influence
- Pāṇini's work also influenced logician Emil Post and his theory of *canonical systems* (formal languages with string rewrite rules)

Pāṇini and the Aṣṭādhyāyī

- lexical lists (Dhatupāṭha, Gaṇapāṭha)
- algorithms to be applied to inputs from lexical lists to form
well formed words (morphology)
well posed grammatical sentences (syntax)
syntax is less developed than morphology and phonetics
- introduced notions of phoneme, morpheme, root and word forms
- distinguishes between syntax, morphology, and lexicology
- text organized in 3,959 *sūtrāṇi* (rules) across 8 chapters
- 3 associated texts: Śivasūtrāṇi (a list of all Sanskrit phonemes with suitable notation); Dhatupāṭha (a lexical list of Saskrit verbal roots, organized in ten classes); Gaṇapāṭha (a lexical list of nominal stems)

Śivasūtrāṇi शिवसूत्राणि 14 lists of phonemes

IAST	Devanāgarī
1. a i u ṇ	१. अ इ उ ण्।
2. Ṛḷ k	२. ऋ ऌ क्।
3. e o ṅ	३. ए ओ ङ्।
4. ai au c	४. ऐ औ च्।
5. ha ya va ra ṭ	५. ह य व र ट्।
6. la ṇ	६. ल ण्।
7. ña ma ṅa ṇa na m	७. ञ म ङ ण न म्।
8. jha bha ñ	८. झ भ ञ्।
9. gha ḍha dha ṣ	९. घ ढ ध ष्।
10. ja ba ga ḍa da ś	१०. ज ब ग ड द श्।
11. kha pha cha ṭha tha ca ṭa ta v	११. ख फ छ ठ थ च ट त व्।
12. ka pa y	१२. क प य्।
13. śa ṣa sa r	१३. श ष स र्।
14. ha l	१४. ह ल्।

called Śivasūtrāṇi because of a poetic image describing the list of phonemes of the Sanskrit language as resulting from the drum beats of Shiva's Cosmic Dance

नृत्तावसाने नटराजराजो ननाद ढक्कां नवपञ्चवारम्।
उद्धर्तुकामः सनकादिसिद्धादिनेतद्विमर्शे शिवसूत्रजालम्॥

Phonemes and Phonology in Pāṇini

• phoneme arranged similarly to modern classification by manner of articulation

• each of the 14 groups of phonemes ends with a dummy letter (symbol) *anubandha*

• the *anubandha* distinguishes: vowels, sibilant, nasals, palatals, ...

• phonological rules are then formulated using the *anubandha* for an arbitrary element of the corresponding group of phonemes

Example: the rule *y v r l* replace *i u ṛ ḷ* before a vowel stated as *iKo yaṆ aCi*; iK={ *i, u, ṛ, ḷ* }, iKo=genitive; yaṆ={*y*, *v*, *r*, *l*} semivowels; aC=vowels, aCi=locative

• suprasegmental structures and connections to modern Feature Geometry (Kornai)

Mahābhāṣya of Patañjali

- later commentary on Pāṇini's Aṣṭādhyāyī with further elaboration on Sanskrit grammar (2nd century BCE)
- considers further level of structure: semantics

Bhartṛhari and the theory of Sphoṭa स्फोट

- later development (5th century CE); term *sphoṭa* already used in Patañjali (and perhaps in Pāṇini) for a notion analogous to *phoneme*
- finer notion of *sphoṭa* in Bhartṛhari: *varṇa-sphoṭa* (phoneme) *pada-sphoṭaa* (lexeme, morpheme); *vakya-sphoṭa* (unit of structure at sentence level: syntactic)
- sign, signifier (*vācaka*), and signified (*vācya*)

Ferdinand de Saussure and Structural Linguistics

- emphasis on *synchronic* instead of *diachronic* view of language
- *sign* as foundation: *signified* (semantic level) and *signifier* (mean of expressing it)
- paradigmatic relations between sets of units (grouped by common properties)
- syntagmatic relations: rules for chaining units selected by paradigmatic rules into larger structures

- Bloomfield's American Structuralism: less emphasis on semantics, more on mechanics of phonology, morphology

From de Saussure to Chomsky's generative grammar

- criticism of structuralist approach: maybe OK for phonology and morphology, inadequate for syntax
- Chomsky claims first "generative grammar" was Pāṇini's Aṣṭādhyāyī
- General idea: a set of rules produced in an algorithmic way that predict grammaticality of sentences (including morphological level)
- in second half of XX century, structural linguistics superseded by generative grammar

What's so special about Sanskrit?

- morphologically and syntactically richest Indo-European language
- large body of literature spanning millennia of language evolution
- organized scientifically: work of Pāṇini and the ancient linguists
- considered very suitable for Computational Linguistics (look at the series of volumes on *Sanskrit Computational Linguistics* in the Lecture Notes in Artificial Intelligence series of Springer)
- contact with Sanskrit had a massive impact on European culture

The Sanskrit language, whatever its antiquity, is of a wonderful structure; more perfect than the Greek, more copious than the Latin, and more exquisitely refined than either, yet bearing to both of them a stronger affinity both in the roots of verbs and in the forms of grammar, than could possibly have been produced by accident; so strong indeed that no philologer could examine them all three, without believing them to have sprung from a common source...

Sir William Jones, 1786

References:

1. Otto Böhtlingk, *Panini's Grammatik*, 1887
2. András Kornai, *The generative power of feature geometry*, Annals of Mathematics and Artificial Intelligence, 8 (1993) N.1-2, 37–46
3. Ferdinand de Sassure, *Course in General Linguistics* (1916), Open Court, 1983
4. Noam Chomsky, *Aspects of the theory of syntax*, MIT Press, 1965