

---

# Cognate Identification and Phylogenetic Inference: Search for a Better Past

---

Abhaya Agarwal & Jason Adams

Language and Statistics II

Fall 2007

---

# Historical Linguistics

- Study of language change
    - Identify genetic relationships between languages
    - Identify corresponding language features like recurrent sound correspondences
  - Cognate identification finds words which correspond to the same ancestor word
  - Phylogenetic Inference uses this information to identify genetic relationships
-

---

# So how do languages change?

- Two primary modes
  - Phonological shifts
    - Simplification of consonant clusters is a universal property in languages
    - Regular in nature, same sound in same context will always change
  - Semantic Shift
    - Words change their meaning over time and fall out of use
-

---

# Cognate Identification

- Cognates

- Strict

- Words in a language pair that have descended from the same word in a parent language
    - Does not include loan words (borrowings)

- Orthographic

- Words with similar spellings/phonetic transcriptions
-

---

# Two Approaches to Cognate Identification

- Orthographic methods
    - Cognates are determined by string and pattern matching
  - Phonetic methods
    - Cognates are determined by
      - Matching phonetic patterns
      - Finding sound correspondences
  - Distinguishing factor: phonological information
-

# Orthographic Measures

- Simple string matching
  - Simard et al. (1993) compared first four characters to measure *cognateness*
- Longest Common Subsequence Ratio

$$LCSR(x, y) = \frac{\|LCS(x, y)\|}{\max(\|x\|, \|y\|)}$$

- Dice's coefficient

$$Dice(x, y) = \frac{2f_{x,y}}{f_x + f_y}$$

# Similarity Measures

- Tiedemann (1999)
  - Three weighted string similarity measures
    1. Single corresponding vowels and consonants
    2. Sequences of vowels and consonants
    3. Non-matching parts of strings that are otherwise the same
      - Used to find systematic differences
      - Example: c:ska *asymmetric* : *asymmetriska*

---

# Orthographic Approaches

- Advantages

- Methods may be applied to standard text
  - Does not require phonetic transcription
- Most algorithms are relatively simple
  - LCSR
  - Dice's coefficient



# Orthographic Approaches

- Disadvantages

- False positives

- *faux amis* – homograms (or homophones if phonetic) but different meanings or origins

- English: *billion* German: *Billion* (= trillion)

- Different Orthographies

- e.g. Cyrillic versus Latin alphabets

---

# Phonetic Approaches

- Uses phonetic transcriptions of words
    - Usually manual transcriptions
    - Can be automatic when available
-

---

# Phonetic Similarity

- Phonetic similarity measures
  - Attempt to quantify closeness of sounds
  - Not straightforward
    - Edit distance misses important information
      - /d/ would be the same distance from /a/ as from /t/

---

# Inexact String Matching

- Covington (1996)
    - Inexact string matching of phonetic transcriptions
    - Operations: skip and match
      - Cost assigned to each
    - Minimizing cost maximizes phonetic similarity score of characters in the word-pair
-

---

# Phonological Features

- Kondrak (2000)
    - ALINE system
      - Phonetic similarity determined by multi-valued phonological features
      - *ad hoc* salience weights applied to features
    - Not empirical: uses insights from linguistics
-

# Phonological Features

Feature name	Phonological term	Numerical value	Manner			
Place	[bilabial]	1.0		[stop]	1.0	
	[labiodental]	0.95		[affricate]	0.9	
	[dental]	0.9		[fricative]	0.8	
	[alveolar]	0.85		[approximant]	0.6	
	[retroflex]	0.8		[high vowel]	0.4	
	[palato-alveolar]	0.75		[mid vowel]	0.2	
	[palatal]	0.7		[low vowel]	0.0	
	[velar]	0.6		High	[high]	1.0
	[uvular]	0.5			[mid]	0.5
	[pharyngeal]	0.3			[low]	0.0
	[glottal]	0.1		Back	[front]	1.0
					[central]	0.5
				[back]	0.0	

Example phonological features used by Kondrak (2000).

---

# Adding Semantic Knowledge

- Kondrak (2001)
    - Uses ALINE system to generate phonetic similarity
    - Adds measure of semantic similarity from WordNet
    - Outperformed orthographic measures on precision
-

---

# Learning Approaches

- Learning approaches are popular in the recent literature
    - Dynamic Bayesian Networks
      - Filali and Bilmes (2005)
    - Pair HMMs
      - MacKay and Kondrak (2005)
    - Support Vector Machines
      - Mulloni (2007)
    - Semi-supervised learning of partial cognates
      - Frunza and Inkpen (2006)
-

---

# Dynamic Bayesian Networks

- Filali and Bilmes (2005)
  - Type of Graphical Model
    - HMMs being the simplest form
    - Means of representing factored probability distribution as a graph
-

---

# Dynamic Bayesian Networks

- Builds a stochastic model of edit distance
    - Each node has a probability function
      - maps parent values to the possible values the node can take
  - Empirical – model learned from phonetically transcribed data
  - Produces a model of phonetic similarity
-

---

# Method Comparison

- Kondrak and Sherif (2006)
  - Evaluation of many methods for computing phonetic similarity applied to cognate identification
    - Dynamic Bayesian Networks
    - Pair HMMs
    - ALINE
    - Levenshtein distance with learned weights
    - Edit distance with uniform costs
    - others
-

# Method Comparison

Languages		Proportion of cognates	Method						
			EDIT	MIEL	ALINE	R&Y	LLW	PHMM	DBN
English	German	0.590	0.906	0.909	0.912	0.894	0.918	0.930	0.927
French	Latin	0.560	0.828	0.819	0.862	0.889	0.922	0.934	0.923
English	Latin	0.290	0.619	0.664	0.732	0.728	0.725	0.803	0.822
German	Latin	0.290	0.558	0.623	0.705	0.642	0.645	0.730	0.772
English	French	0.275	0.624	0.623	0.623	0.684	0.720	0.812	0.802
French	German	0.245	0.501	0.510	0.534	0.475	0.569	0.734	0.645
Albanian	Latin	0.195	0.597	0.617	0.630	0.568	0.602	0.680	0.676
Albanian	French	0.165	0.643	0.575	0.610	0.446	0.545	0.653	0.658
Albanian	German	0.125	0.298	0.340	0.369	0.376	0.345	0.379	0.420
Albanian	English	0.100	0.184	0.287	0.302	0.312	0.378	0.382	0.446
AVERAGE		0.2835	<b>0.576</b>	<b>0.597</b>	<b>0.628</b>	<b>0.601</b>	<b>0.637</b>	<b>0.704</b>	<b>0.709</b>

Precision averaged across 11 test sets using Comparative Indo-European Data Corpus. (Kondrak and Sherif, 2006)

---

# Phonetic Approaches

- Advantages

- Linguistic knowledge taken into account
- Phonetic information closer to “ground truth” of spoken language

- Disadvantages

- Phonetic transcriptions required
    - Manual → time-consuming
    - Automatic → noisy
  - Algorithms tend to be significantly more complicated
-

---

# Cognate Identification: Conclusions

- Phonetic currently outperforming orthographic
  - Learning currently outperforming rule-based
  - Other information sources (e.g. semantic) not yet being exploited in learning approaches
-

---

# Cognate Identification: Future Work

- Better approaches for learning from practical orthographies



---

# Reconstructing the evolutionary History

- Once the cognates are identified and regular sound changes discovered,

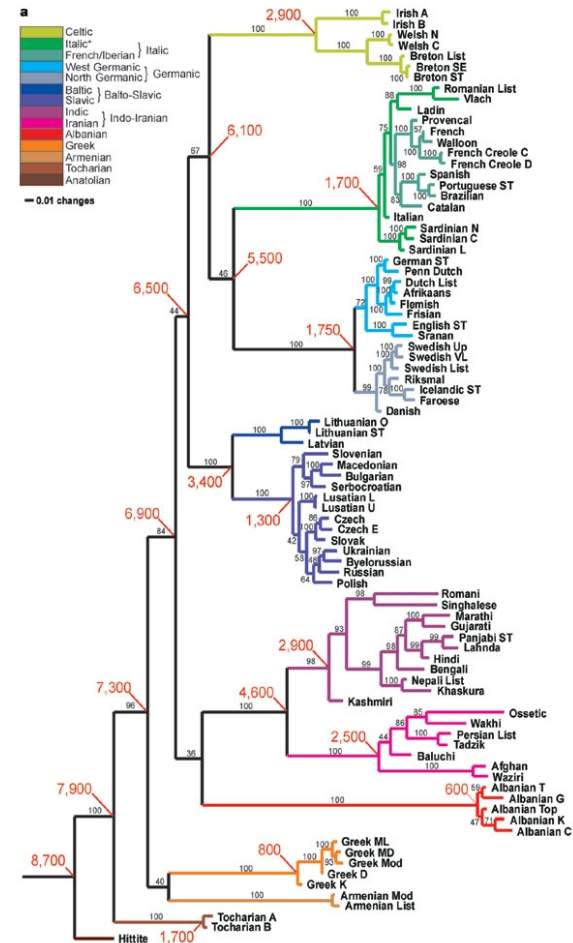
**it's time for  
building the tree !**

---



# Phylogenetic Trees

- How many rivers do we have in Pittsburgh?
  - 2 or 3 ?
- Assumption in the tree: Languages develop independently after diverging



---

# Character(istics) of Languages

- These are features describing the languages
  - A character assigns a whole number to every language depending on some characteristics of the language like lexical or phonological
  - A language can be represented as a vector of character states
-

# Lexical Characters

languages	English	German	French	Italian	Spanish	Russian
words	hand	Hand	main	mano	mano	ruk'a
states	1	1	2	2	2	3

- Languages that have cognate words for the same concept are assigned the same state
- If there would be no semantic shift, all the languages would have same state.

# Phonological Characters

Sanskrit	Greek	Latin	English	Gothic
pita	pater	pater	father	fadar
padam	poda	pedem	foot	fotu
0	0	0	1	1

- Encodes if a particular sound change happened in a language or not
- Here the change of initial p to f seems to have happened in English and Gothic only

---

# Methods for Phylogenetic Reconstruction

- Given a language set, how do we reconstruct the tree?
  - Distance Based methods
    - Measure some sort of distance between languages and cluster based on that
  - Character based method
    - Take into account the individual state changes
-

---

# Distance Based Methods

- UPGMA

- Cluster languages based on the percentage of shared cognates

- Neighbour-Joining\*

- Start with star like topology and greedily merge the two nodes that reduce the tree length most at each step

---

\* First introduced by Saitou and Nei (1987)

---

# Character Based Methods

- Work better than distance based methods in general.
  - Maximum Parsimony
    - Principle of Minimum Evolution
    - Find the tree on which minimum number of state changes take place
  - Doesn't have an explicit model of character evolution
-

---

# Models of Evolution

- Describe how characters evolve along the branches of trees
  - Ideally every character can evolve at a different rate along every branch.
  - We can make additional assumptions like *a character evolves at the same rate across all the branches*
-

---

# Maximum Likelihood\*

- Parametric method
- Search for the tree typology that maximizes the likelihood of observed character states
- It is not widely used because it is slow
- With language data, Bayesian techniques have been used

---

\* See J. Felsenstein (1981)

---

# Homoplasy and Borrowing

- The independence assumption made on the tree can be broken by language contact
  - Also Parallel evolution and back mutation are two confounding effects that need to be modelled
  - Compatible character
    - A character that evolves on the tree without homoplasy
-

---

# Accounting for Homoplasy

- Maximum Compatibility (Ringe et al. 2002)
    - Some researchers have argued that it is easy to identify and eliminate homoplastic characters for language data
    - Try to find the tree that has maximum characters compatible on it
  - We can explicitly account for homoplasy and borrowing in our models
    - Trees -> networks (borrowing)
    - Models of evolution allow homoplasy
-

# Comparison of Methods

- Studies have been done comparing the methods on
  - On real data (Indo-European languages Nakhleh et al. 2005)
  - On synthetic data (Barbancon et al. 2007)
- In general, distance based methods perform worse than character based methods.
- There is no clear winner among character based methods for language data.

---

# Directions being explored

- Dating of internal nodes (Atkinson et al. 2005)
  - Reconstructing trees for languages that diverged long ago (which is around 10k years) (Dunn et al. 2005, Ryder 2006)
  - Try to take the linguist out of the loop (Bouchard et al. 2007)
-

---

# Future Work

- More realistic and informed models of evolution
  - Working with raw language data is desirable
-

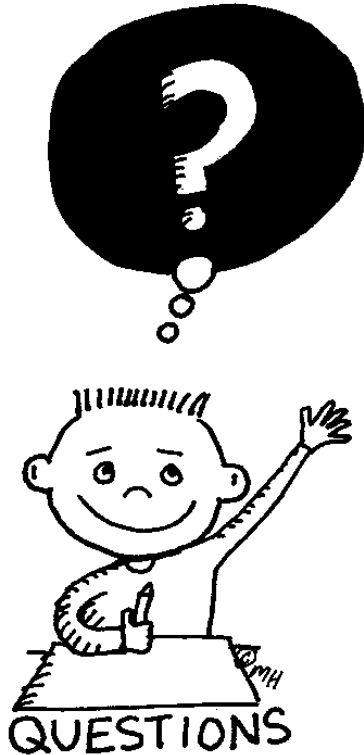
---

# Conclusion

- It is an area that has recently opened up to computational techniques
  - Field is still in a exploratory mode, there are no proven state of the art methods yet
-

---

# Questions?



Email us your comments:  
[jmadams@cs.cmu.edu](mailto:jmadams@cs.cmu.edu)  
[abhayaa@cs.cmu.edu](mailto:abhayaa@cs.cmu.edu)

---

# References

- M.A. Covington. An algorithm to align words for historical comparison. *Computational Linguistics*, 22(4):481-496, 1996.
- K. Filali and J. Bilmes. A Dynamic Bayesian Framework to Model Context and Memory in Edit Distance Learning: An Application to Pronunciation Classification. *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 338-345, 2005.
- O. Frunza and D. Inkpen. Semi-supervised learning of partial cognates using bilingual bootstrapping. *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the ACL*, pages 441-448, 2006.
- G. Kondrak. A new algorithm for the alignment of phonetic sequences. *Proceedings of the First Meeting of the North American Chapter of the Association for Computational Linguistics*, pages 288-295, 2000.

---

# References

- G. Kondrak. Identifying cognates by phonetic and semantic similarity. *North American Chapter Of The Association For Computational Linguistics*, pages 1-8, 2001.
  - Grzegorz Kondrak. Determining recurrent sound correspondences by inducing translation models. In *Proceedings of the 19th international conference on Computational linguistics*, pages 17, Morristown, NJ, USA, 2002.
  - G. Kondrak and T. Sherif. Evaluation of Several Phonetic Similarity Algorithms on the Task of Cognate Identification. *Proceedings of the Workshop on Linguistic Distances*, pages 43-50, 2006.
  - W. Mackay and G. Kondrak. Computing Word Similarity and Identifying Cognates with Pair Hidden Markov Models. *Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL-2005)*, pages 40-47, 2005.
-

# References

- I.D. Melamed. Bitext maps and alignment via pattern recognition. *Computational Linguistics*, 25(1):107-130, 1999.
- A. Mulloni. Automatic Prediction of Cognate Orthography Using Support Vector Machines. *Proceedings of the ACL 2007 Student Research Workshop*, pages 25-30, 2007.
- D. Ringe, Tandy Warnow, and A. Taylor. Indo-european and computational cladistics. *Transactions of the Philological Society*, 100(1):59-129, 2002.
- N. Saitou and M. Nei. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol*, 4:406-425, July 1987.
- M. Simard, G.F. Foster, P. Isabelle. Using cognates to align sentences in bilingual corpora. *Proceedings of the 1993 Conference of the Centre for Advanced Studies on Collaborative Research: Distributed Computing – Volume 2*, pages 1071-1082, 1993.
- J. Tiedemann. Automatic construction of weighted string similarity measures. *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, 1999.

---

# Dating the nodes

- Glottochronology
    - Assume lexical clock
    - A language loses ~20% of its vocabulary in 1000 years
  - BI
    - Some model of evolutions allow for the estimation of dates on internal nodes
    - Need to calibrate nodes by taking clue from historical evidence
  - Argument against dating
-

---

# Looking deeper into the past

- With lexical and phonological characters, we can detect relationships to a time depth of ~10k years
  - Grammatical characters
    - Have been shown to contain phylogenetic signal even when all lexical clues are lost
  - Rates of lexical evolution correlate negatively with frequency of use
    - This suggests that there are words whose half life is much larger than 10k years
-

---

# Can we avoid the linguist?

- No since the aim is to provide them tools to analyse language history
  - But we can try to lower the entry barrier if we can automate the comparative method
  - Putting together cognate identification and phylogenetic inference, we can make the whole thing faster and allow the linguist to focus on finer details
-