Inferring phylogenetic graphs of Natural Languages using Minimum Message Length

Jane N. Ooi and David L. Dowe,
Monash University, Australia,
www.csse.monash.edu.au/~dld

Table of Contents

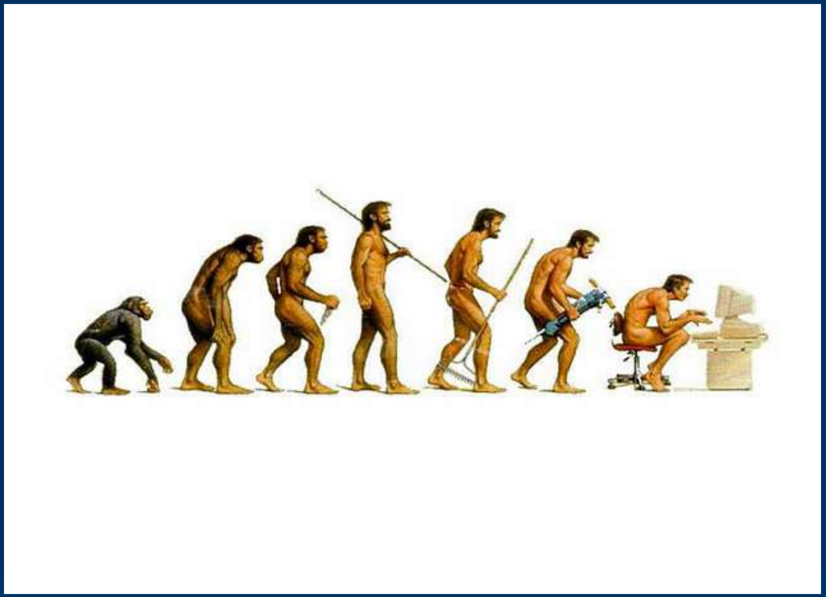
- Motivation and Background
- What is a phylogenetic model?
- Phylogenetic Trees and Graphs
- Types of evolution of languages
- Minimum Message Length (MML)
  - Multistate distribution – modelling of mutations
- Results/Discussion
- Conclusion and future work

Motivation

- To study how languages have evolved (Phylogeny of languages).
  - e.g. Artificial languages,
  - European languages.
- To refine natural language compression method.

Evolution of languages

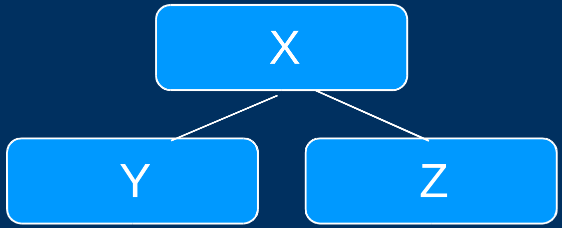
- What is phylogeny?
  - Phylogeny means Evolution



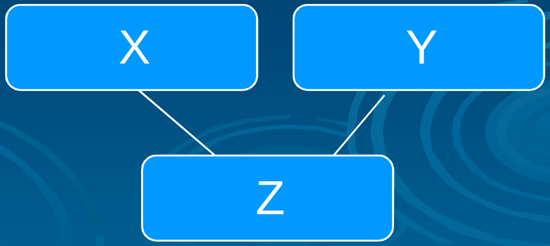
- What is a phylogenetic model?
  - A **phylogenetic tree/graph** is a tree/graph showing the evolutionary interrelationships among various species or other entities that are believed to have a common ancestor.

Difference between a phylogenetic tree and a phylogenetic graph

- Phylogenetic trees
  - Each child node has exactly one parent node.



- Phylogenetic graphs (new concept)
  - Each child node can descend from one or more parent node(s).



Evolution of languages

- 3 types of evolution
  - Evolution of phonology/pronunciation

Words	US	UK
schedule	skedule	shedule
leisure	leezhure	lezhure

  - Evolution of written script/spelling

English	Malay
Mobile	Mobil
Television	Televisyen

  - Evolution of grammatical structures

Minimum Message Length (MML)

- What is MML?
  - A measure of goodness of classification based on information theory (Wallace and Boulton, 1968; Wallace and Dowe, 1999a; Wallace, 2005).
- Data can be described using “models”
- MML methods favour the “best” description of data where
  - “best” = shortest overall two-part message length
- Two part message
  - Msglength = Msglength(model) + msglength(data|model)

Minimum Message Length (MML)

- Degree of similarity between languages can be measured by compressing them in terms of one another.
- Example :
  - Language A Language B
    - 3 possibilities –
      - Unrelated – shortest message length when compressed separately.
      - A descended from B – shortest message length when B compressed and then A compressed in terms of B.
      - B descended from A – shortest message length when A compressed and then B compressed in terms of A.

Minimum Message Length (MML)

The best phylogenetic model is the tree/graph that achieves the shortest overall two-part message length.

Modelling mutation between words

- Root language
  - Equal frequencies for all characters.
    - Log(size of alphabet) * no. of chars.
  - Some characters occur more frequently than others.
    - e.g.: English “x” compared with “a”.
    - Multi-state (multinomial) distribution of characters.

Modelling mutation between words

- Child languages
  - Muti-state distribution
    - 4 states.
      - Insert
      - Delete
      - Copy
      - Change
  - Use string alignment techniques to find the best alignment between words.
  - Dynamic Programming Algorithm to find alignment between strings.
  - MML favors the alignment between words that produces the shortest overall message length.

Example:

r e c o m m a n d e r

| | | | | | | | | | |

r e c o m m e n d - -

Work to date

- Preliminary model
  - Only copy and change mutations
  - Words of the same length
  - artificial and some European languages.
- Expanded model
  - Copy, change, insert and delete mutations
  - Words of different length
  - artificial and some European languages.

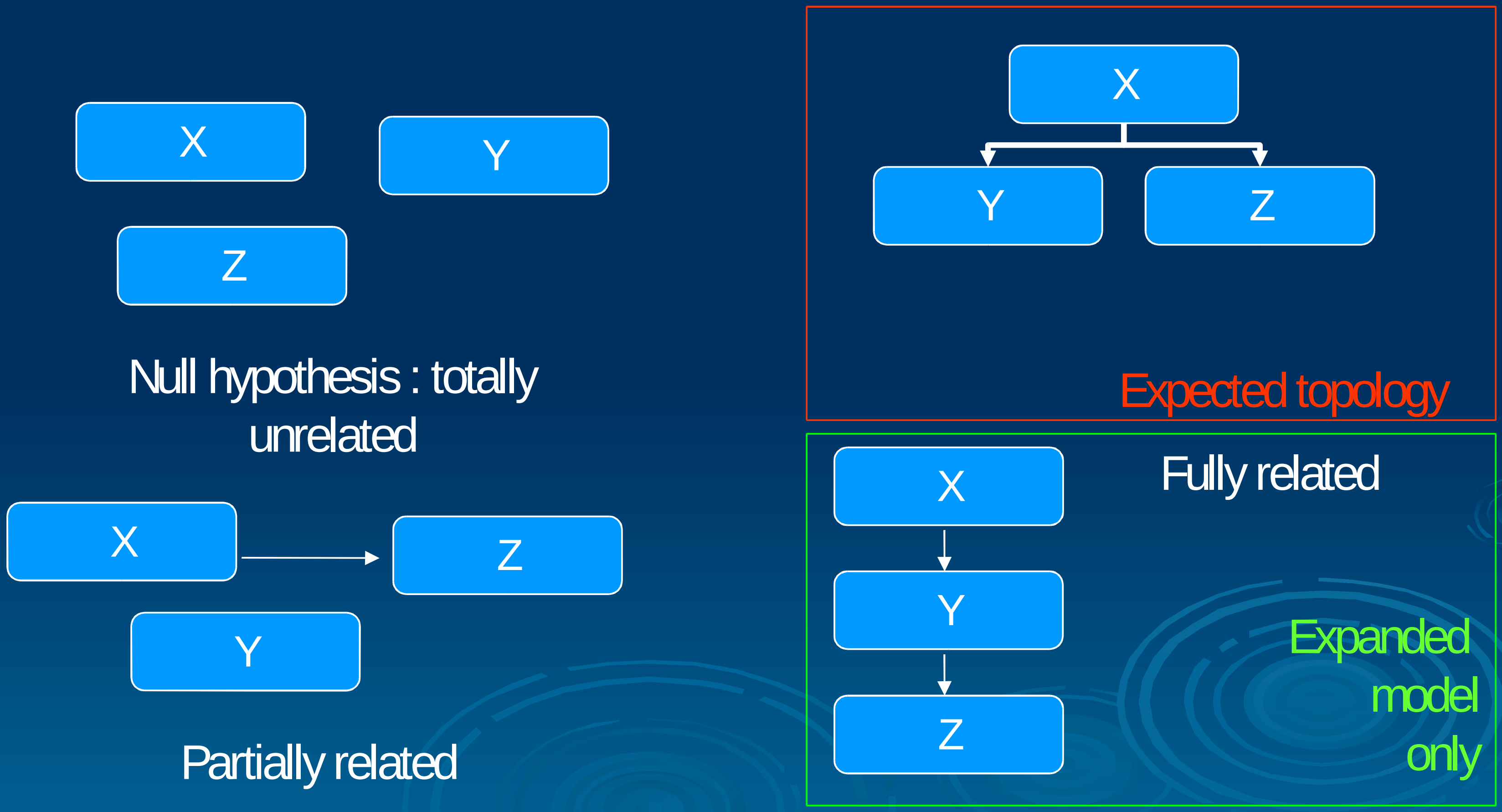
Results – Preliminary model

- Artificial languages
- A – random
- B – 5% mutation from A
- C – 5% mutation from B
- Full stop "." marks the end of string.

	A	B	C
1	asdfge.	assfge.	assfge.
2	zlsdrya.	zlcdrya.	zlchrya.
3	wet.	wet.	wbt.
4	vsert.	vsegt.	vsagt.
…	…....	…....	…....
50	…....	…....	…....

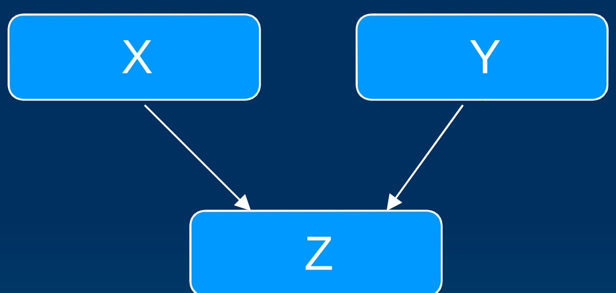
Results – Preliminary model

- Possible tree topologies for 3 languages :

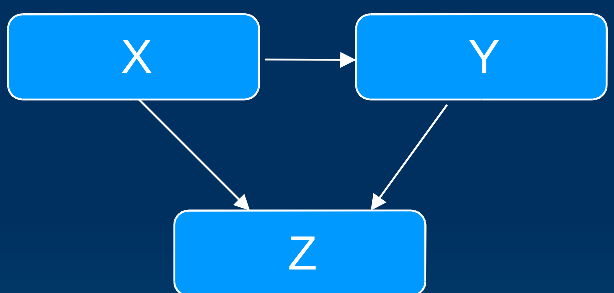


Results – Preliminary model

- Possible graph topologies for 3 languages:



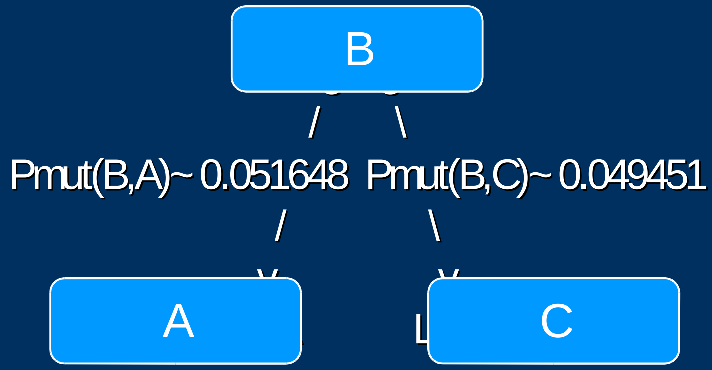
Non-related parents



Related parents

Results – Preliminary model

- Results :
  - Best tree =



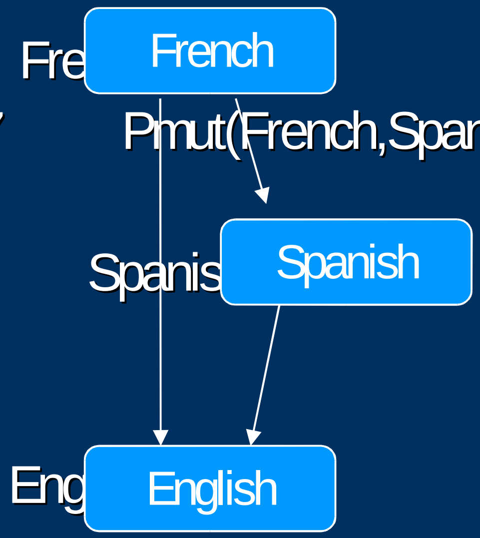
  - Overall Message Length = 2933.26 bits
    - Cost of topology = log(5)
    - Cost of fixing root language (B) = log(3)
    - Cost of root language = 2158.7186 bits
    - Branch 1
      - Cost of child language (Lang. A) binomial distribution = 392.069784 bits
    - Branch 2
      - Cost of child language (Lang. C) binomial distribution = 378.562159 bits

Results – Preliminary model

- European Languages (with accents removed)
  - French
  - English
  - Spanish

	English	French	Spanish
1	baby.	bebe.	nene.
2	beach.	plage.	playa.
3	biscuits.	biscuits.	bizcocho.
4	cream.	creme.	crema.
…	…..	…..	…..
30	…..	…..	…..

Results – Preliminary model



P(from French)~ 0.834297

P(from Spanish not French) ~ 0.090559

P(from neither)~ 0.075145

Pmut(French,Spanish) ~ 0.245174

Cost of "parent" language (French) =1226.76 bits
Cost of language (Spanish) binomial distribution = 734.59 bits
Cost of child language (English) trinomial distribution = 537.70 bits
Total tree cost = log(5) + log(3) + log(2) + 1226.76 + 734.59 + 537.70
= 2503.95 bits

Results – Expanded model

- 16 sets of 4 languages
- Different length vocabularies
  - A – randomly generated
  - B – mutated from A
  - C – mutated from A
  - D – mutated from B
- Mutation probabilities
  - Copy – 0.65
  - Change – 0.20
  - Insert – 0.05
  - Delete – 0.10

Results – Expanded model

	Language A	Language B	Language C	Language D
1	awjmv.	afjmv.	wqmv.	afjnv.
2	bauke.	baxke.	auke.	bave.
3	doinet.	domnit	deoinet.	domnit.
4	eni.	eol.	enc.	eol.
5	foijgnw.	fiogw.	foijnw.	fidgw.
…...	……....	……....	……....	……....
…...	……....	……....	……....	……....
50	……....	……....	……....	……....

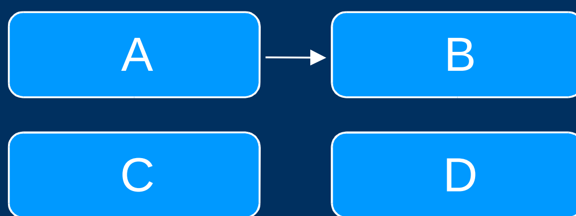
Examples of a set of 4 vocabularies used

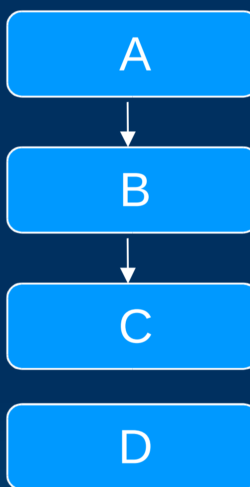
Results – Expanded model

- Possible tree structures for 4 languages:

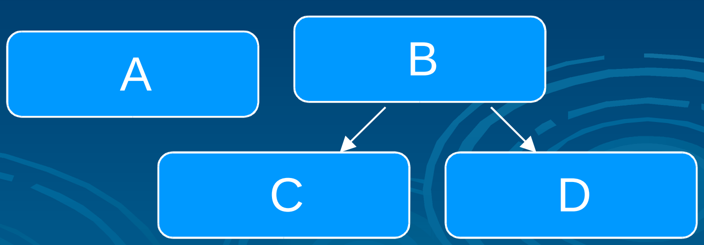
Null hypothesis :

totally unrelated

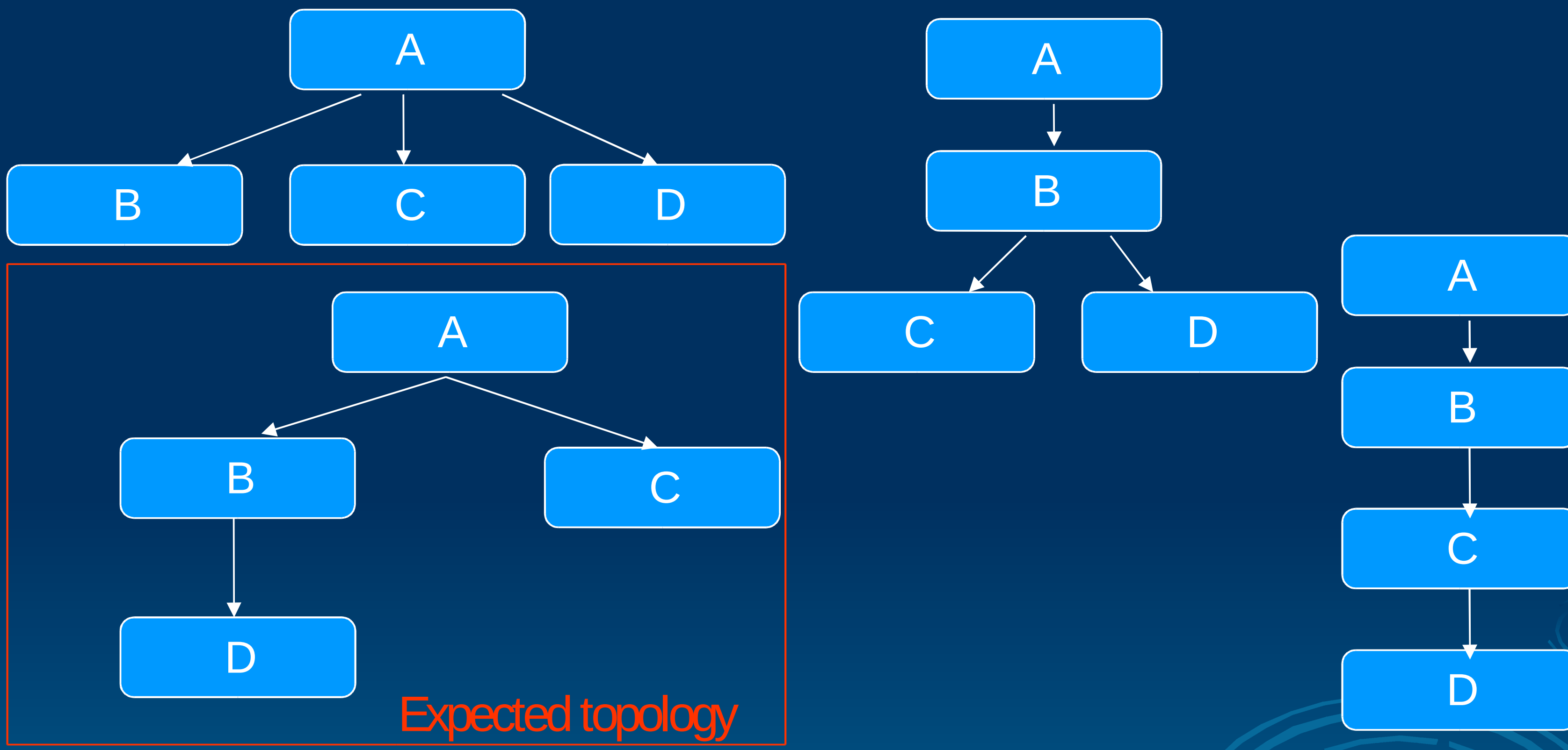




Partially related



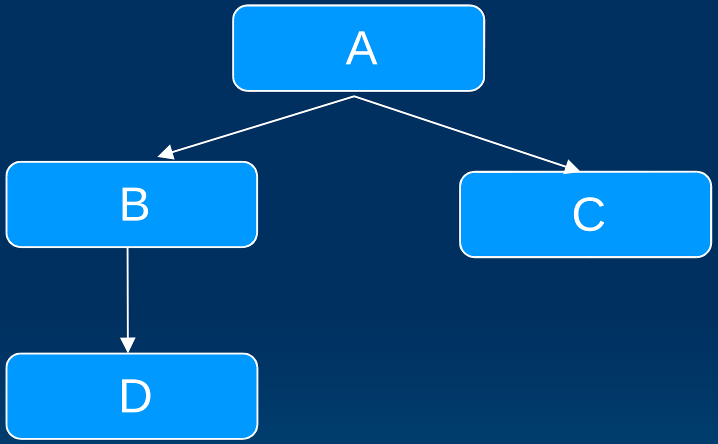
Results – Expanded model



Fully related

Results – Expanded model

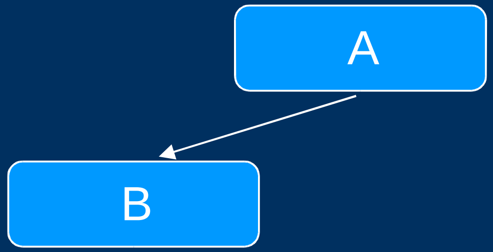
- Correct tree structure 100% of the time.
- Sample of inferred tree and cost :



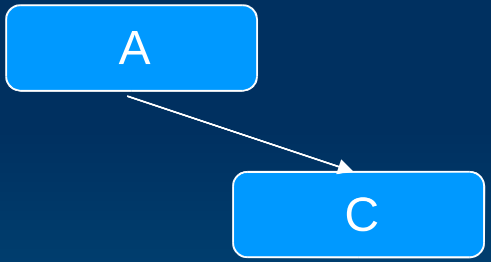
- Language A : size = 383 chars, cost = 1821.121913 bits

Results – Expanded model

- Pr(Delete) = 0.076250
- Pr(Insert) = 0.038750
- Pr(Mismatch) = 0.186250
- Pr(Match) = 0.698750
- 4 state Multinomial cost = 930.108894 bits



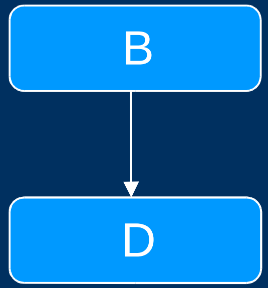
- Pr(Delete) = 0.071250
- Pr(Insert) = 0.038750
- Pr(Mismatch) = 0.183750
- Pr(Match) = 0.706250
- 4 state Multinomial cost = 916.979371 bits



- *Note that all multinomial cost includes and extra cost of log(26) to state the new character for mismatch and insert *

Results – Expanded model

- Pr(Delete) = 0.066580
- Pr(Insert) = 0.035248
- Pr(Mismatch) = 0.189295
- Pr(Match) = 0.708877
- 4 state Multinomial cost = 873.869382 bits



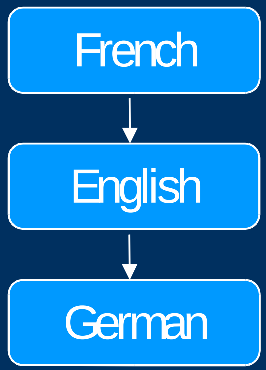
- Cost of fixing topology = log(7) = 2.81 bits
- Total tree cost = 930.11 + 916.98 + 873.87 +
  1821.11 + log(7) + log(4) + log(3) + log(2)
  = 4549.46 bits

Results – Expanded model

- European Languages
  - French
  - English
  - German

	English	French	German
1	even.	meme.	sogar.
2	eyes.	oeil.	auge.
3	false.	faux.	falsch.
4	fear.	peur.	angst.
…	…...	…...	…...
601	…...	…...	…...

Results – Expanded model



Total cost of this tree = 56807.155 bits

Cost of fixing topology = log(4) = 2 bits

Cost of fixing root language (French) = log(3) = 1.585 bits

- Cost of French = no. of chars * log(27) = 21054.64 bits

Results – Expanded model

- Cost of fixing parent/child language (English) = log(2) = 1 bit
- Cost of multistate distribution (French -> English) = 15567.98 bits
- MML inferred probabilities:
  - Pr(Delete) = 0.164322
  - Pr(Insert) = 0.071429
  - Pr(Mismatch) = 0.357143
  - Pr(Match) = 0.407106

- Cost of multistate distribution (English -> German) = 20179.95 bits
- MML inferred probabilities:
  - Pr(Delete) = 0.069480
  - Pr(Insert) = 0.189866
  - Pr(Mismatch) = 0.442394
  - Pr(Match) = 0.298260

- Note that an extra cost of log(26) is needed for each mismatch and log(27) for each insert to state the new character.

Conclusion

- MML methods have managed to
  - infer the correct phylogenetic tree/graphs for artificial languages.
  - infer phylogenetic trees/graphs for languages by encoding them in terms of one another.
- We can not (or can we?) conclude that one language really descends from another language. We can only conclude that they are related.

Future work :

- Compression – grammar and vocabulary.
- Compression – phonemes of languages.
- Endangered languages – Indigenous languages.
- Refine coding scheme.
  - Some characters occur more frequently than others.
    E.g.: English - “x” compared with “a”.
  - Some characters are more likely to mutate from one language to another language.

Questions?

Some further reading on MML

- C. S. Wallace and P. R. Freeman. Single factor analysis by MML estimation. Journal of the Royal Statistical Society. Series B, 54(1):195-209, 1992.
- C. S.Wallace. Multiple factor analysis by MML estimation. Technical Report CS 95/218, Department of Computer Science, Monash University, 1995.
- C. S. Wallace and D. L. Dowe. MML estimation of the von Mises concentration parameter. Technical Report CS 93/193, Department of Computer Science, Monash University,1993.
- C. S. Wallace and D. L. Dowe. Refinements of MDL and MML coding. The Computer Journal, 42(4):330-337, 1999.
- P. J. Tan and D. L. Dowe. MML inference of decision graphs with multi-way joins. In Proceedings of the 15th Australian Joint Conference on Artificial Intelligence, Canberra, Australia, 2-6 December 2002, published in Lecture Notes in Artificial Intelligence (LNAI) 2557, pages 131-142. Springer-Verlag, 2002.
- S. L. Needham and D. L. Dowe. Message length as an effective Ockham's razor in decision tree induction. In Proceedings of the 8th International Workshop on Artificial Intelligence and Statistics (AI+STATS 2001), Key West, Florida, U.S.A., January 2001, pages 253-260, 2001
- Y. Agusta and D. L. Dowe. Unsupervised learning of correlated multivariate Gaussian mixture models using MML. In Proceedings of the Australian Conference on Artificial Intelligence 2003, Lecture Notes in Artificial Intelligence (LNAI) 2903, pages 477-489. Springer-Verlag, 2003.
- J. W. Comley and D. L. Dowe. General Bayesian networks and asymmetric languages. In Proceedings of the Hawaii International Conference on Statistics and Related Fields, June 5-8, 2003, 2003.
- J. W. Comley and D. L. Dowe. Minimum Message Length, MDL and Generalised Bayesian Networks with Asymmetric Languages, chapter 11, pages 265-294. M.I.T. Press, 2005. [Camera ready copy submitted October 2003].
- P. J. Tan and D. L. Dowe. MML inference of oblique decision trees. In Proc. 17th Australian Joint Conference on Artificial Intelligence (AI04), Cairns, Qld., Australia, pages 1082-1088. Springer-Verlag, December 2004.

- www.csse.monash.edu.au/~dld/CSWallacePublications