Back to the roots

using regular sound correspondences for linguistic phylogeny (as one should)

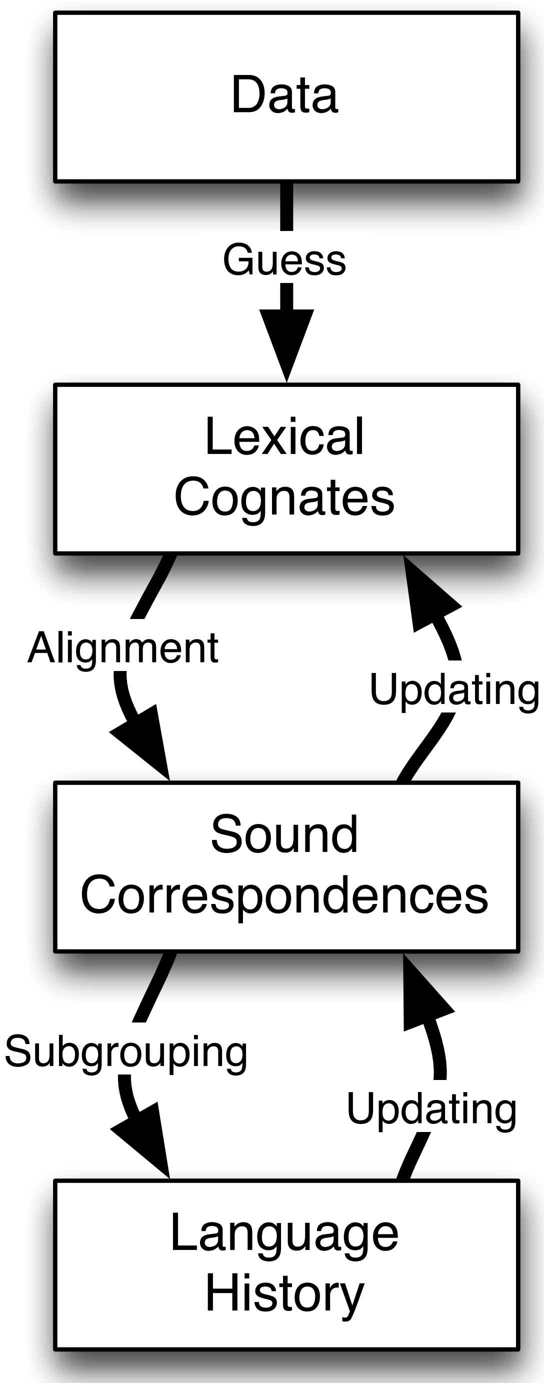
Michael Cysouw
Philipps Universität Marburg

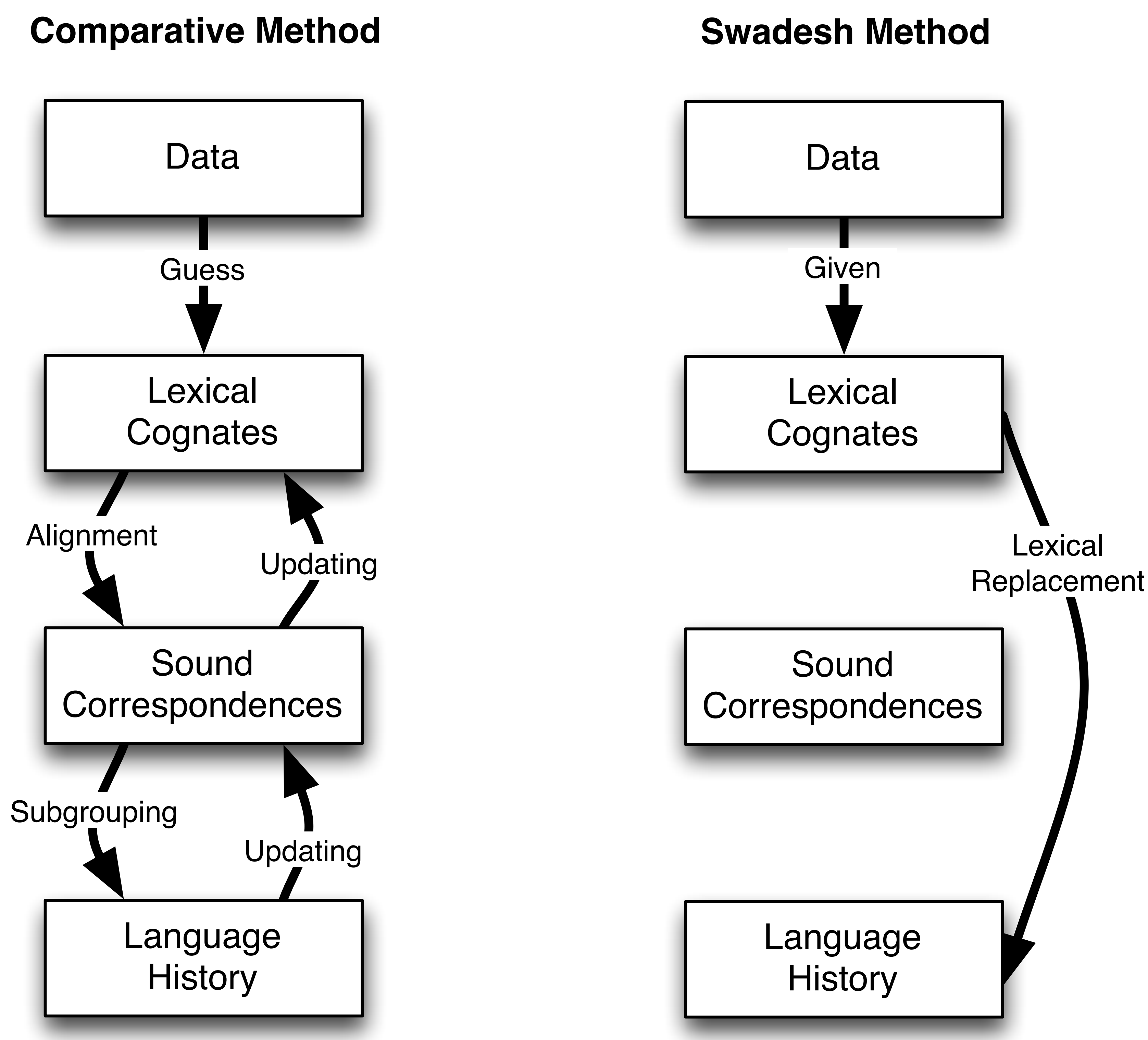
Bern, 24 November 2012

Methods for Lexical Comparison

- Comparative Method
- Swadesh Method
- Black Box Method
- Alignment Method
- Sound Change Method

Comparative Method




Comparative Method
Data
Guess
Lexical Cognates
Alignment
Updating
Sound Correspondences
Subgrouping
Updating
Language History
Swadesh Method
Data
Given
Lexical Cognates
Lexical Replacement
Sound Correspondences
Language History

Swadesh Method

- **Necessary prerequisite**
  known cognacy in parallel wordlist
- **Method**
  Reconstruct history from
  distribution of cognates

Swadesh Method

- **Language distances**
  Swadesh (1952)
- **Stochastic model with first tree**
  Sankoff (1969), Dobson (1969)
- **Modern replacement model**
  Gray & Jordan (2000), (Gleason 1959)

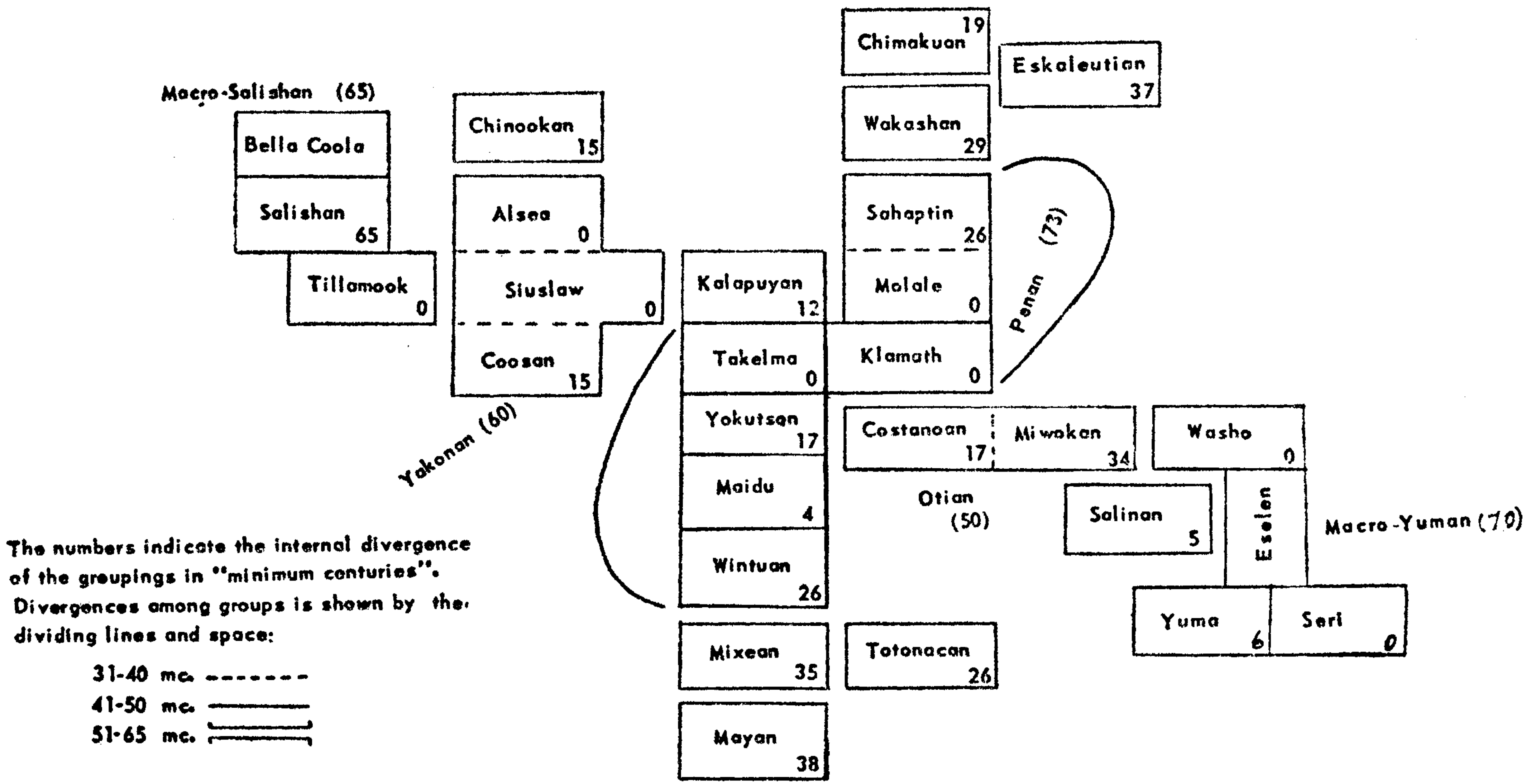


FIG. 2. Internal and external relationships of the Penan micro-phylum.

Swadesh, Morris. 1952. Lexico-Statistic Dating of Prehistoric Ethnic Contacts: With Special Reference to North American Indians and Eskimos. Proceedings of the American Philosophical Society 96(4). 452-463.

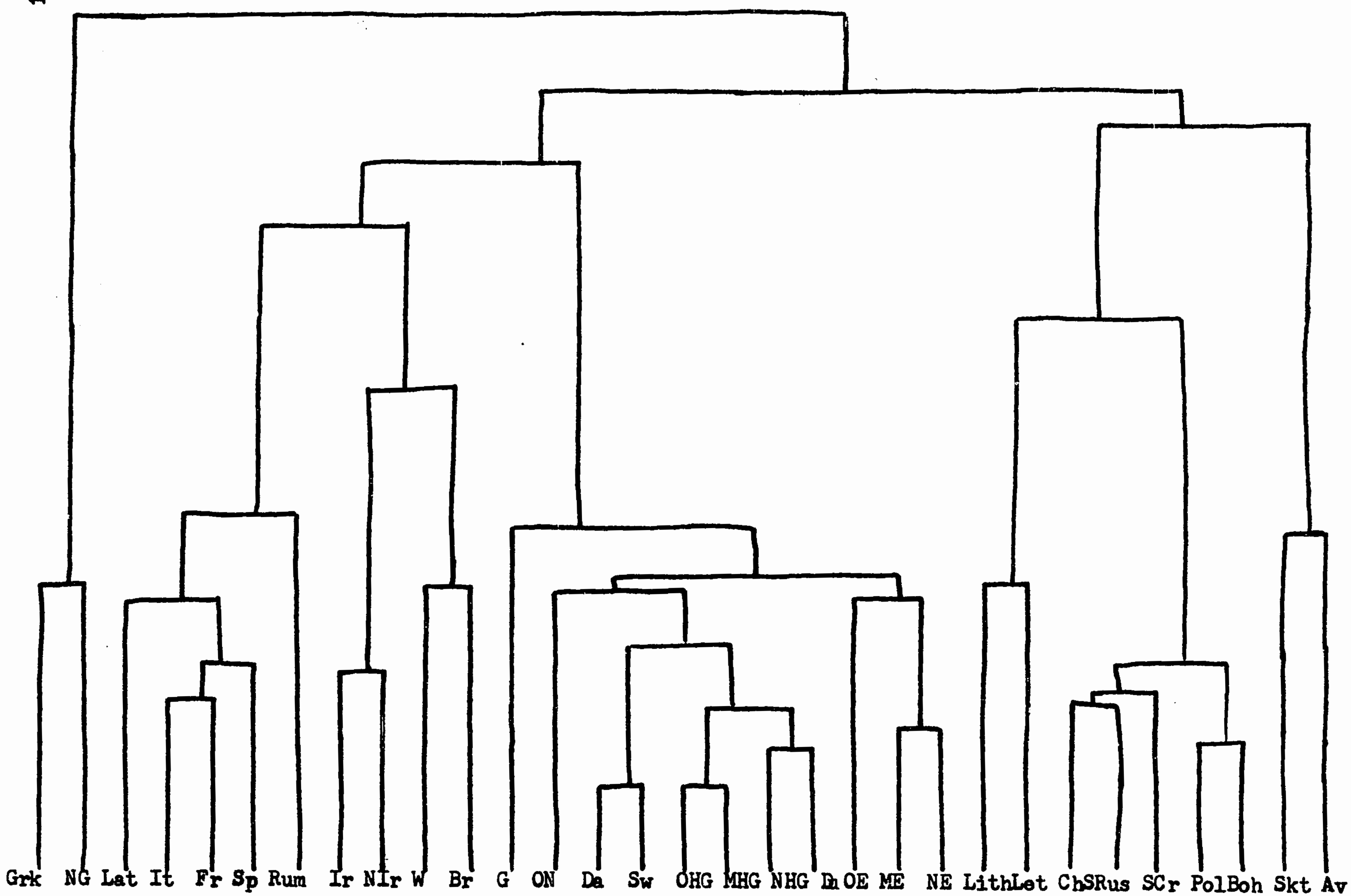
Swadesh, Morris. 1959, "Linguistics As An Instrument of Prehistory. Southwestern Journal of Anthropology 15, no. 1: 20-35.

David Sankoff

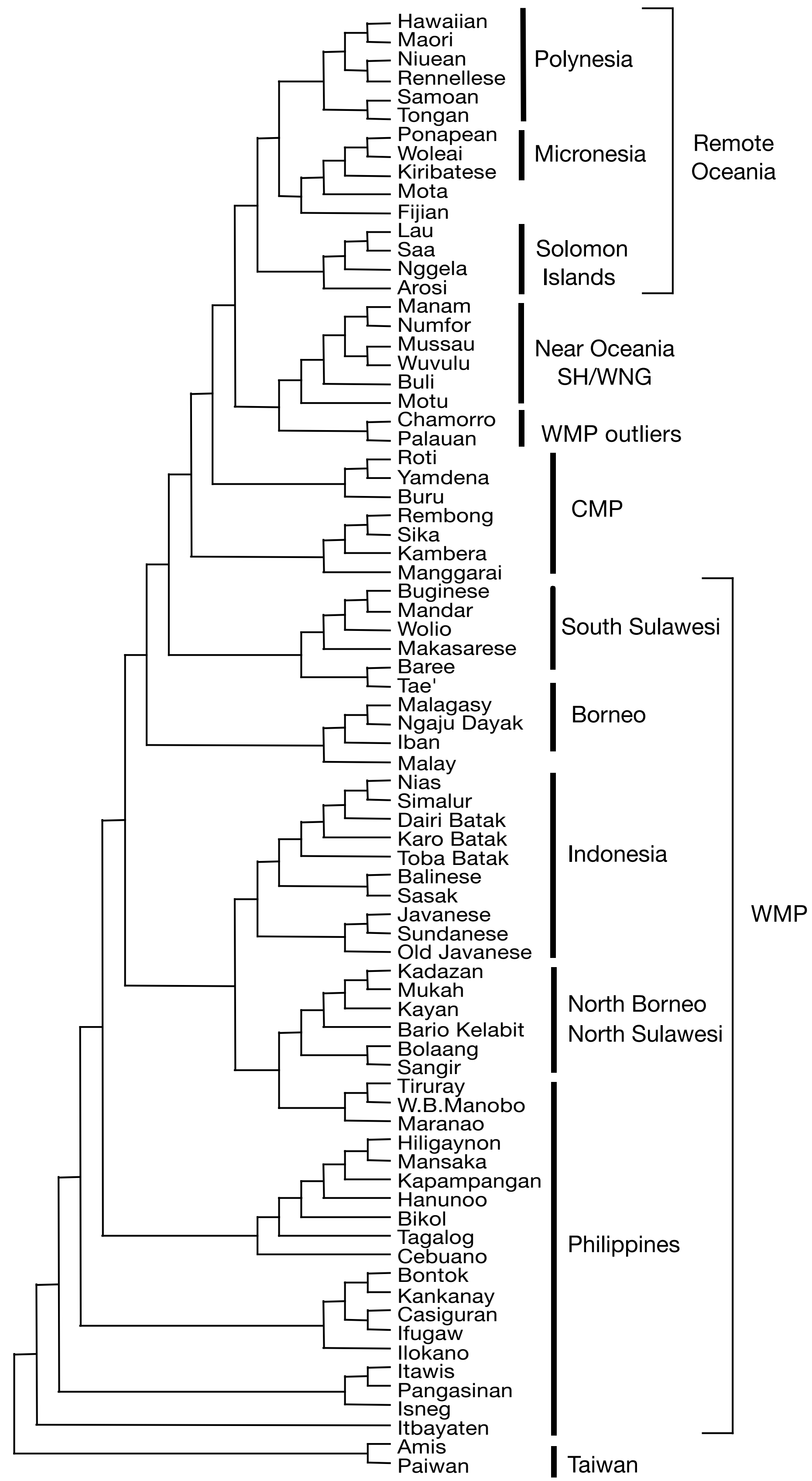
Annette J. Dobson

Dobson, A J. "Lexicostatistical Grouping." Anthropological Linguistics 11, no. 7 (1969): 216-221.
Dobson, Annette J. "Unrooted Trees for Numerical Taxonomy." Journal of Applied Probability 11, no. 1 (1974).

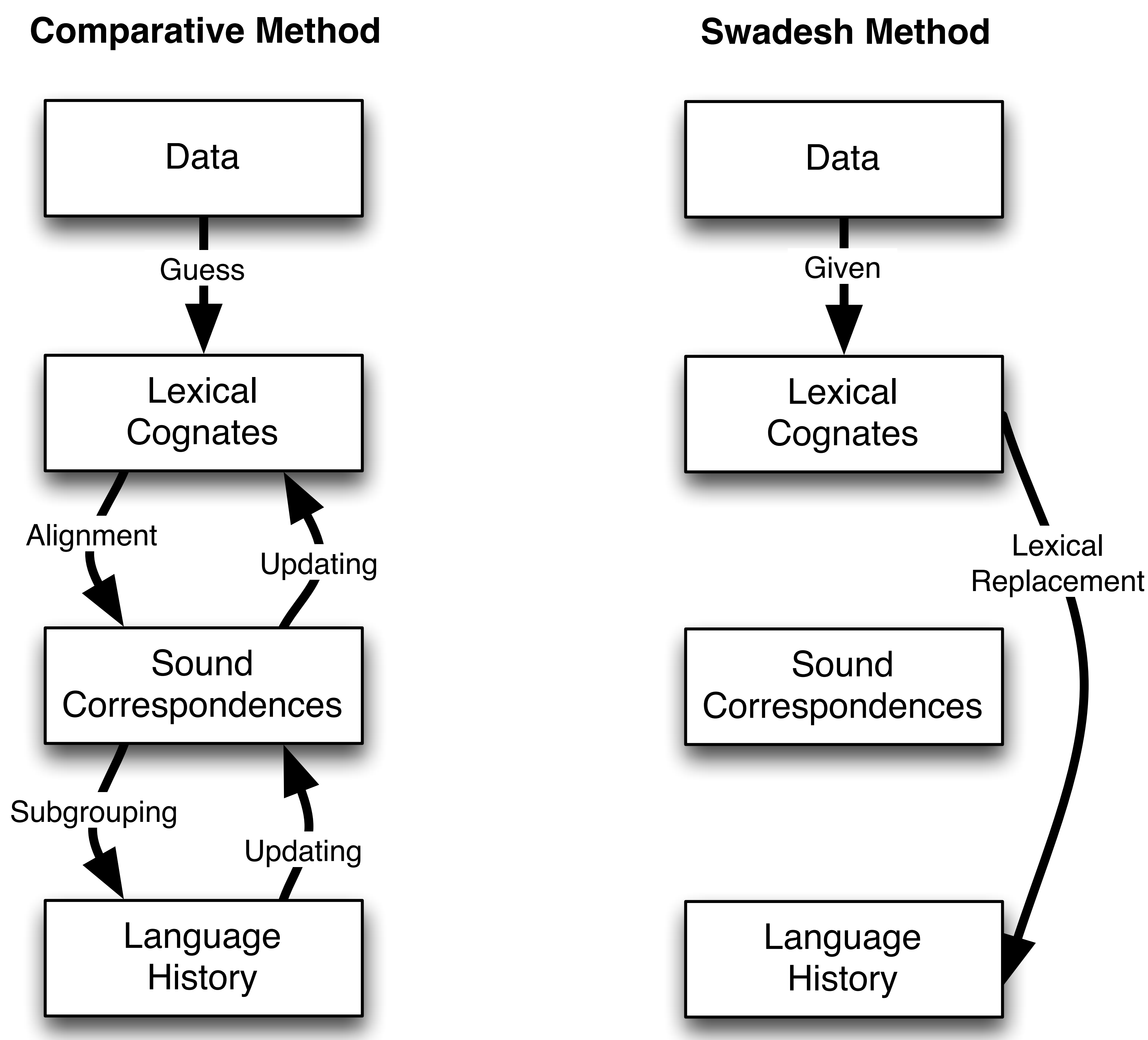
Fig. 16. Tree of Indo-European languages as reconstructed from data in Table 4.

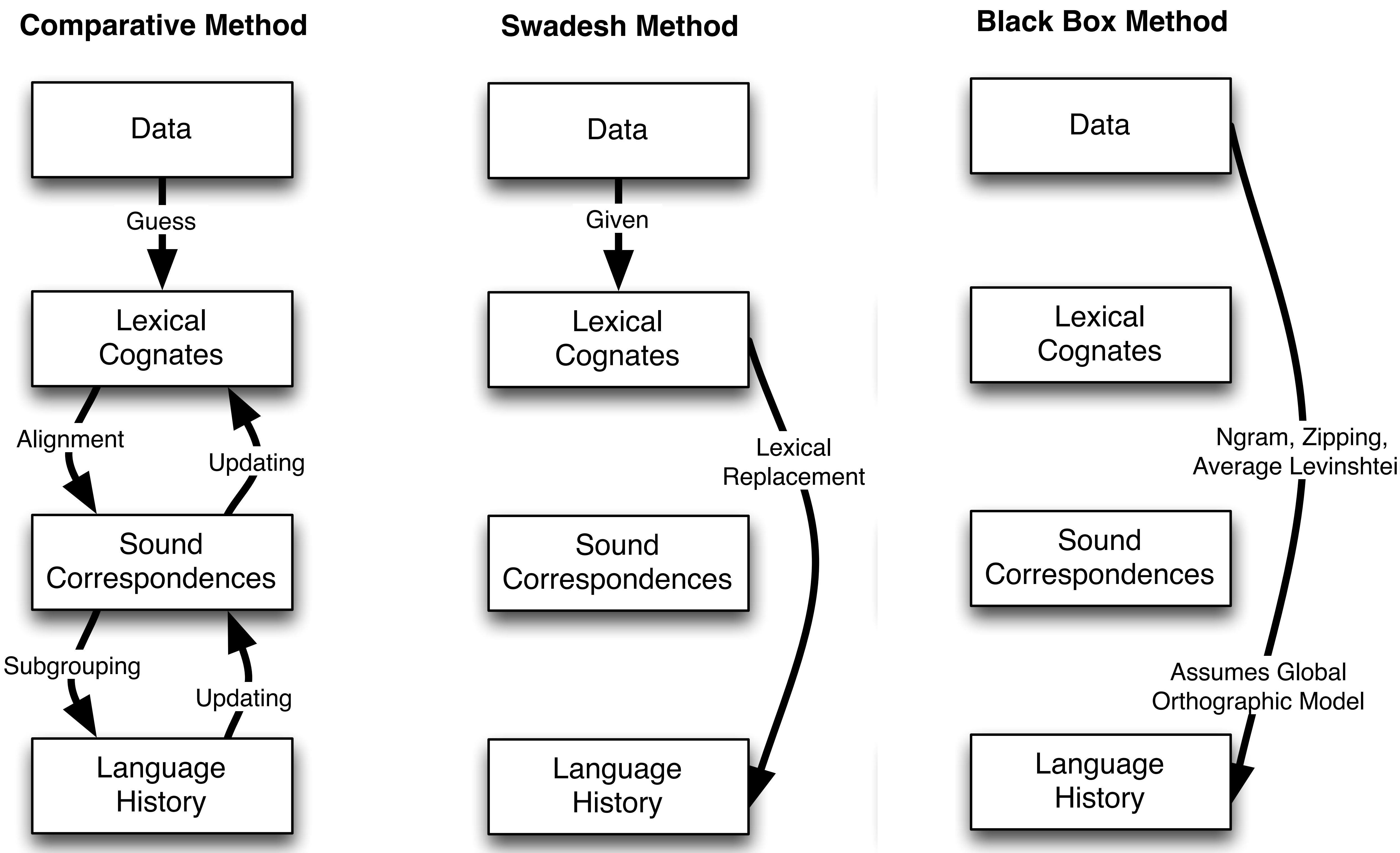


Sankoff, D. "Historical Linguistics As Stochastic Process." PhD Dissertation. McGill University, 1969.



Gray, Russell D, and Fiona M Jordan. "Language Trees Support the Express-train Sequence of Austronesian Expansion." Nature 405 (2000): 1052-1055.
Gleason, H A. "Counting and Calculating for Historical Reconstruction." Anthropological Linguistics 1, no. 2 (1959): 22-32.


Comparative Method
Data
Guess
Lexical Cognates
Alignment
Updating
Sound Correspondences
Subgrouping
Updating
Language History
Swadesh Method
Data
Given
Lexical Cognates
Lexical Replacement
Sound Correspondences
Language History


Comparative Method
Data
Guess
Lexical Cognates
Alignment
Updating
Sound Correspondences
Subgrouping
Updating
Language History
Swadesh Method
Data
Given
Lexical Cognates
Lexical Replacement
Sound Correspondences
Language History
Black Box Method
Data
Lexical Cognates
Ngram, Zipping, Average Levinshtein
Sound Correspondences
Assumes Global Orthographic Model
Language History

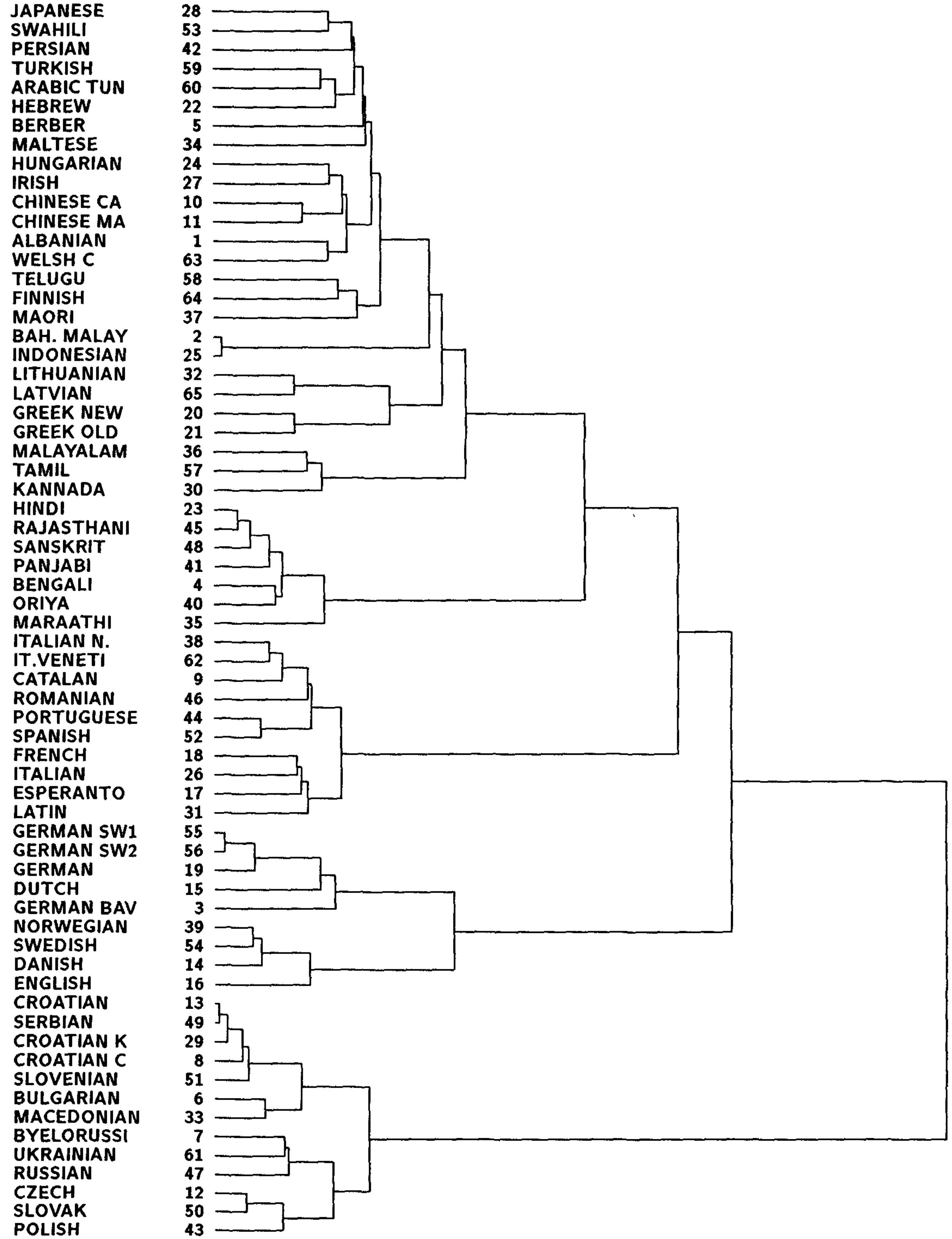
Black Box Methods

- **Necessary prerequisite**
  Global orthographic model
  ("same alphabet for all data")
- **Method**
  Compare global orthographic similarity

Black Box Methods

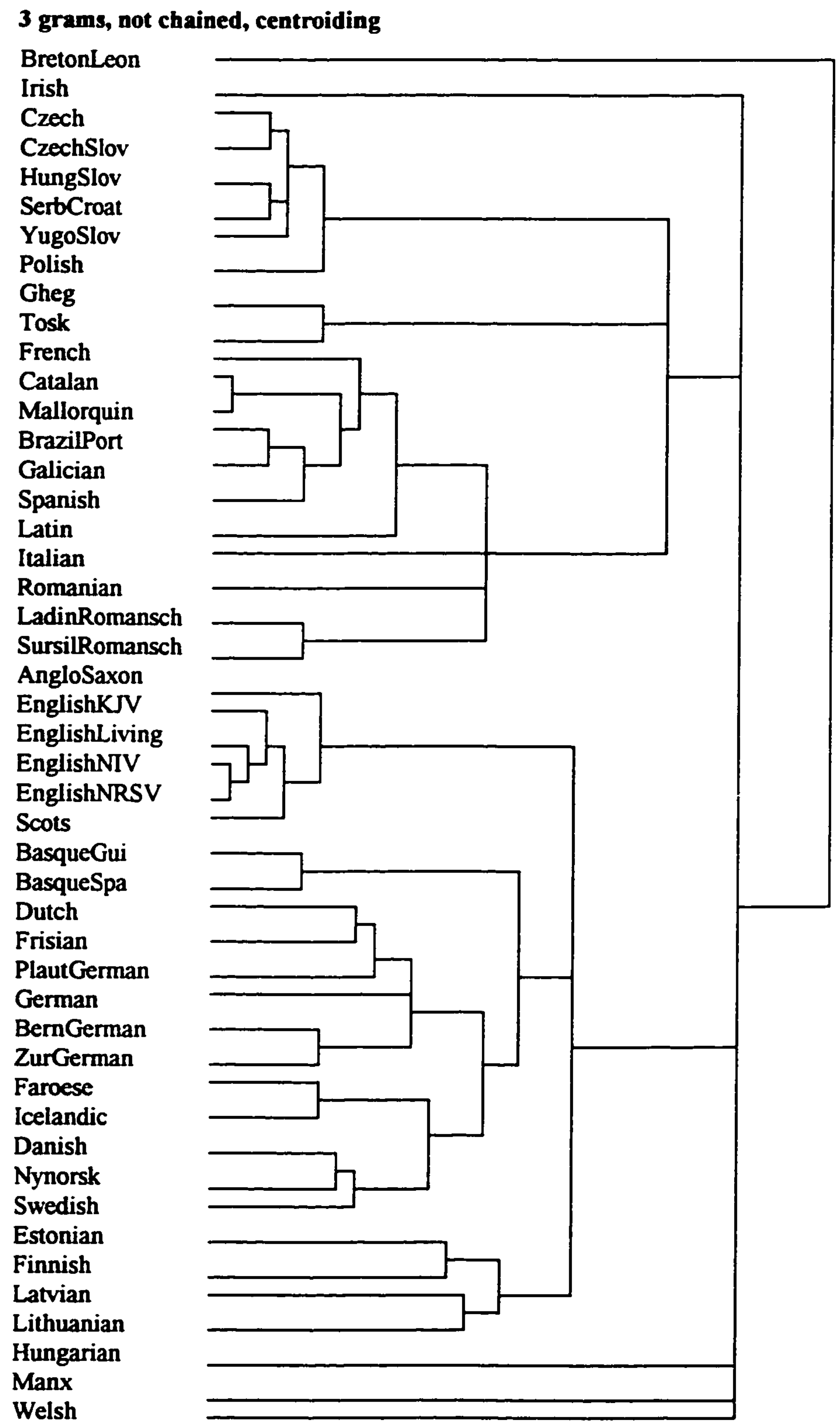
- **Aggregate Levenshtein Distance**
  Batagelj et al. (1992), Kessler (1995)
- **N-Gram Similarity**
  Huffman (1998)
- **Zipping Distance**
  Benedetto et al. (2002)

Aggregate Levenshtein Distance



Batagelj, V, T Pisanski & D Kerži. 1992. Automatic clustering of languages. Computational Linguistics 18(3). 339-352.

N-Gram Similarity



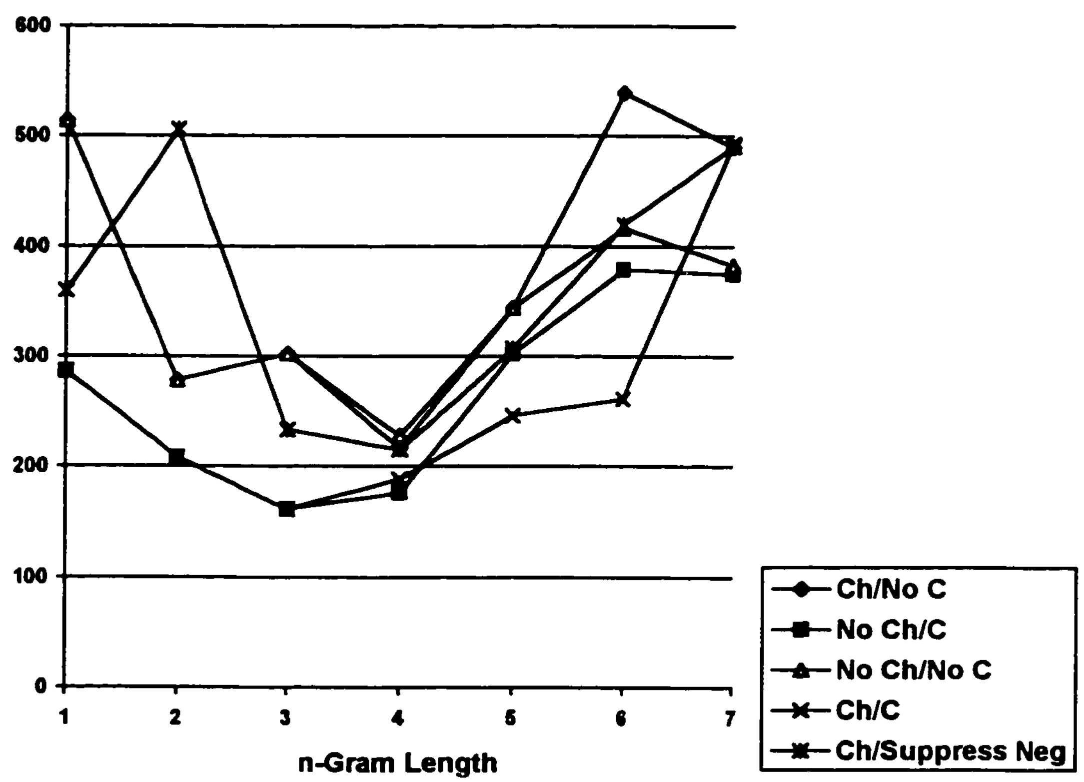
Huffman, Stephen M. "The Genetic Classification of Languages by N-Gram Analysis." PhD Thesis, Washington, D.C.: Georgetown University, 1998.

Appendix 9. Divergence from reference classification

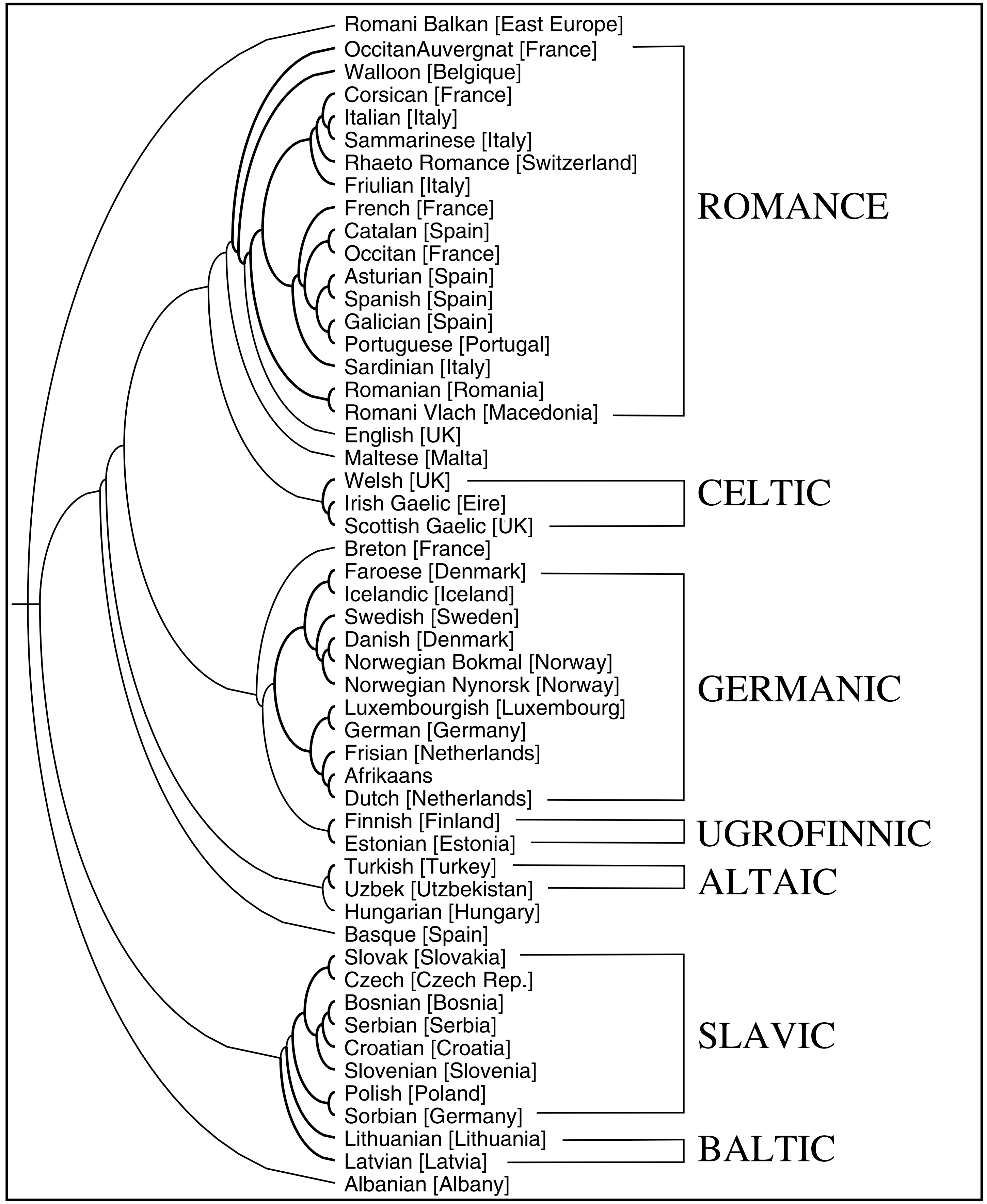
Text display:

N-gram Length	1	2	3	4	5	6	7
Chained, No centroid	514	278	302	217	344	539	490
Not Chained, Centroid	286	208	161	176	302	379	375
Not Chained, No Centroid	514	278	302	228	344	417	383
Chained, Centroid	286	208	161	188	246	261	493
Suppress Negative	360	506	233	215	307	420	491

Divergence from Reference Classification vs. n-Gram Length



Zipping



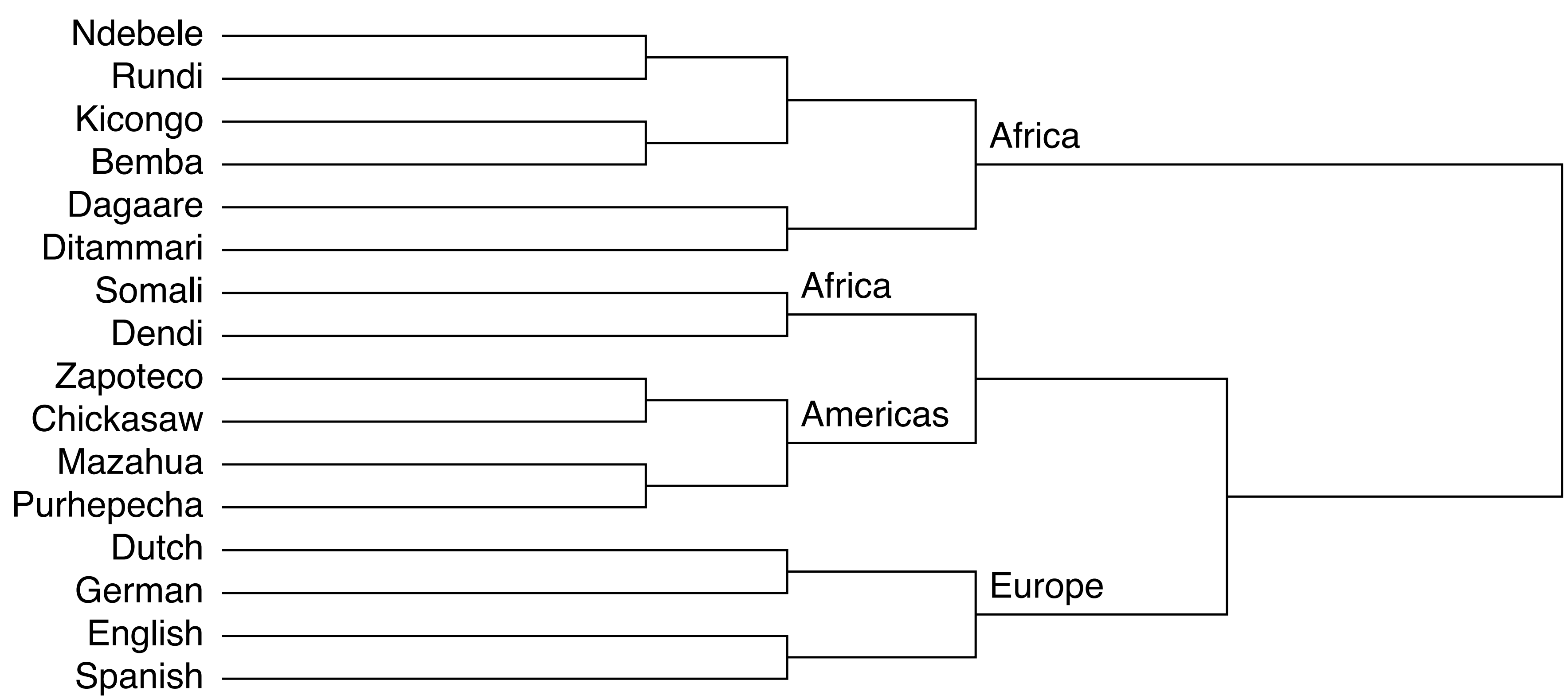
Benedetto, Dario, Emanuele Caglioti, and Vittorio Loreto. "Language Trees and Zipping." Phys. Rev. Lett. 88, no. 4 (2002)

Zipping

Normalized Compression Distance

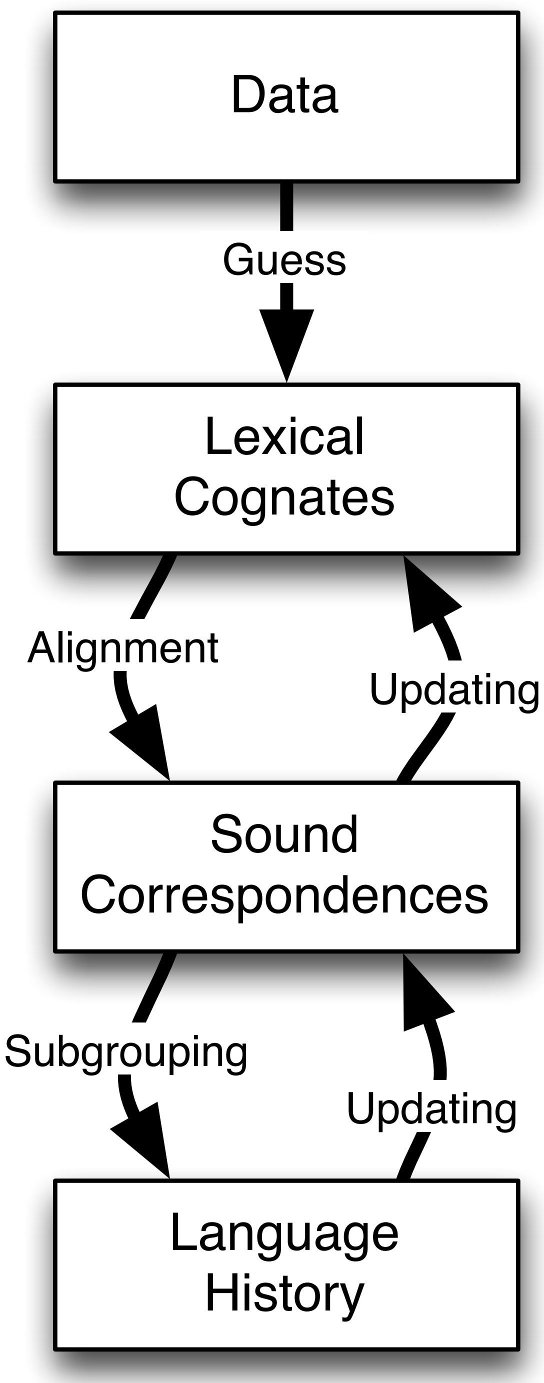
(C = compressed size)

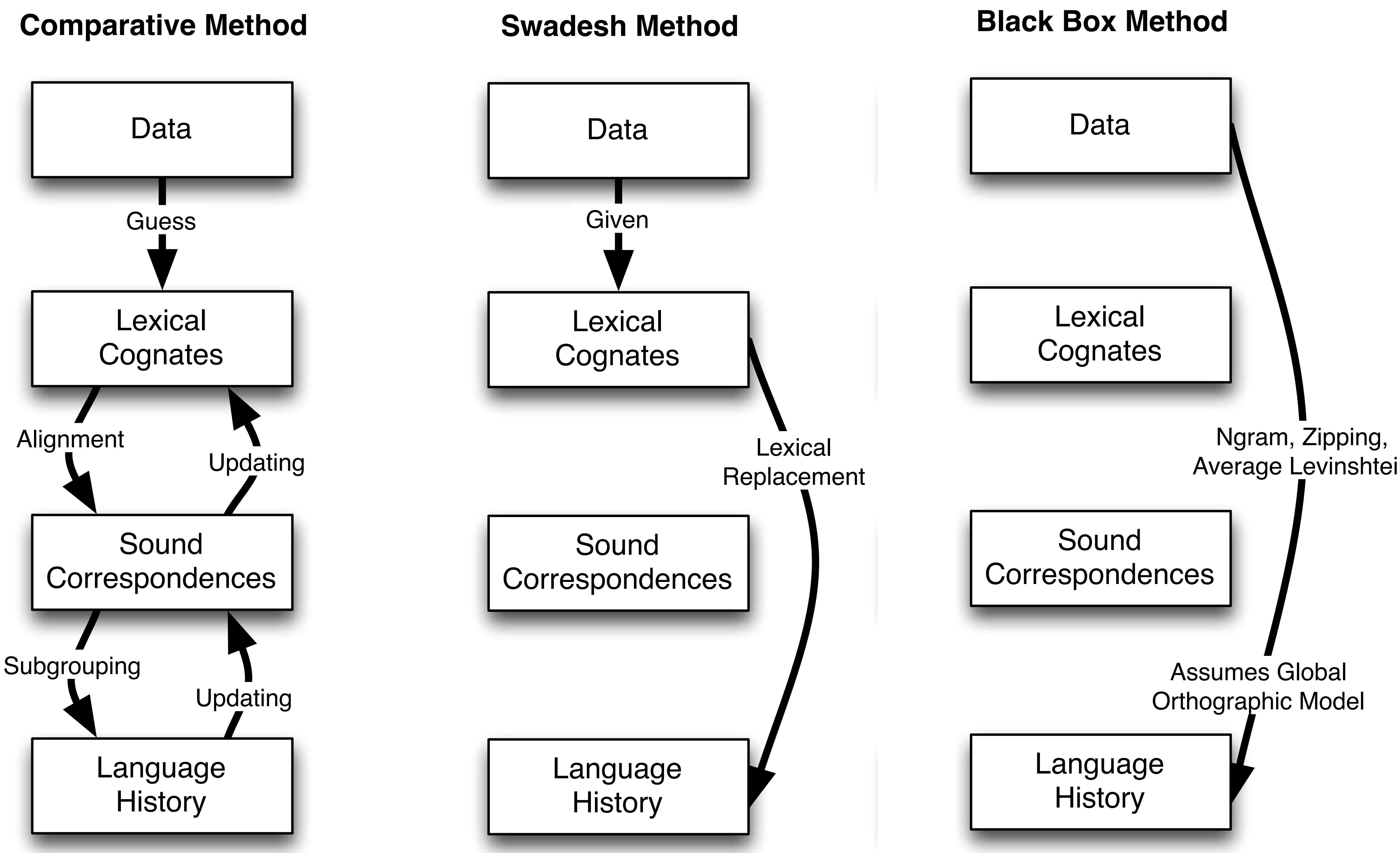
$$NCD(x,y) = \frac{C(xy) - \min\{C(x), C(y)\}}{\max\{C(x), C(y)\}}.$$

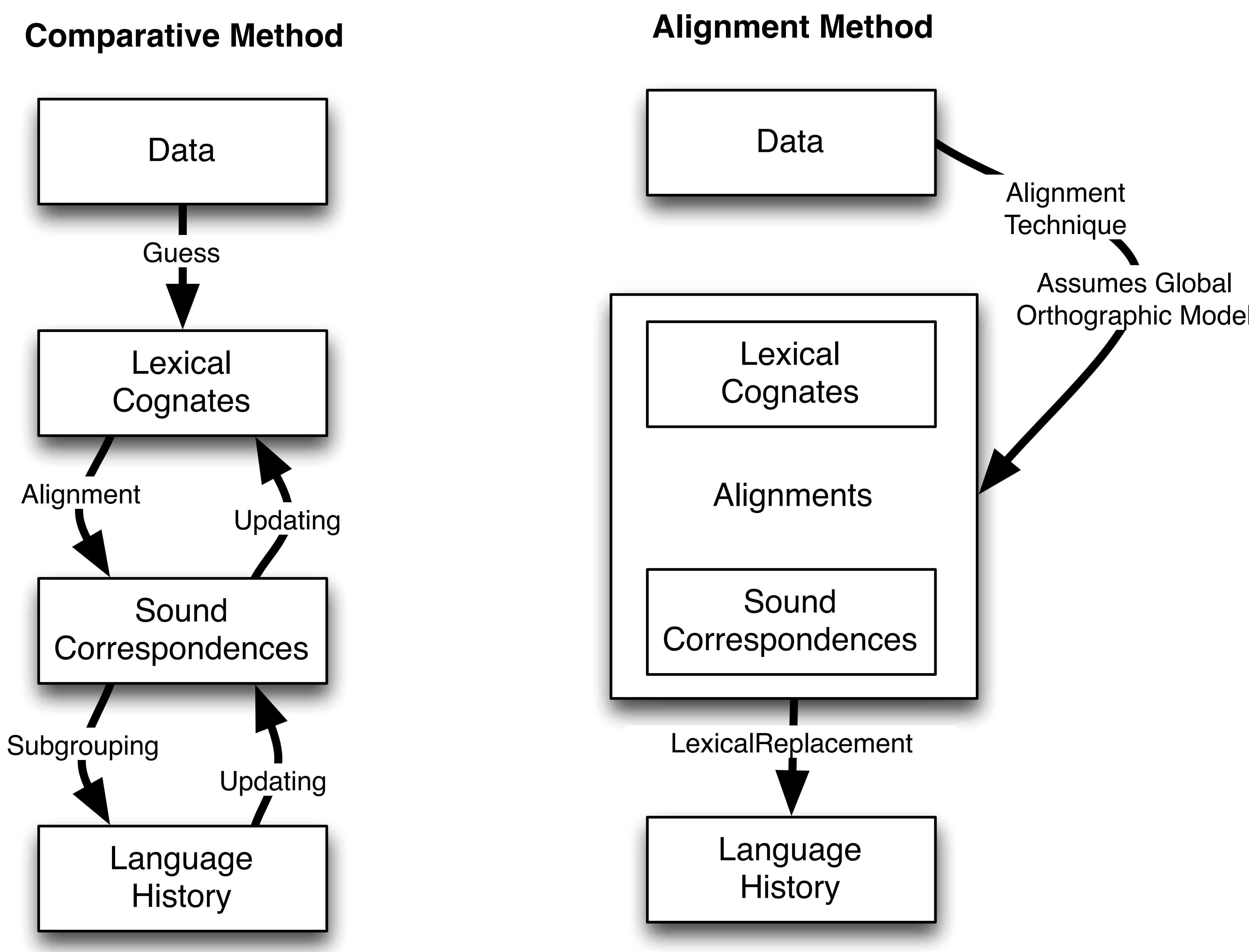


Cilibrasi, R, and P.M.B. Vitanyi. "Clustering by Compression." IEEE Transactions on Information Theory 51, no. 4 (2005),

Comparative Method




Comparative Method
Data
Guess
Lexical Cognates
Alignment
Updating
Sound Correspondences
Subgrouping
Updating
Language History
Swadesh Method
Data
Given
Lexical Cognates
Lexical Replacement
Sound Correspondences
Language History
Black Box Method
Data
Lexical Cognates
Ngram, Zipping, Average Levinshtein
Sound Correspondences
Assumes Global Orthographic Model
Language History


Comparative Method
Data
Guess
Lexical Cognates
Alignment
Updating
Sound Correspondences
Subgrouping
Updating
Language History
Alignment Method
Data
Alignment Technique
Assumes Global Orthographic Model
Lexical Cognates
Alignments
Sound Correspondences
LexicalReplacement
Language History

Alignment Method

- **Necessary prerequisite**
  Global orthography model and parallel wordlist
- **Method**
  Discover sound correspondences, possibly together with decision on cognacy

Alignment Method

- **Alignment**
  Covington (1996, 1998)
- **Relation to Levenshtein**
  Kondrak (2000)
- **Multiple Alignment**
  Bhargava & Kondrak (2009), Prokić et al. (2009)

Spanish/Italian/French ‘five’:

θ	i	ŋ	k	-	o
c	i	ŋ	k	w	e
s	ɛ̃	-	k	-	-

Covington, Michael A. "Alignment of Multiple Languages for Historical Comparison." In Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics. 1998.

	Covington's alignments	*ALINE's alignments*
three : trēs	θ r i y	‖ θ r iy ‖
	t r ē s	‖ t r ē ‖ s
blow : flāre	b l - - o w	‖ b l o ‖ w
	f l ā r e -	‖ f l ā ‖ re
full : plēnus	f - - - u l	‖ f u l ‖
	p l ē n u s	‖ p - l ‖ ēnus
fish : piscis	f - - - i š	‖ f i š ‖
	p i s k i s	‖ p i s ‖ kis
I : ego	- - a y	‖ ay ‖
	e g o -	‖ e ‖ go
tooth : dentis	- - - t u w θ	‖ t uw θ ‖
	d e n t i - s	den ‖ t i s ‖

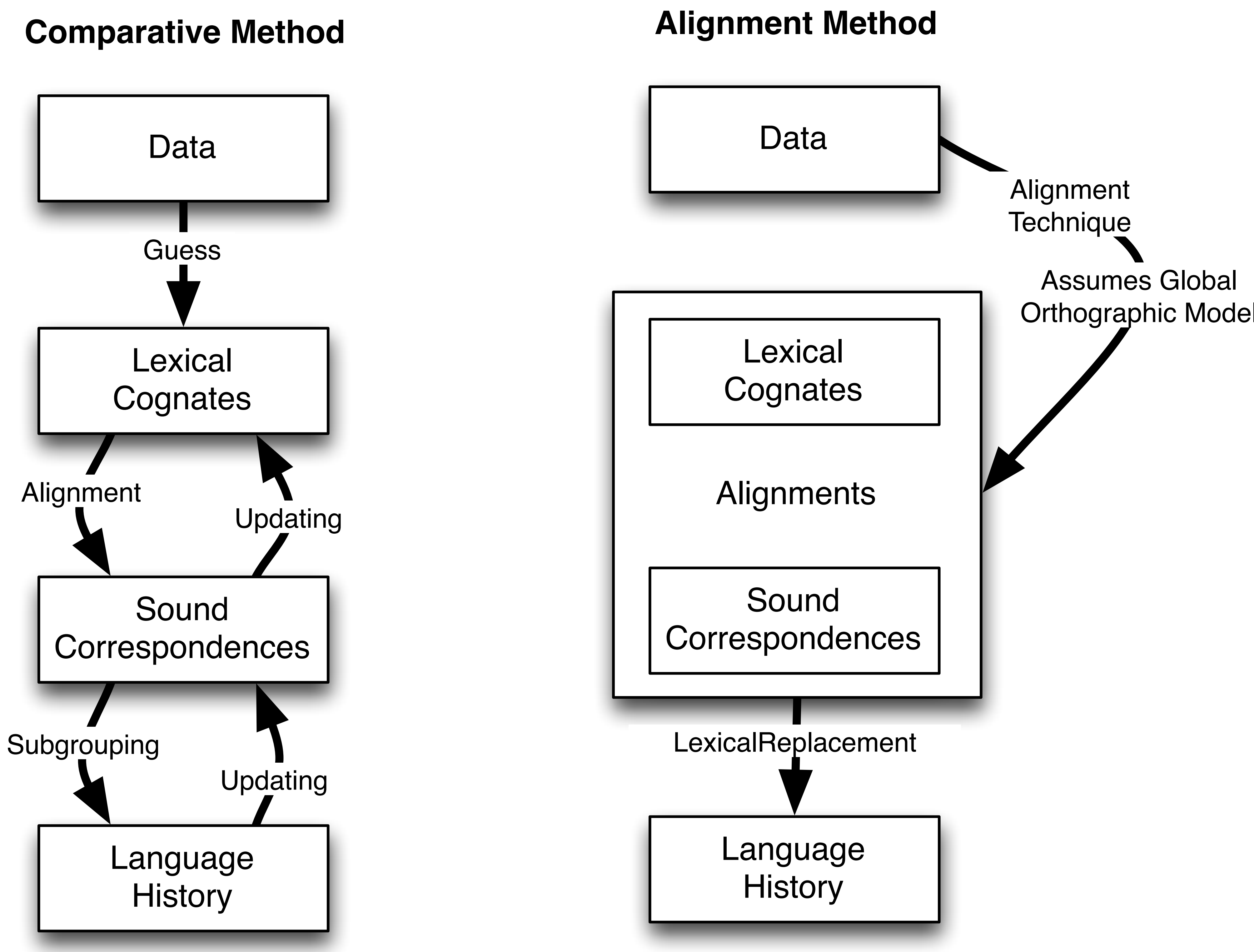
Kondrak, Grzegorz. "A New Algorithm for the Alignment of Phonetic Sequences." In Proceedings of the 1st North American Chapter of the Association for Computational Linguistics Conference. Morgan Kaufmann Publishers Inc., 2000.

```
D--E--N-
Z--E--N-
DZIE--N-
DI-E--NA
D--I--A-
D--I--E-
Z--U--E-
J--O--UR
DJ-O--U-
```

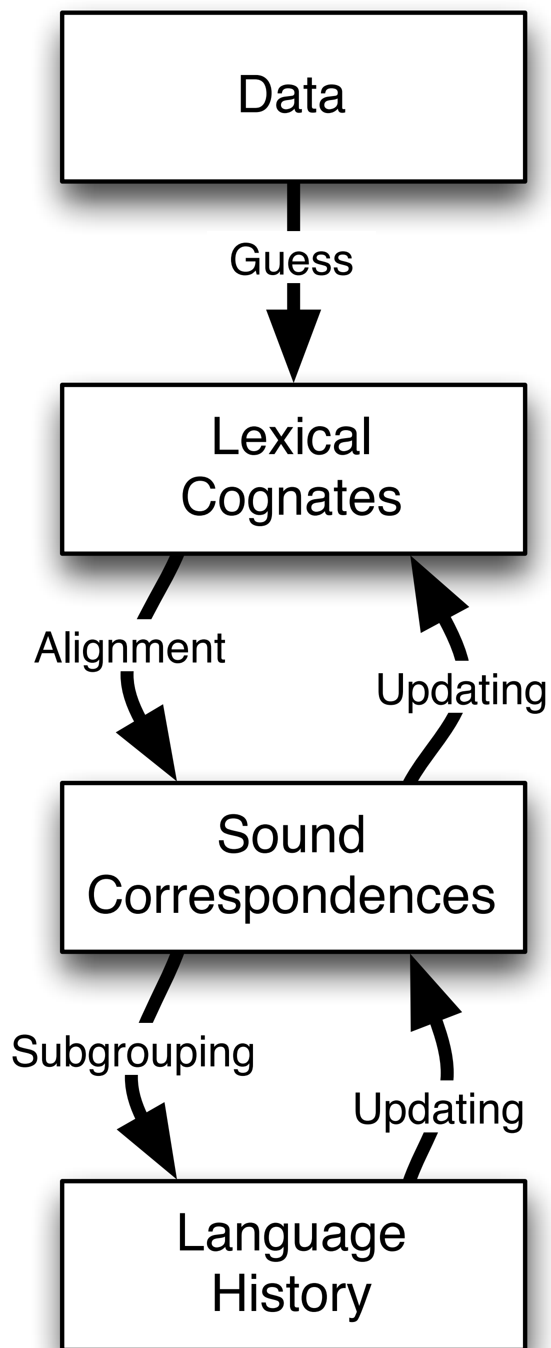
Bhargava, Aditya, and Grzegorz Kondrak. "Multiple Word Alignment with Profile Hidden Markov Models." In Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Student Research Workshop and Doctoral Consortium (NAACL-HLT 2009). Boulder, Colorado: Association for Computational Linguistics, 2009.

j	ˈɑ	-	-	-	-
j	ˈɑ	z	e	-	-
-	ˈɑ	s	-	-	-
j	ˈɑ	s	-	-	-
j	ˈɑ	z	e	k	a
j	ˈɛ	-	-	-	-
-	ˈɒ	s	-	-	-

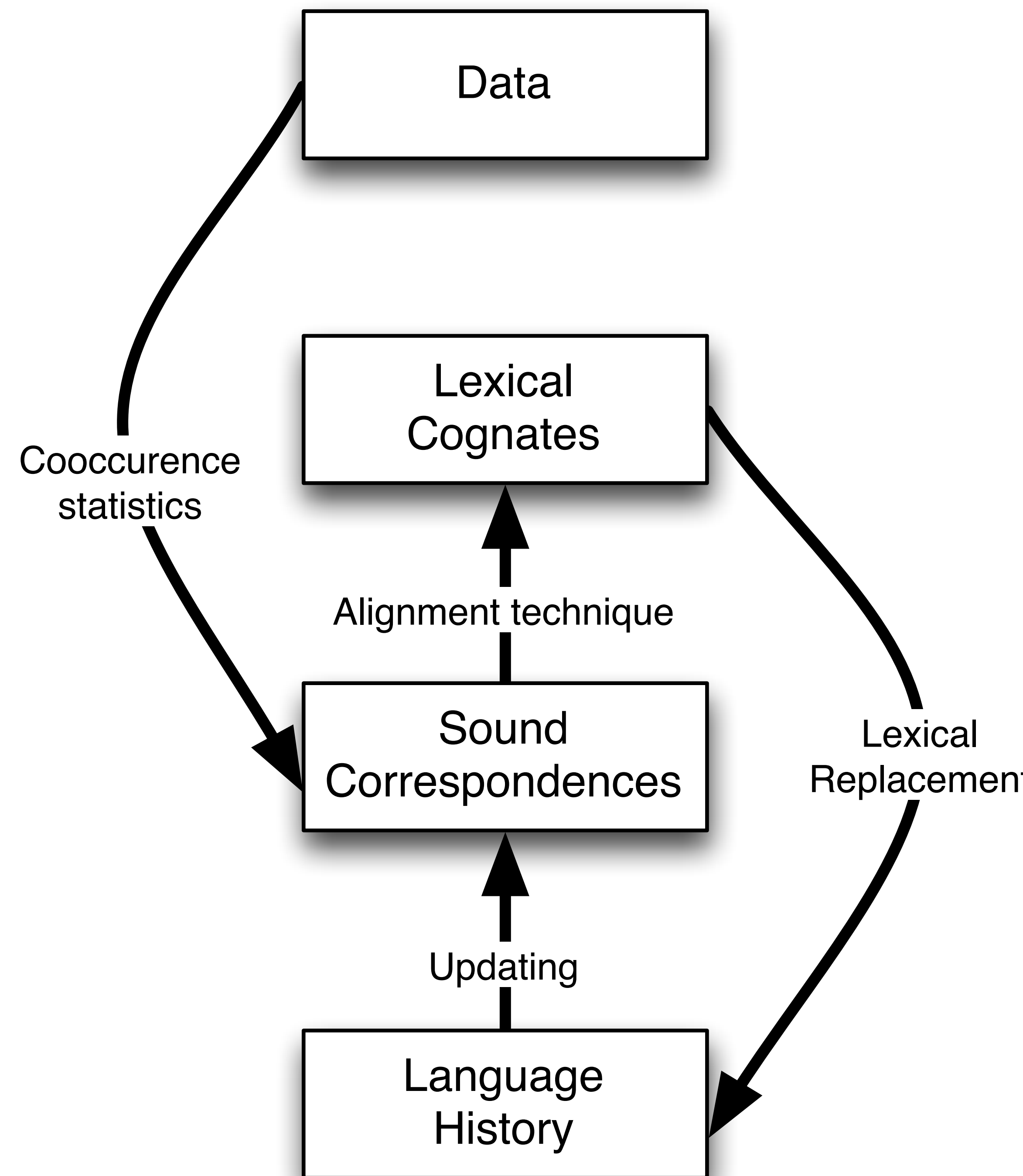
Prokić, Jelena, Martijn Wieling, and John Nerbonne. "Multiple Sequence Alignments in Linguistics." In Proceedings of the EACL 2009 Workshop on Language Technology and Resources for Cultural Heritage, Social Sciences, Humanities, and Education (LaTeCH-SHELT&R 2009). Association for Computational Linguistics, 2009.


Comparative Method
Data
Guess
Lexical Cognates
Alignment
Updating
Sound Correspondences
Subgrouping
Updating
Language History
Alignment Method
Data
Alignment Technique
Assumes Global Orthographic Model
Lexical Cognates
Alignments
Sound Correspondences
LexicalReplacement
Language History

Comparative Method



Sound Change Method

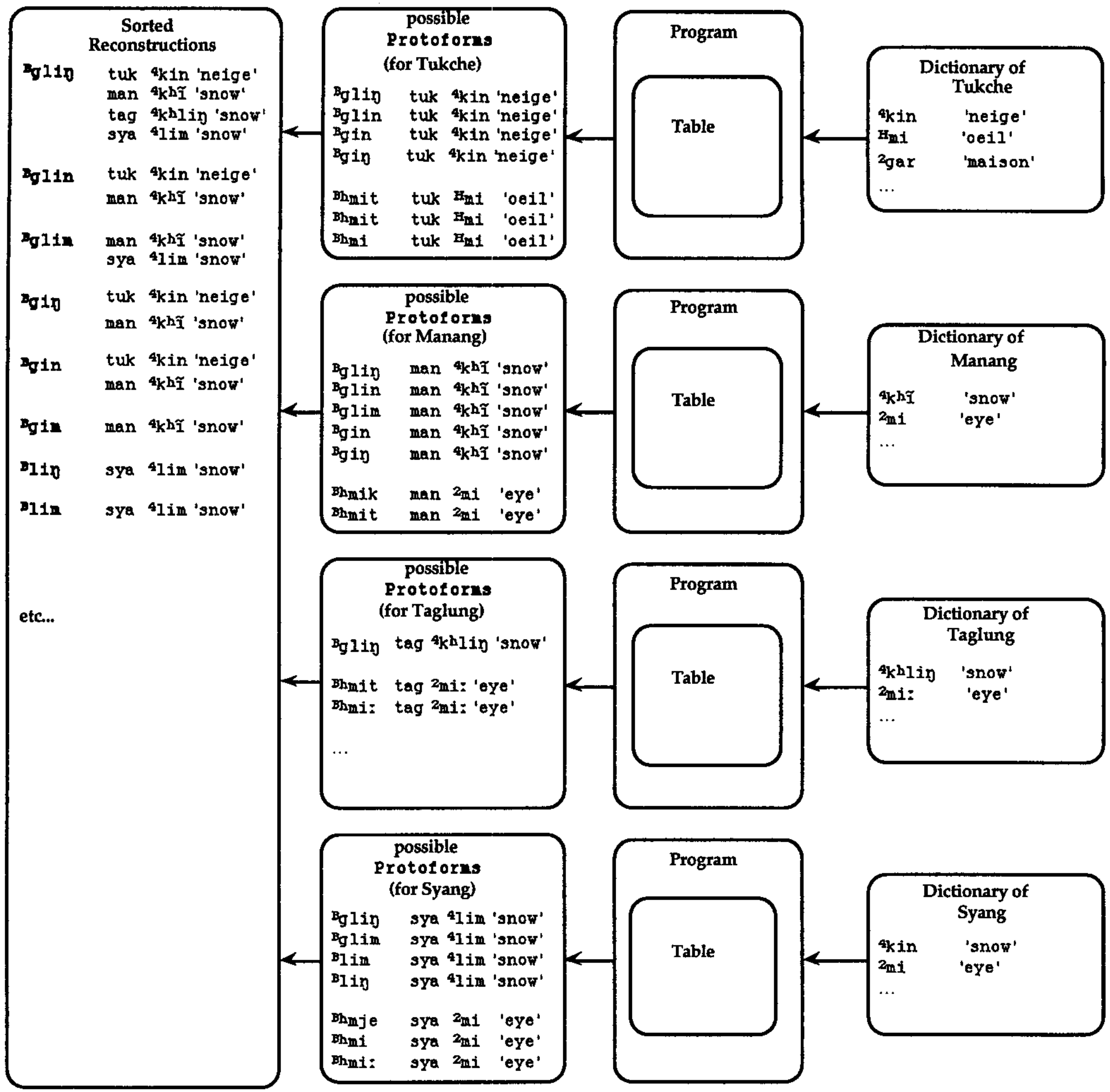


Sound Change Method

- **Necessary prerequisite**
  (Large) parallel wordlist,
  but ***no*** global harmonized orthography
- **Method**
  Discover sound correspondences through regularity, and use that to discover cognacy

Sound Change Method

- **Rule-based automatic reconstruction**
  (Hewson 1973, Lowe & Mazaudon 1994)
- **Sound co-occurrence**
  Ross (1947)
- **Cross-script mapping**
  Cysouw & Jung (2007)



Lowe, John Brandon, and Martine Mazaudon. "The Reconstruction Engine: A Computer Implementation of the Comparative Method." Computational Linguistics 20, no. 3 (1994): 381-417.

Now let me put the matter more rigidly. Suppose I take 1000 common ideas, numerals, parts of the body, names of relatives, etc. and express them in English and German. I now prepare a Table of the following kind:

English ↓ ← German

	[pf]	[ts]	[t]	[f]	[h]..	....
[p]	x x x x x		x			
[f]		x		x x x[1] x x x		
[t]	x	x x x[2] x x x x			x	
[d]			x x x[3] x		x[4]	
. . . .						

Ross, Alan S C. "Philologica Mathematica." Časopis Pro Moderní Filologii 31 (1947): 102-108, 185-191.

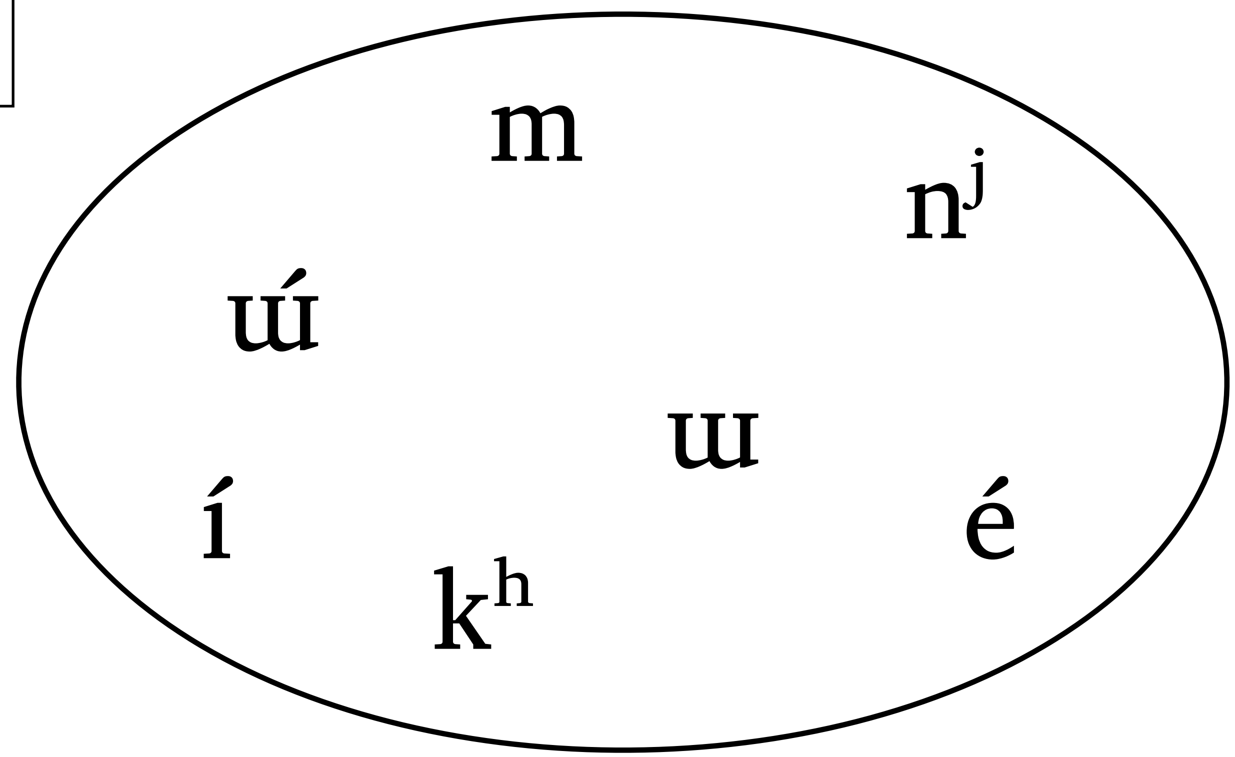
Cross-script mapping

E	R	freq	dice
r	р	184	0.88874745
n	н	115	0.8461936
l	л	104	0.79646295
s	с	114	0.7927922
t	т	165	0.7701921
m	м	47	0.7699933
o	о	184	0.7510106
k	ть	21	0.74458015
p	п	50	0.7388723
i	и	102	0.7034591
a	а	221	0.6866478
u	у	40	0.6449104
c	к	77	0.6251676
e	е	219	0.59066784
b	б	32	0.525643
w	в	46	0.46787763
d	д	42	0.381996
⋮	⋮	⋮	⋮

Cysouw, Michael & Hagen Jung. 2007. Cognate identification and alignment using practical orthographies. Proceedings of Ninth Meeting of the ACL Special Interest Group in Computational Morphology and Phonology, 109-116.

‘bag of symbol” approach

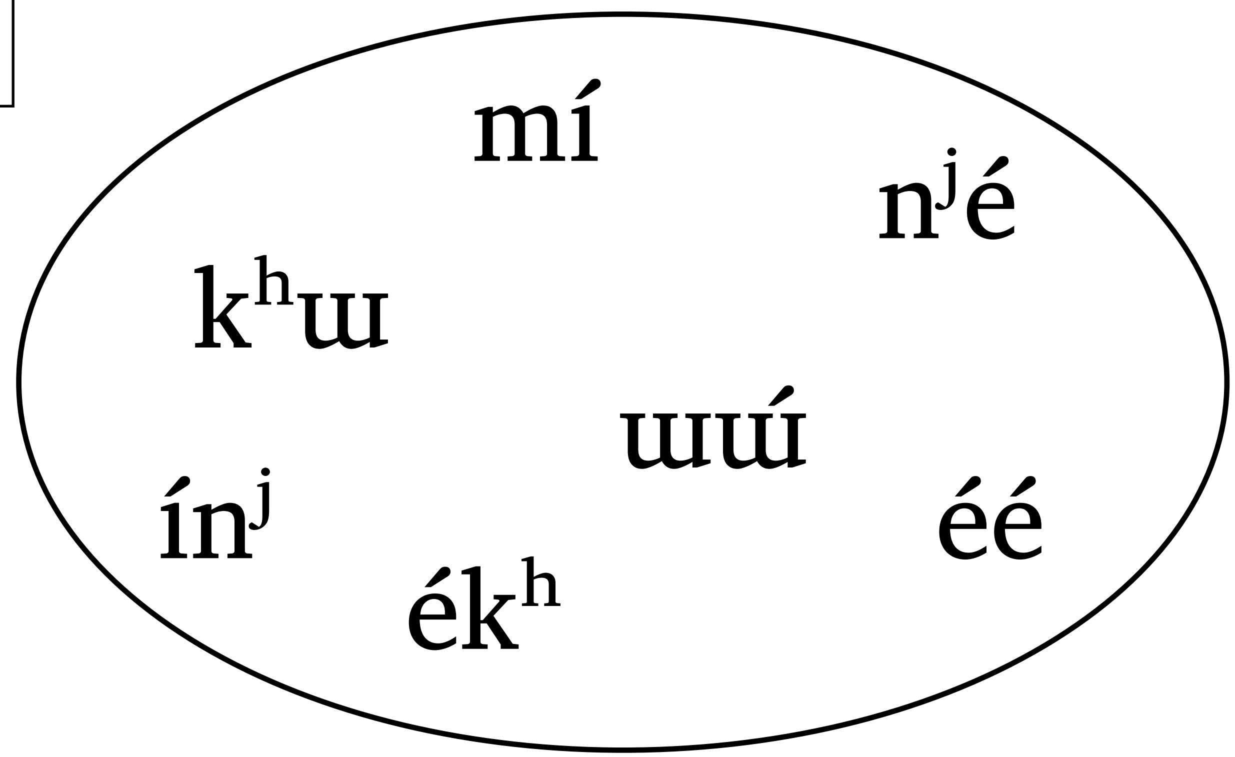
mín^{j}éék^{h}ɯɯ́



(1-grams)

‘bag of symbol” approach

mín^{j}éék^{h}ɯɯ́



(2-grams)

Cross-script mapping

Cross-script mapping

- Ignore linear structure of words
  "bag of symbols" approach

Cross-script mapping

- Ignore linear structure of words
  "bag of symbols" approach
- Use parallel wordlist to estimate
  co-occurrences of n-grams

Cross-script mapping

- Ignore linear structure of words
  “bag of symbols” approach
- Use parallel wordlist to estimate co-occurrences of n-grams
- N-grams that have a high probability of co-occurrence in parallel meaning are interested for historical linguistics

	Bora	Muinane
down	tʃɨ̂n^{j}e, paári	báari, gíino
bee	íímɯ́ʔóexp^{h}i, téʔts^{h}ipa	nɨ̂ɨ̂bɨri, mɨ̂ɨbɨriʔɨ
sharp	ts^{h}ɯ́ʔxɨβáne	sɨ̂ɨ̂xéβano
…	…	…

Bora	Muinane	Bora	Muinane	Bora	Muinane
#k	#k^h	#i	#i	#n	#n
kɨ	k^{h}ɯ	#a	#a	#m	#m
se	ts^{h}ɨ	di	ti	mɨ	mɯ
xe	xɨ	du	to	nɨ	nɯ
ga	k^{w}a	#d	#t	us	ts^{h}ɨ
ba	pa	#s	#ts^h	#t	#t^h
#b	#p	gi	tʃi	ɨg	ɯk^w
e#	ɨ#	ni	ni	#ɸ	#p^h

Using bigram matching for cognate detection

Using bigram matching for cognate detection

- Bora ‘two’: mín^{j}éék^{h}ɯɯ́
  Muinane ‘two’: míínokɨ

Using bigram matching for cognate detection

- Bora ‘two’: mín^{j}éék^{h}ɯɯ́
  Muinane ‘two’: míínokɨ
- Extension of Levenshtein (1966): Needleman-Wunsch (1970)

	#m	mi	ii	in	no	ok	kɨ	ɨ#
#m								
mi								
in^j								
n^{j}e								
ee								
ek^h								
k^{h}ɯ								
ɯɯ								
ɯ#								

Levenshtein

	#m	mi	ii	in	no	ok	kɨ	ɨ#
#m	1							
mi		1						
in^j			?					
n^{j}e				?				
ee					?			
ek^h								
k^{h}ɯ								
ɯɯ								
ɯ#								

Levenshtein, V I. "Binary Codes Capable of Correcting Deletions, Insertions, and Reversals." Soviet Physics Doklady 10 (1966): 707–710.

Needleman-Wunsch

	#m	mi	ii	in	no	ok	kɨ	ɨ#
#m	22	3	2	2	2	2	2	2
mi	4	12	2	2	5	1	1	1
in^j	2	1	5	9	3	1	1	2
n^je	1	1	5	5	4	1	1	2
ee	3	3	3	3	6	2	2	2
ek^h	1	2	1	1	4	2	3	2
k^hɯ	2	2	2	2	2	1	23	2
ɯɯ	2	2	3	3	2	2	4	4
ɯ#	2	2	3	2	3	1	3	4

Needleman, S B, and C D Wunsch. "A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins." Journal of Molecular Biology 48, no. 3 (1970): 443-453.

	#m	mi	ii	in	no	ok	kɨ	ɨ#
#m	22	3	2	2	2	2	2	2
mi	4	12	2	2	5	1	1	1
in^j	2	1	5	9	3	1	1	2
n^je	1	1	5	5	4	1	1	2
ee	3	3	3	3	6	2	2	2
ek^h	1	2	1	1	4	2	3	2
$k^hɯ$	2	2	2	2	2	1	23	2
ɯɯ	2	2	3	3	2	2	4	4
ɯ#	2	2	3	2	3	1	3	4

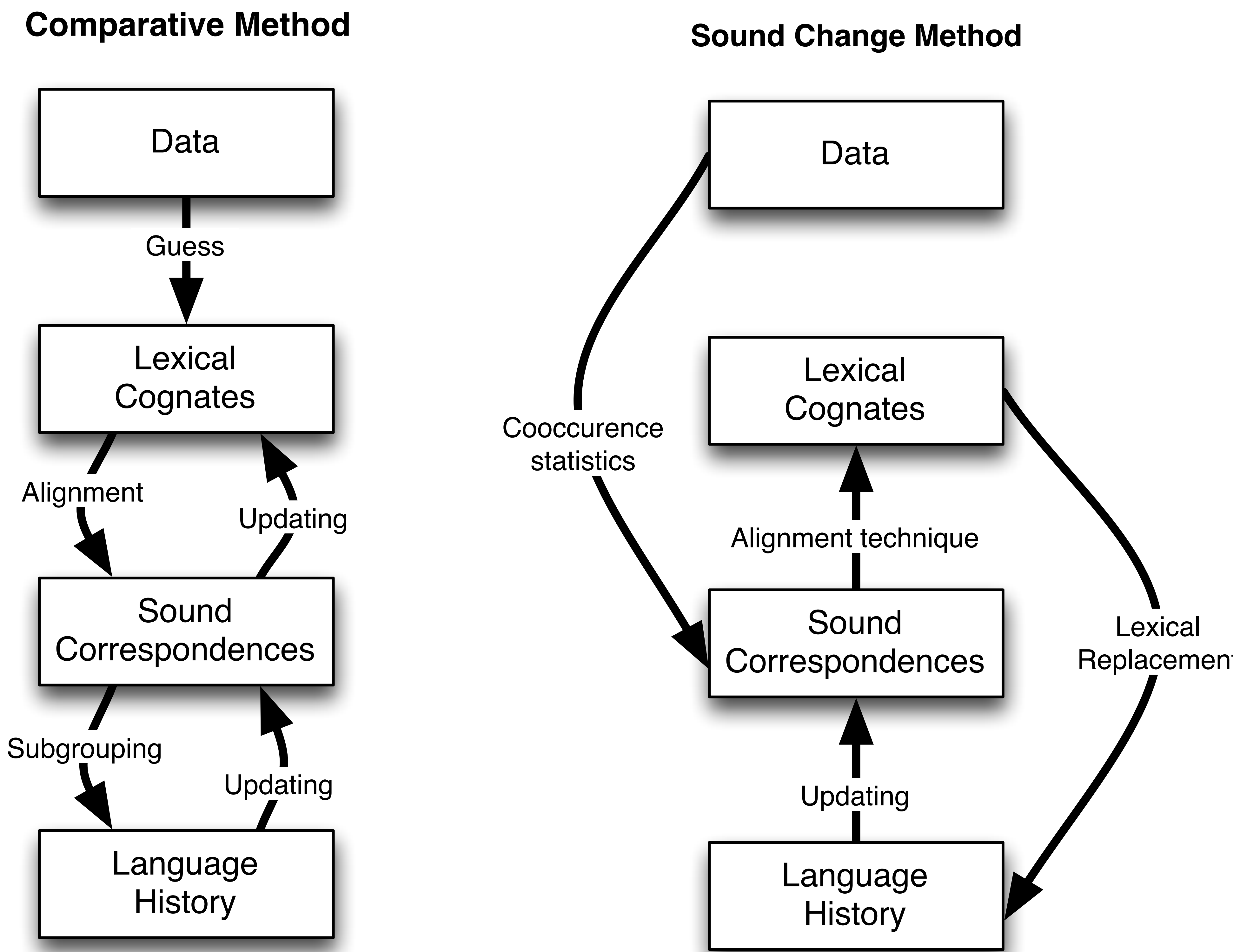
	#m	mi	ii	in	no	ok	kɨ	ɨ#
#m	22	3	2	2	2	2	2	2
mi	4	12	2	2	5	1	1	1
in^j	2	1	5	9	3	1	1	2
n^je	1	1	5	5	4	1	1	2
ee	3	3	3	3	6	2	2	2
ek^h	1	2	1	1	4	2	3	2
$k^hɯ$	2	2	2	2	2	1	23	2
ɯɯ	2	2	3	3	2	2	4	4
ɯ#	2	2	3	2	3	1	3	4

<table>
<tr><td>#m</td><td>mi</td><td>ii</td><td>in</td><td colspan="2">no</td><td>ok</td><td>kɨ</td><td colspan="2">ɨ#</td></tr>
<tr><td>#m</td><td>mi</td><td colspan="2">in^j</td><td>n^{j}e</td><td>ee</td><td>ek^h</td><td>k^{h}ɯ</td><td>ɯɯ</td><td>ɯ#</td></tr>
</table>

	#m	mi	ii	in	no	ok	kɨ	ɨ#
#m	22	3	2	2	2	2	2	2
mi	4	12	2	2	5	1	1	1
in^j	2	1	5	9	3	1	1	2
n^{j}e	1	1	5	5	4	1	1	2
ee	3	3	3	3	6	2	2	2
ek^h	1	2	1	1	4	2	3	2
k^{h}ɯ	2	2	2	2	2	1	23	2
ɯɯ	2	2	3	3	2	2	4	4
ɯ#	2	2	3	2	3	1	3	4

	Ocaina	Witoto Murui
HAND	on̥õõ	onoʤɯ
WE	xoxo	koko
HOUSE	ɸoo	ɸo
DOG	hõʔxo	hɯko
JAGUAR	hõʔxo	hɯko
FATHER	mõõ	moo
HUMMINGBIRD	ɸaʔtííʔtʲo	ɸiθido
TREE	am̥ɯ̃ɯ̃ɲ̥a	amena
STICK	am̥ɯ̃ɯ̃ɲ̥a	amena
WHO	bṍ	bu
SLEEP	ɯ́ɯ́nõ	ɯnɯ
AGOUTI	ɸɯ́ɯ́tʲo	ɸɯdo
THIS	bĩ́ĩ	bie
THIS	baʔi	bie
NAME	maam̥ɯ	mamekɯ
DAY	mooɲ̥a	aremona
BOW	tsipóxatʲa	θɯkuira
HEAR	xaaxa	kakade
DAY	moɲ̥amó	aremona

	Ocaina	Witoto Murui
GREASE	ɸahĩĩ	ɸare
YOU (PLURAL)	mõʔ	omoɯ
THIS	bɯ	bie
ARROW	oɯdʲáátʲa	dukɯraθɯ
SPEAR	oɯdʲáátʲa	dukɯraθɯ
LIP	ɸaʔóóʔko	ɸue igoɯ
GREEN	moxóóso	mokorede
I	xõ	kue
ONE	tʲa	dahe
WE	xo	kaɯ
TOOTH	aʔtiiʔtʲo	iθido
MOUTH	ɸooɯ	ɸue
BELLY	gááho	hebe
FATHER	mõõhõ	moo
YOU (PLURAL)	mõʔxo	omɯko
SWAMP	xonɯ́ɯ́βaga	kɯnere
RAT	mɯɲṍṍko	miɲɯe
PATH, TRAIL	naahõ	naɯθo
OWL	mṍṍn̥õhõ	monuiθɯ


Comparative Method
Data
Guess
Lexical Cognates
Alignment
Updating
Sound Correspondences
Subgrouping
Updating
Language History
Sound Change Method
Data
Cooccurence statistics
Lexical Cognates
Alignment technique
Sound Correspondences
Lexical Replacement
Updating
Language History

Conclusion

Conclusion

- *Regular sound change* is a very powerful notion to investigate language history

Conclusion

- *Regular sound change* is a very powerful notion to investigate language history
- **We knew that for a long time !**
  we just seem to have forgotten about it in computational approaches

Conclusion

- *Regular sound change* is a very powerful notion to investigate language history
- **We knew that for a long time !**
  we just seem to have forgotten about it in computational approaches
- *Regular symbol correspondence* can relatively easily be discovered statistically **before** cognate identification

Conclusion

- *Regular sound change* is a very powerful notion to investigate language history
- **We knew that for a long time !**
  we just seem to have forgotten about it in computational approaches
- *Regular symbol correspondence* can relatively easily be discovered statistically **before** cognate identification
- **Reversing the comparative method**
  first sound change, then cognacy judgement