諸言語の歴史的変化に対する数理的取り組み

九州大学
村脇 有吾

2015年1月19日 NL研 招待講演@九大医学部百年講堂

日本語の起源

日本語の起源、同系語は?

- 朝鮮語 [Aston, 1879][金澤, 1910][Martin, 1966]
- アルタイ語族 [Miller, 1971]
- ノストラ語族 [Starostin, 1989]
- ユーラシア語族 [Greenberg, 2000]
- オーストロネシア語族 [川本, 1980][Benedict, 1990]
- タミル語 (ドラヴィダ語族) [大野, 1980]
- レプチャ語 [安田, 1955]
- 高句麗地名 [新村, 1916]

代表的な文献
必ずしも初出ではない

日本語系統論はなぜはやらなくなったのか[1]

ー日本語系統論の現在・過去・未来ー

長 田 俊 樹
総合地球環境学研究所・国際日本文化研究センター

キーワード：日本語系統論、混淆言語、言語類型論、服部四郎、上田万年。

0．序

2001年4月から1年間、 国際日本文化研究センターにおいて、共同研究会『日本語系統論の現在』がおこなわれた。この共同研究会がなぜ開催されるようになったのか。またどのような発表がおこなわれ、どのような成果があったのか。それらについては、この報告書の「はじめに」のところでのべたので、ここではくりかえさない。ただ、ここで指摘したいのは共同研究員をあつめるのに苦労したということだ。「日本語系統論だけはかんべんしてくれ」とか、「そんなものには興味はない」とか、正直いって、かなりの研究者にことわられた。なぜ共同研究員あつめに苦労したのか。共同研究会の幹事として、おもいをめぐらしてみたところ、つぎのような結論にいたった。それは、共同研究員になりたがらない理由を大別すると、つぎのふたつにわけられるということだ。日本語系統論にはふれたくないというタイプと、日本語系統論には関心がないというタイプのふたつだ。前者は、日本語系統論に関心があるが、そのことにふれるのはまともな言語学者がやるべきではないとかんがえ、後者は日本語系統論などさいしょから眼中になく、言語学者の仕事ともおもっていない。いずれにせよ、日本語系統論がいか

[長田, 2003]

日本語系統論はなぜはやらなくなったのか[1]

ー日本語系統論の現在・過去・未来ー

長 田 俊 樹

総合地球環境学研究所・国際日本文化研究センター

キーワード：日本語系統論、混淆言語、言語類型論、服部四郎、上田万年。

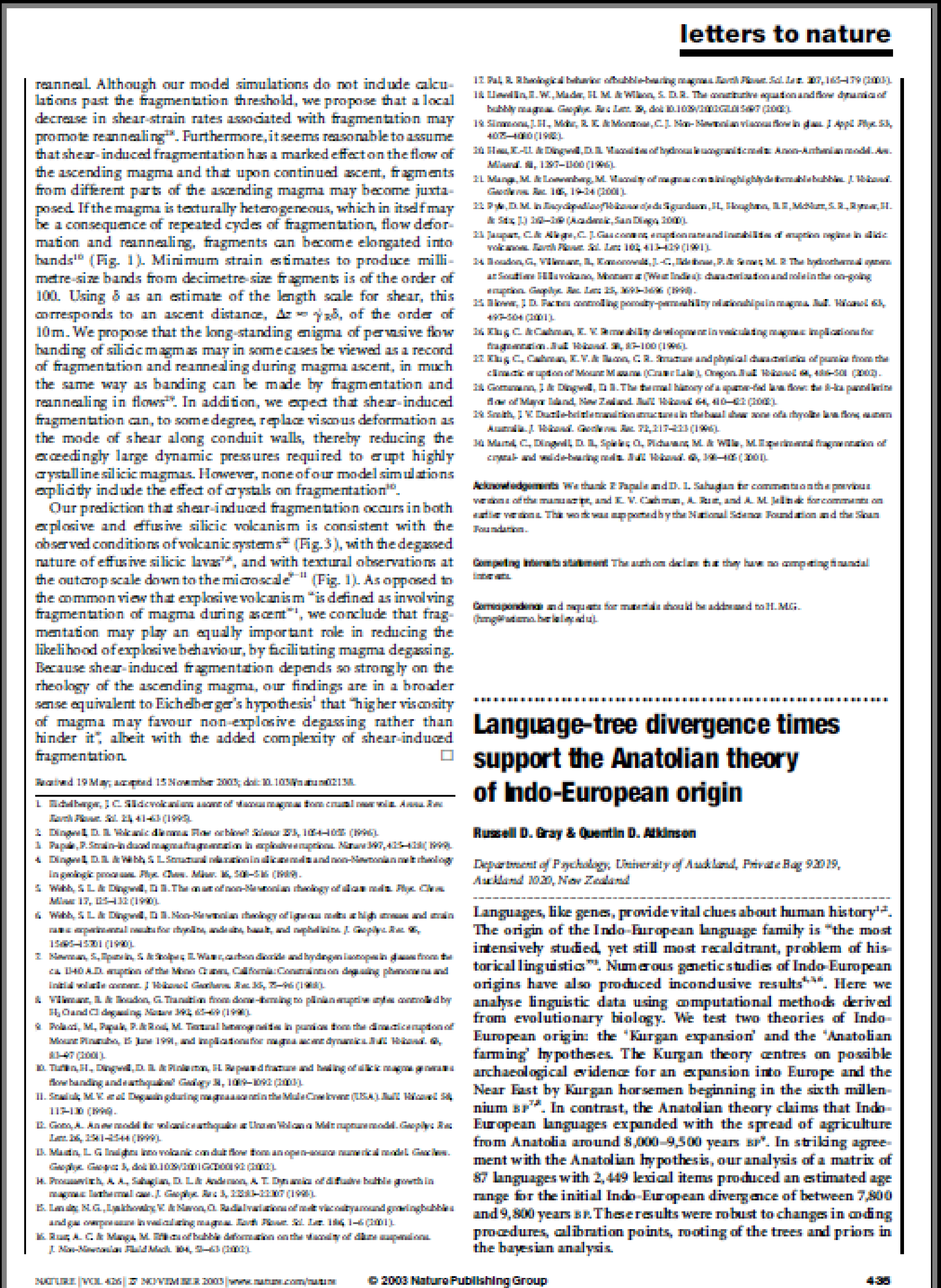
0．序

2001年4月から1年間、国際日本文化研究センターにおいて、共同研究会『日本語系統論の現在』がおこなわれた。この共同研究会がなぜ開催されるようになったのか。またどのような発表がおこなわれ、どのような成果があったのか。それらについては、この報告書の「はじめに」のところでのべたので、ここではくりかえさない。ただ、ここで指摘したいのは共同研究員をあつめるのに苦労したということだ。「日本語系統論だけはかんべんしてくれ」とか、「そんなものには興味はない」とか、正直いって、かなりの研究者にことわられた。なぜ共同研究員あつめに苦労したのか。共同研究会の幹事として、おもいをめぐらしてみたところ、つぎのような結論にいたった。それは、共同研究員になりたがらない理由を大別すると、つぎのふたつにわけられるということだ。日本語系統論にはふれたくないというタイプと、日本語系統論には関心がないというタイプのふたつだ。前者は、日本語系統論に関心があるが、そのことにふれるのはまともな言語学者がやるべきではないとかんがえ、後者は日本語系統論などさいしょから眼中になく、言語学者の仕事ともおもっていない。いずれにせよ、日本語系統論がいか

[長田, 2003]

系統論への計算的取り組み

[Gray+, Nature 2003]



letters to nature

reanneal. Although our model simulations do not include calculations past the fragmentation threshold, we propose that a local decrease in shear-strain rates associated with fragmentation may promote reannealing[28]. Furthermore, it seems reasonable to assume that shear-induced fragmentation has a marked effect on the flow of the ascending magma and that upon continued ascent, fragments from different parts of the ascending magma may become juxtaposed. If the magma is texturally heterogeneous, which in itself may be a consequence of repeated cycles of fragmentation, flow deformation and reannealing, fragments can become elongated into bands[10] (Fig. 1). Minimum strain estimates to produce millimetre-size bands from decimetre-size fragments is of the order of 100. Using δ as an estimate of the length scale for shear, this corresponds to an ascent distance, $\Delta z = \dot{\gamma}_R \delta$, of the order of 10 m. We propose that the long-standing enigma of pervasive flow banding of silicic magmas may in some cases be viewed as a record of fragmentation and reannealing during magma ascent, in much the same way as banding can be made by fragmentation and reannealing in flows[29]. In addition, we expect that shear-induced fragmentation can, to some degree, replace viscous deformation as the mode of shear along conduit walls, thereby reducing the exceedingly large dynamic pressures required to erupt highly crystalline silicic magmas. However, none of our model simulations explicitly include the effect of crystals on fragmentation[30].

Our prediction that shear-induced fragmentation occurs in both explosive and effusive silicic volcanism is consistent with the observed conditions of volcanic systems[22] (Fig. 3), with the degassed nature of effusive silicic lavas[7,8], and with textural observations at the outcrop scale down to the microscale[9–11] (Fig. 1). As opposed to the common view that explosive volcanism "is defined as involving fragmentation of magma during ascent"[1], we conclude that fragmentation may play an equally important role in reducing the likelihood of explosive behaviour, by facilitating magma degassing. Because shear-induced fragmentation depends so strongly on the rheology of the ascending magma, our findings are in a broader sense equivalent to Eichelberger's hypothesis[1] that "higher viscosity of magma may favour non-explosive degassing rather than hinder it", albeit with the added complexity of shear-induced fragmentation.

Received 19 May; accepted 15 November 2003; doi:10.1038/nature02138.

Language-tree divergence times support the Anatolian theory of Indo-European origin

Russell D. Gray & Quentin D. Atkinson

Department of Psychology, University of Auckland, Private Bag 92019, Auckland 1020, New Zealand

Languages, like genes, provide vital clues about human history[1,2]. The origin of the Indo-European language family is "the most intensively studied, yet still most recalcitrant, problem of historical linguistics"[3]. Numerous genetic studies of Indo-European origins have also produced inconclusive results[4,5,6]. Here we analyse linguistic data using computational methods derived from evolutionary biology. We test two theories of Indo-European origin: the 'Kurgan expansion' and the 'Anatolian farming' hypotheses. The Kurgan theory centres on possible archaeological evidence for an expansion into Europe and the Near East by Kurgan horsemen beginning in the sixth millennium BP[7,8]. In contrast, the Anatolian theory claims that Indo-European languages expanded with the spread of agriculture from Anatolia around 8,000–9,500 years BP[9]. In striking agreement with the Anatolian hypothesis, our analysis of a matrix of 87 languages with 2,449 lexical items produced an estimated age range for the initial Indo-European divergence of between 7,800 and 9,800 years BP. These results were robust to changes in coding procedures, calibration points, rooting of the trees and priors in the bayesian analysis.

NATURE | VOL 426 | 27 NOVEMBER 2003 | www.nature.com/nature © 2003 Nature Publishing Group 435

系統論への計算的取り組み

[Gray+, Nature 2003]

[Atkinson, Science 2011]

ability to control the intricate structure of DNA nano-architectures and create more diverse building blocks for molecular engineering.

References and Notes

1. N. C. Seeman, *J. Theor. Biol.* 99, 237 (1982).
2. N. C. Seeman, *Nature* 421, 427 (2003).
3. T. J. Fu, N. C. Seeman, *Biochemistry* 32, 3211 (1993).
4. J. H. Chen, N. C. Seeman, *Nature* 350, 631 (1991).
5. E. Winfree, F. Liu, L. A. Wenzler, N. C. Seeman, *Nature* 394, 539 (1998).
6. H. Yan, S. H. Park, G. Finkelstein, J. H. Reif, T. H. LaBean, *Science* 301, 1882 (2003).
7. W. M. Shih, J. D. Quispe, G. F. Joyce, *Nature* 427, 618 (2004).
8. F. Mathieu *et al.*, *Nano Lett.* 5, 661 (2005).
9. R. P. Goodman *et al.*, *Science* 310, 1661 (2005).
10. F. A. Aldaye, H. F. Sleiman, *J. Am. Chem. Soc.* 129, 13376 (2007).
11. Y. He *et al.*, *Nature* 452, 198 (2008).
12. C. Zhang *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* 105, 10665 (2008).
13. C. Lin, Y. Liu, H. Yan, *Biochemistry* 48, 1663 (2009).
14. P. W. K. Rothemund, *Nature* 440, 297 (2006).
15. E. S. Andersen *et al.*, *Nature* 459, 73 (2009).
16. S. M. Douglas *et al.*, *Nature* 459, 414 (2009).
17. Y. Ke *et al.*, *J. Am. Chem. Soc.* 131, 15903 (2009).
18. H. Dietz, S. M. Douglas, W. M. Shih, *Science* 325, 725 (2009).
19. Materials and methods are available as supporting material on *Science* Online.
20. P. W. K. Rothemund *et al.*, *J. Am. Chem. Soc.* 126, 16344 (2004).
21. P. J. Hagerman, *Annu. Rev. Biophys. Chem.* 17, 265 (1988).

Acknowledgments: We thank the AFM applications group at Bruker Nanosurfaces for assistance in acquiring some of the high-resolution AFM images, using Peak Force Tapping with ScanAsyst on the MultiMode 8. This research was partly supported by grants from the Office of Naval Research, Army Research Office, National Science Foundation, Department of Energy, and National Institutes of Health to H.Y. and Y.L. and from the Sloan Research Foundation to H.Y. Y.L. and H.Y. were supported by the Technology and Research Initiative Fund from Arizona State University and as part of the Center for Bio-Inspired Solar Fuel Production, an Energy Frontier Research Center funded by the U.S. Department of Energy, Office of Science, Office of Basic Energy Sciences under award DE-SC0001016.

Supporting Online Material
www.sciencemag.org/cgi/content/full/332/6027/342/DC1
Materials and Methods
SOM Text
Figs. S1 to S39
Tables S1 to S19
Reference 22

18 January 2011; accepted 4 March 2011
10.1126/science.1202998

Phonemic Diversity Supports a Serial Founder Effect Model of Language Expansion from Africa

Quentin D. Atkinson[1,2]*

Human genetic and phenotypic diversity declines with distance from Africa, as predicted by a serial founder effect in which successive population bottlenecks during range expansion progressively reduce diversity, underpinning support for an African origin of modern humans. Recent work suggests that a similar founder effect may operate on human culture and language. Here I show that the number of phonemes used in a global sample of 504 languages is also clinal and fits a serial founder–effect model of expansion from an inferred origin in Africa. This result, which is not explained by more recent demographic history, local language diversity, or statistical non-independence within language families, points to parallel mechanisms shaping genetic and linguistic diversity and supports an African origin of modern human languages.

The number of phonemes—perceptually distinct units of sound that differentiate words—in a language is positively correlated with the size of its speaker population (*1*) in such a way that small populations have fewer phonemes. Languages continually gain and lose phonemes because of stochastic processes (*2*, *3*). If phoneme distinctions are more likely to be lost in small founder populations, then a succession of founder events during range expansion should progressively reduce phonemic diversity with increasing distance from the point of origin, paralleling the serial founder effect observed in population genetics (*4*–*9*). A founder effect has already been used to explain patterns of variation in other cultural replicators, including human material culture (*10*–*13*) and birdsong (*14*). A range of possible mechanisms (*15*) predicts similar dynamics governing the evolution of phonemes (*11*, *16*) and language generally (*17*–*20*). This raises the possibility that the serial founder-effect model used to trace our genetic origins to a recent expansion from Africa (*4*–*9*) could also be applied to global phonemic diversity to investigate the origin and expansion of modern human languages. Here I examine geographic variation in phoneme inventory size using data on vowel, consonant, and tone inventories taken from 504 languages in the World Atlas of Language Structures (WALS) (*21*), together with information on language location, taxonomic affiliation, and speaker demography (Fig. 1 and table S1) (*15*).

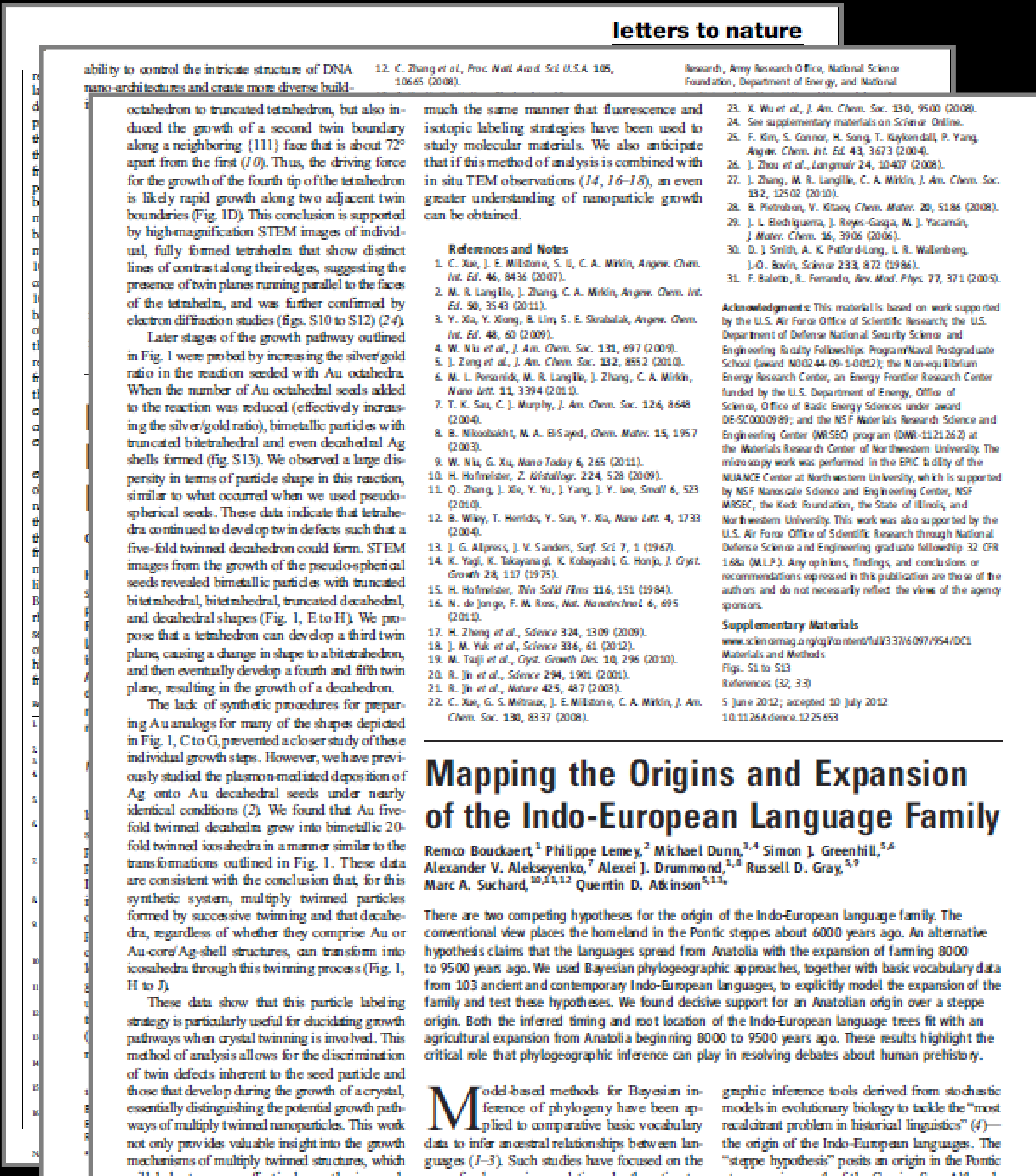
Consistent with previous work (*1*), speaker population size is a significant predictor of phonemic diversity (Pearson's correlation $r = 0.385$, $df = 503$, $P < 0.001$), with smaller population size predicting smaller overall phoneme inventories (fig. S1A). The same relationship holds for vowel ($r = 0.378$, $df = 503$, $P < 0.001$) and tone ($r = 0.230$, $df = 503$, $P < 0.001$) inventories separately, with a weaker, though still significant, effect of population size on consonant diversity ($r = 0.131$, $df = 503$, $P = 0.003$). To account for any non-independence within language families, the analysis was repeated, first using mean values at the language family level (table S2) and then using a hierarchical linear regression framework to model nested dependencies in variation at the family, subfamily, and genus levels (*15*). These analyses confirm that, consistent with a founder effect model, smaller population size predicts reduced phoneme inventory size both between families (family-level analysis $r = 0.468$, $df = 49$, $P < 0.001$; fig. S1B) and within families, controlling for taxonomic affiliation [hierarchical linear model: fixed-effect coefficient (β) = 0.0338 to 0.0985 [95% highest posterior density (HPD)], $P = 0.009$].

Figure 1B shows clear regional differences in phonemic diversity, with the largest phoneme inventories in Africa and the smallest in South America and Oceania. A series of linear regressions was used to predict phoneme inventory size from the log of speaker population size and distance from 2560 potential origin locations around the world (*15*). Incorporating modern speaker population size into the model controls for geographic patterning in population size and means that the analysis is conservative about the amount of variation attributed to ancient demography. Model fit was evaluated with the Bayesian information criterion (BIC) (*22*). Following previous work (*5*, *6*), the set of origin locations within four BIC units of the best-fit location was taken to be the most likely area of origin under a serial founder–effect model.

The origin locations producing the strongest decline in phonemic diversity and best-fit model lie across central and southern Africa (Fig. 2A). This region could represent either a single origin for modern languages or the main origin under a polygenesis scenario. The best-fit model incorporating population size and distance from the origin explains 31% of the variance in phoneme inventory size [correlation coefficient (R) = 0.558, $F_{2,501} = 113.463$, $P < 0.001$] (Fig. 3). Both population size ($r_{population} = 0.146$, $P = 0.002$)

[1]Department of Psychology, University of Auckland, Private Bag 92019, Auckland, New Zealand. [2]Institute of Cognitive and Evolutionary Anthropology, University of Oxford, 64 Banbury Road, Oxford OX2 6PN, UK.

*E-mail: q.atkinson@auckland.ac.nz



letters to nature

ability to control the intricate structure of DNA nano-architectures and create more diverse build-

12. C. Zhang et al., Proc. Natl. Acad. Sci. U.S.A. 105, 10665 (2008).

Research, Army Research Office, National Science Foundation, Department of Energy, and National

octahedron to truncated tetrahedron, but also induced the growth of a second twin boundary along a neighboring {111} face that is about 72° apart from the first (10). Thus, the driving force for the growth of the fourth tip of the tetrahedron is likely rapid growth along two adjacent twin boundaries (Fig. 1D). This conclusion is supported by high-magnification STEM images of individual, fully formed tetrahedra that show distinct lines of contrast along their edges, suggesting the presence of twin planes running parallel to the faces of the tetrahedra, and was further confirmed by electron diffraction studies (figs. S10 to S12) (24).

Later stages of the growth pathway outlined in Fig. 1 were probed by increasing the silver/gold ratio in the reaction seeded with Au octahedra. When the number of Au octahedral seeds added to the reaction was reduced (effectively increasing the silver/gold ratio), bimetallic particles with truncated bitetrahedral and even decahedral Ag shells formed (fig. S13). We observed a large dispersity in terms of particle shape in this reaction, similar to what occurred when we used pseudo-spherical seeds. These data indicate that tetrahedra continued to develop twin defects such that a five-fold twinned decahedron could form. STEM images from the growth of the pseudo-spherical seeds revealed bimetallic particles with truncated bitetrahedral, bitetrahedral, truncated decahedral, and decahedral shapes (Fig. 1, E to H). We propose that a tetrahedron can develop a third twin plane, causing a change in shape to a bitetrahedron, and then eventually develop a fourth and fifth twin plane, resulting in the growth of a decahedron.

The lack of synthetic procedures for preparing Au analogs for many of the shapes depicted in Fig. 1, C to G, prevented a closer study of these individual growth steps. However, we have previously studied the plasmon-mediated deposition of Ag onto Au decahedral seeds under nearly identical conditions (2). We found that Au five-fold twinned decahedra grew into bimetallic 20-fold twinned icosahedra in a manner similar to the transformations outlined in Fig. 1. These data are consistent with the conclusion that, for this synthetic system, multiply twinned particles formed by successive twinning and that decahedra, regardless of whether they comprise Au or Au-core/Ag-shell structures, can transform into icosahedra through this twinning process (Fig. 1, H to J).

These data show that this particle labeling strategy is particularly useful for elucidating growth pathways when crystal twinning is involved. This method of analysis allows for the discrimination of twin defects inherent to the seed particle and those that develop during the growth of a crystal, essentially distinguishing the potential growth pathways of multiply twinned nanoparticles. This work not only provides valuable insight into the growth mechanisms of multiply twinned structures, which

much the same manner that fluorescence and isotopic labeling strategies have been used to study molecular materials. We also anticipate that if this method of analysis is combined with in situ TEM observations (14, 16–18), an even greater understanding of nanoparticle growth can be obtained.

References and Notes

1. C. Xue, J. E. Millstone, S. Li, C. A. Mirkin, Angew. Chem. Int. Ed. 46, 8436 (2007).
2. M. R. Langille, J. Zhang, C. A. Mirkin, Angew. Chem. Int. Ed. 50, 3543 (2011).
3. Y. Xia, Y. Xiong, B. Lim, S. E. Skrabalak, Angew. Chem. Int. Ed. 48, 60 (2009).
4. W. Niu et al., J. Am. Chem. Soc. 131, 697 (2009).
5. J. Zeng et al., J. Am. Chem. Soc. 132, 8552 (2010).
6. M. L. Personick, M. R. Langille, J. Zhang, C. A. Mirkin, Nano Lett. 11, 3394 (2011).
7. T. K. Sau, C. J. Murphy, J. Am. Chem. Soc. 126, 8648 (2004).
8. B. Nikoobakht, M. A. El-Sayed, Chem. Mater. 15, 1957 (2003).
9. W. Niu, G. Xu, Nano Today 6, 265 (2011).
10. H. Hofmeister, Z. Kristallogr. 224, 528 (2009).
11. Q. Zhang, J. Xie, Y. Yu, J. Yang, J. Y. Lee, Small 6, 523 (2010).
12. B. Wiley, T. Herricks, Y. Sun, Y. Xia, Nano Lett. 4, 1733 (2004).
13. J. G. Allpress, J. V. Sanders, Surf. Sci. 7, 1 (1967).
14. K. Yagi, K. Takayanagi, K. Kobayashi, G. Honjo, J. Cryst. Growth 28, 117 (1975).
15. H. Hofmeister, Thin Solid Films 116, 151 (1984).
16. N. de Jonge, F. M. Ross, Nat. Nanotechnol. 6, 695 (2011).
17. H. Zheng et al., Science 324, 1309 (2009).
18. J. M. Yuk et al., Science 336, 61 (2012).
19. M. Tsuji et al., Cryst. Growth Des. 10, 296 (2010).
20. R. Jin et al., Science 294, 1901 (2001).
21. R. Jin et al., Nature 425, 487 (2003).
22. C. Xue, G. S. Métraux, J. E. Millstone, C. A. Mirkin, J. Am. Chem. Soc. 130, 8337 (2008).
23. X. Wu et al., J. Am. Chem. Soc. 130, 9500 (2008).
24. See supplementary materials on Science Online.
25. F. Kim, S. Connor, H. Song, T. Kuykendall, P. Yang, Angew. Chem. Int. Ed. 43, 3673 (2004).
26. J. Zhou et al., Langmuir 24, 10407 (2008).
27. J. Zhang, M. R. Langille, C. A. Mirkin, J. Am. Chem. Soc. 132, 12502 (2010).
28. B. Pietrobon, V. Kitaev, Chem. Mater. 20, 5186 (2008).
29. J. L. Elechiguerra, J. Reyes-Gasga, M. J. Yacamán, J. Mater. Chem. 16, 3906 (2006).
30. D. J. Smith, A. K. Petford-Long, L. R. Wallenberg, J.-O. Bovin, Science 233, 872 (1986).
31. F. Baletto, R. Ferrando, Rev. Mod. Phys. 77, 371 (2005).

Acknowledgments: This material is based on work supported by the U.S. Air Force Office of Scientific Research; the U.S. Department of Defense National Security Science and Engineering Faculty Fellowships Program/Naval Postgraduate School (award N00244-09-1-0012); the Non-equilibrium Energy Research Center, an Energy Frontier Research Center funded by the U.S. Department of Energy, Office of Science, Office of Basic Energy Sciences under award DE-SC0000989; and the NSF Materials Research Science and Engineering Center (MRSEC) program (DMR-1121262) at the Materials Research Center of Northwestern University. The microscopy work was performed in the EPIC facility of the NUANCE Center at Northwestern University, which is supported by NSF Nanoscale Science and Engineering Center, NSF MRSEC, the Keck Foundation, the State of Illinois, and Northwestern University. This work was also supported by the U.S. Air Force Office of Scientific Research through National Defense Science and Engineering graduate fellowship 32 CFR 168a (M.L.P.). Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the agency sponsors.

Supplementary Materials
www.sciencemag.org/cgi/content/full/337/6097/954/DC1
Materials and Methods
Figs. S1 to S13
References (32, 33)

5 June 2012; accepted 10 July 2012
10.1126/science.1225653

Mapping the Origins and Expansion of the Indo-European Language Family

Remco Bouckaert,[1] Philippe Lemey,[2] Michael Dunn,[3,4] Simon J. Greenhill,[5,6] Alexander V. Alekseyenko,[7] Alexei J. Drummond,[1,8] Russell D. Gray,[5,9] Marc A. Suchard,[10,11,12] Quentin D. Atkinson[5,13]*

There are two competing hypotheses for the origin of the Indo-European language family. The conventional view places the homeland in the Pontic steppes about 6000 years ago. An alternative hypothesis claims that the languages spread from Anatolia with the expansion of farming 8000 to 9500 years ago. We used Bayesian phylogeographic approaches, together with basic vocabulary data from 103 ancient and contemporary Indo-European languages, to explicitly model the expansion of the family and test these hypotheses. We found decisive support for an Anatolian origin over a steppe origin. Both the inferred timing and root location of the Indo-European language trees fit with an agricultural expansion from Anatolia beginning 8000 to 9500 years ago. These results highlight the critical role that phylogeographic inference can play in resolving debates about human prehistory.

Model-based methods for Bayesian inference of phylogeny have been applied to comparative basic vocabulary data to infer ancestral relationships between languages (1–3). Such studies have focused on the

graphic inference tools derived from stochastic models in evolutionary biology to tackle the "most recalcitrant problem in historical linguistics" (4)—the origin of the Indo-European languages. The "steppe hypothesis" posits an origin in the Pontic

系統論への計算的取り組み

[Gray+, Nature 2003]

[Atkinson, Science 2011]

[Bouckaert+, Science 2012]

ability to control the intricate structure of DNA nano-architectures and create more diverse build-

12. C. Zhang et al., Proc. Natl. Acad. Sci. U.S.A. 105, 10665 (2008).

Research, Army Research Office, National Science Foundation, Department of Energy, and National

octahedron to truncated tetrahedron, but also induced the growth of a second twin boundary along a neighboring {111} face that is about 72° apart from the first (10). Thus, the driving force for the growth of the fourth tip of the tetrahedron is likely rapid growth along two adjacent twin boundaries (Fig. 1D). This conclusion is supported by high-magnification STEM images of individual, fully formed tetrahedra that show distinct lines of contrast along their edges, suggesting the presence of twin planes running parallel to the faces of the tetrahedra, and was further confirmed by electron diffraction studies (figs. S10 to S12) (24).

Later stages of the growth pathway outlined in Fig. 1 were probed by increasing the silver/gold ratio in the reaction seeded with Au octahedra. When the number of Au octahedral seeds added to the reaction was reduced (effectively increasing the silver/gold ratio), bimetallic particles with truncated bitetrahedral and even decahedral Ag shells formed (fig. S13). We observed a large dispersity in terms of particle shape in this reaction, similar to what occurred when we used pseudo-spherical seeds. These data indicate that tetrahedra continued to develop twin defects such that a five-fold twinned decahedron could form. STEM images from the growth of the pseudo-spherical seeds revealed bimetallic particles with truncated bitetrahedral, bitetrahedral, truncated decahedral, and decahedral shapes (Fig. 1, E to H). We propose that a tetrahedron can develop a third twin plane, causing a change in shape to a bitetrahedron, and then eventually develop a fourth and fifth twin plane, resulting in the growth of a decahedron.

The lack of synthetic procedures for preparing Au analogs for many of the shapes depicted in Fig. 1, C to G, prevented a closer study of these individual growth steps. However, we have previously studied the plasmon-mediated deposition of Ag onto Au decahedral seeds under nearly identical conditions (2). We found that Au five-fold twinned decahedra grew into bimetallic 20-fold twinned icosahedra in a manner similar to the transformations outlined in Fig. 1. These data are consistent with the conclusion that, for this synthetic system, multiply twinned particles formed by successive twinning and that decahedra, regardless of whether they comprise Au or Au-core/Ag-shell structures, can transform into icosahedra through this twinning process (Fig. 1, H to J).

These data show that this particle labeling strategy is particularly useful for elucidating growth pathways when crystal twinning is involved. This method of analysis allows for the discrimination of twin defects inherent to the seed particle and those that develop during the growth of a crystal, essentially distinguishing the potential growth pathways of multiply twinned nanoparticles. This work not only provides valuable insight into the growth mechanisms of multiply twinned structures, which

much the same manner that fluorescence and isotopic labeling strategies have been used to study molecular materials. We also anticipate that if this method of analysis is combined with in situ TEM observations (14, 16–18), an even greater understanding of nanoparticle growth can be obtained.

References and Notes
1. C. Xue, J. E. Millstone, S. Li, C. A. Mirkin, Angew. Chem. Int. Ed. 46, 8436 (2007).
2. M. R. Langille, J. Zhang, C. A. Mirkin, Angew. Chem. Int. Ed. 50, 3543 (2011).
3. Y. Xia, Y. Xiong, B. Lim, S. E. Skrabalak, Angew. Chem. Int. Ed. 48, 60 (2009).
4. W. Niu et al., J. Am. Chem. Soc. 131, 697 (2009).
5. J. Zeng et al., J. Am. Chem. Soc. 132, 8552 (2010).
6. M. L. Personick, M. R. Langille, J. Zhang, C. A. Mirkin, Nano Lett. 11, 3394 (2011).
7. T. K. Sau, C. J. Murphy, J. Am. Chem. Soc. 126, 8648 (2004).
8. B. Nikoobakht, M. A. El-Sayed, Chem. Mater. 15, 1957 (2003).
9. W. Niu, G. Xu, Nano Today 6, 265 (2011).
10. H. Hofmeister, Z. Kristallogr. 224, 528 (2009).
11. Q. Zhang, J. Xie, Y. Yu, J. Yang, J. Y. Lee, Small 6, 523 (2010).
12. B. Wiley, T. Herricks, Y. Sun, Y. Xia, Nano Lett. 4, 1733 (2004).
13. J. G. Allpress, J. V. Sanders, Surf. Sci. 7, 1 (1967).
14. K. Yagi, K. Takayanagi, K. Kobayashi, G. Honjo, J. Cryst. Growth 28, 117 (1975).
15. H. Hofmeister, Thin Solid Films 116, 151 (1984).
16. N. de Jonge, F. M. Ross, Nat. Nanotechnol. 6, 695 (2011).
17. H. Zheng et al., Science 324, 1309 (2009).
18. J. M. Yuk et al., Science 336, 61 (2012).
19. M. Tsuji et al., Cryst. Growth Des. 10, 296 (2010).
20. R. Jin et al., Science 294, 1901 (2001).
21. R. Jin et al., Nature 425, 487 (2003).
22. C. Xue, G. S. Métraux, J. E. Millstone, C. A. Mirkin, J. Am. Chem. Soc. 130, 8337 (2008).
23. X. Wu et al., J. Am. Chem. Soc. 130, 9500 (2008).
24. See supplementary materials on Science Online.
25. F. Kim, S. Connor, H. Song, T. Kuykendall, P. Yang, Angew. Chem. Int. Ed. 43, 3673 (2004).
26. J. Zhou et al., Langmuir 24, 10407 (2008).
27. J. Zhang, M. R. Langille, C. A. Mirkin, J. Am. Chem. Soc. 132, 12502 (2010).
28. B. Pietrobon, V. Kitaev, Chem. Mater. 20, 5186 (2008).
29. J. L. Elechiguerra, J. Reyes-Gasga, M. J. Yacamán, J. Mater. Chem. 16, 3906 (2006).
30. D. J. Smith, A. K. Petford-Long, L. R. Wallenberg, J.-O. Bovin, Science 233, 872 (1986).
31. F. Baletto, R. Ferrando, Rev. Mod. Phys. 77, 371 (2005).

Acknowledgments: This material is based on work supported by the U.S. Air Force Office of Scientific Research; the U.S. Department of Defense National Security Science and Engineering Faculty Fellowships Program/Naval Postgraduate School (award N00244-09-1-0012); the Non-equilibrium Energy Research Center, an Energy Frontier Research Center funded by the U.S. Department of Energy, Office of Science, Office of Basic Energy Sciences under award DE-SC0000989; and the NSF Materials Research Science and Engineering Center (MRSEC) program (DMR-1121262) at the Materials Research Center of Northwestern University. The microscopy work was performed in the EPIC facility of the NUANCE Center at Northwestern University, which is supported by NSF Nanoscale Science and Engineering Center, NSF MRSEC, the Keck Foundation, the State of Illinois, and Northwestern University. This work was also supported by the U.S. Air Force Office of Scientific Research through National Defense Science and Engineering graduate fellowship 32 CFR 168a (M.L.P.). Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the agency sponsors.

Supplementary Materials
www.sciencemag.org/cgi/content/full/337/6097/954/DC1
Materials and Methods
Figs. S1 to S13
References (32, 33)

5 June 2012; accepted 10 July 2012
10.1126/science.1225653

Mapping the Origins and Expansion of the Indo-European Language Family

Remco Bouckaert,[1] Philippe Lemey,[2] Michael Dunn,[3,4] Simon J. Greenhill,[5,6] Alexander V. Alekseyenko,[7] Alexei J. Drummond,[1,8] Russell D. Gray,[5,9] Marc A. Suchard,[10,11,12] Quentin D. Atkinson[5,13]

There are two competing hypotheses for the origin of the Indo-European language family. The conventional view places the homeland in the Pontic steppes about 6000 years ago. An alternative hypothesis claims that the languages spread from Anatolia with the expansion of farming 8000 to 9500 years ago. We used Bayesian phylogeographic approaches, together with basic vocabulary data from 103 ancient and contemporary Indo-European languages, to explicitly model the expansion of the family and test these hypotheses. We found decisive support for an Anatolian origin over a steppe origin. Both the inferred timing and root location of the Indo-European language trees fit with an agricultural expansion from Anatolia beginning 8000 to 9500 years ago. These results highlight the critical role that phylogeographic inference can play in resolving debates about human prehistory.

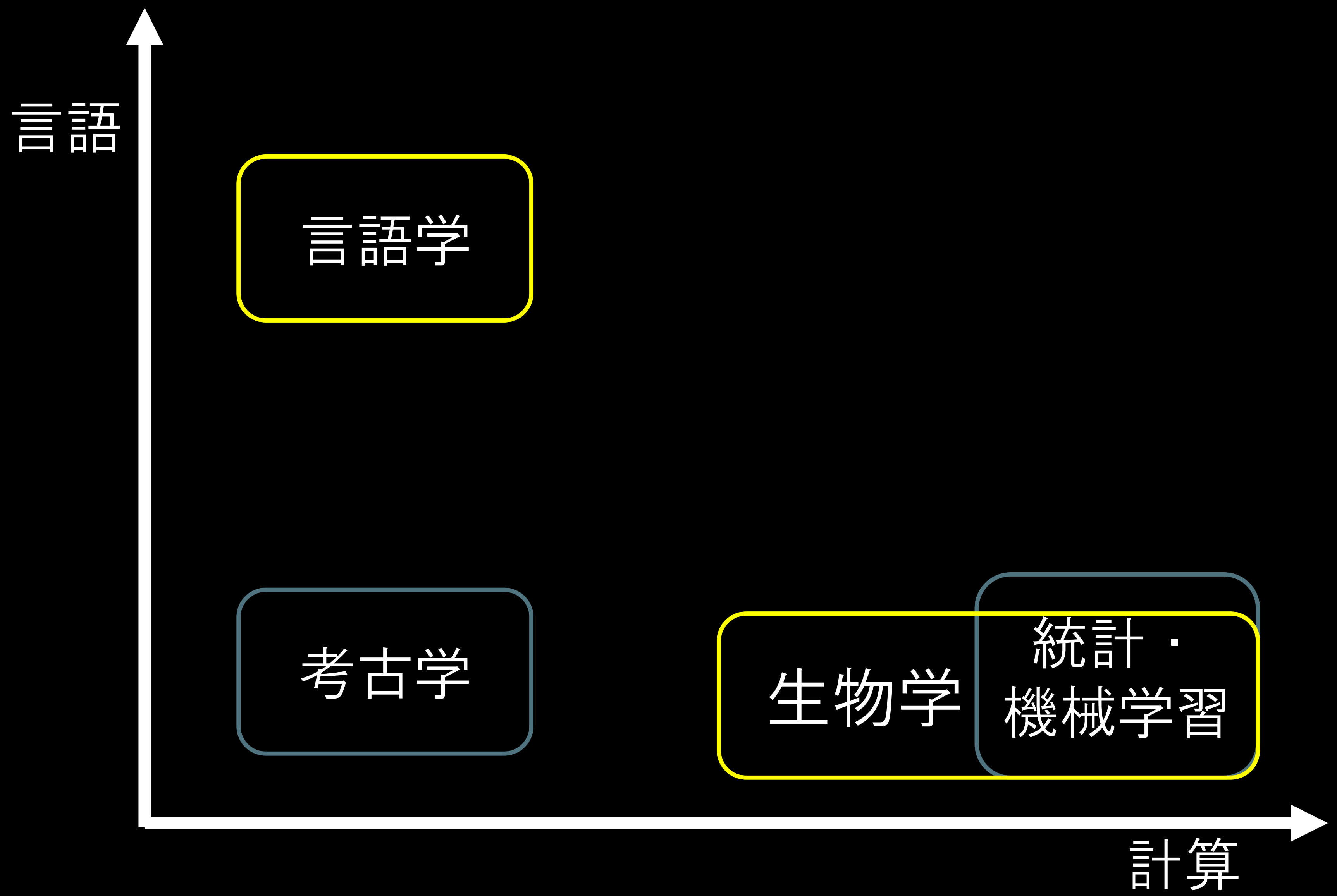
Model-based methods for Bayesian inference of phylogeny have been applied to comparative basic vocabulary data to infer ancestral relationships between languages (1–3). Such studies have focused on the

graphic inference tools derived from stochastic models in evolutionary biology to tackle the "most recalcitrant problem in historical linguistics" (4)—the origin of the Indo-European languages. The "steppe hypothesis" posits an origin in the Pontic

系統論への計算的取り組み

[Gray+, Nature 2003]
[Atkinson, Science 2011]
[Bouckaert+, Science 2012]

Taken from http://www.fos.auckland.ac.nz/~quentinatkinson/Quentin_Atkinsons_Website/Home.html


言語
言語学
考古学
生物学
統計・
機械学習
計算

言語

言語学

考古学

計算

Photo by Tom & Katrien

言語

考古学

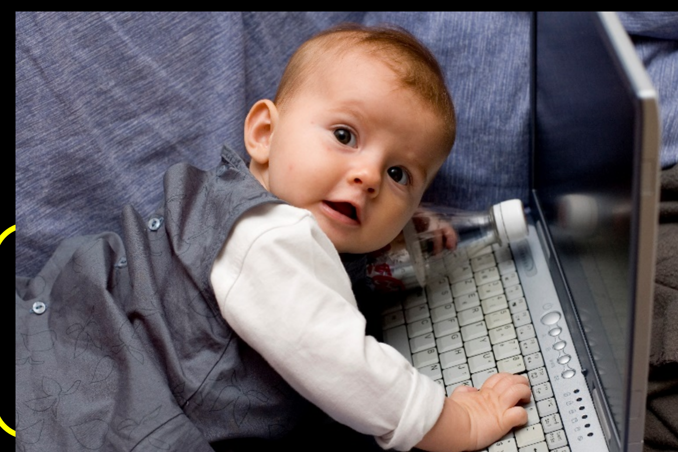
計算

Photo by stuart anthony
Licensed under CC BY-NC 2.0

Photo by Tom & Katrien
Licensed under CC BY-SA

言語

計算
言語学

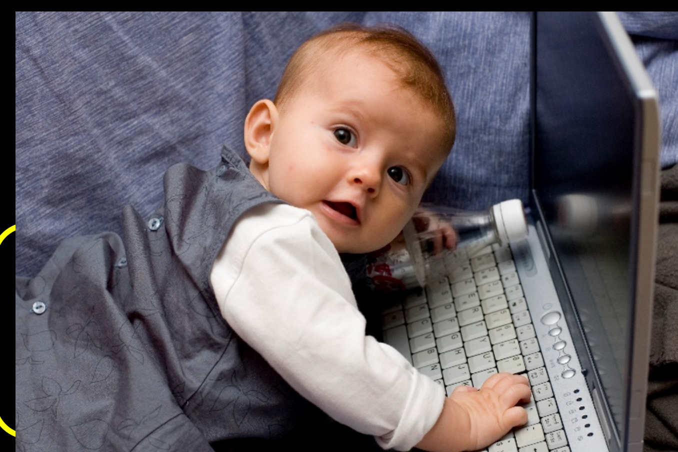


考古学

計算

言語

考古学



計算

Photo by stuart anthony

Photo by Tom & Katrien

今日のおはなし

1. 音法則
2. 基礎語彙
3. 類型論

今日のおはなし

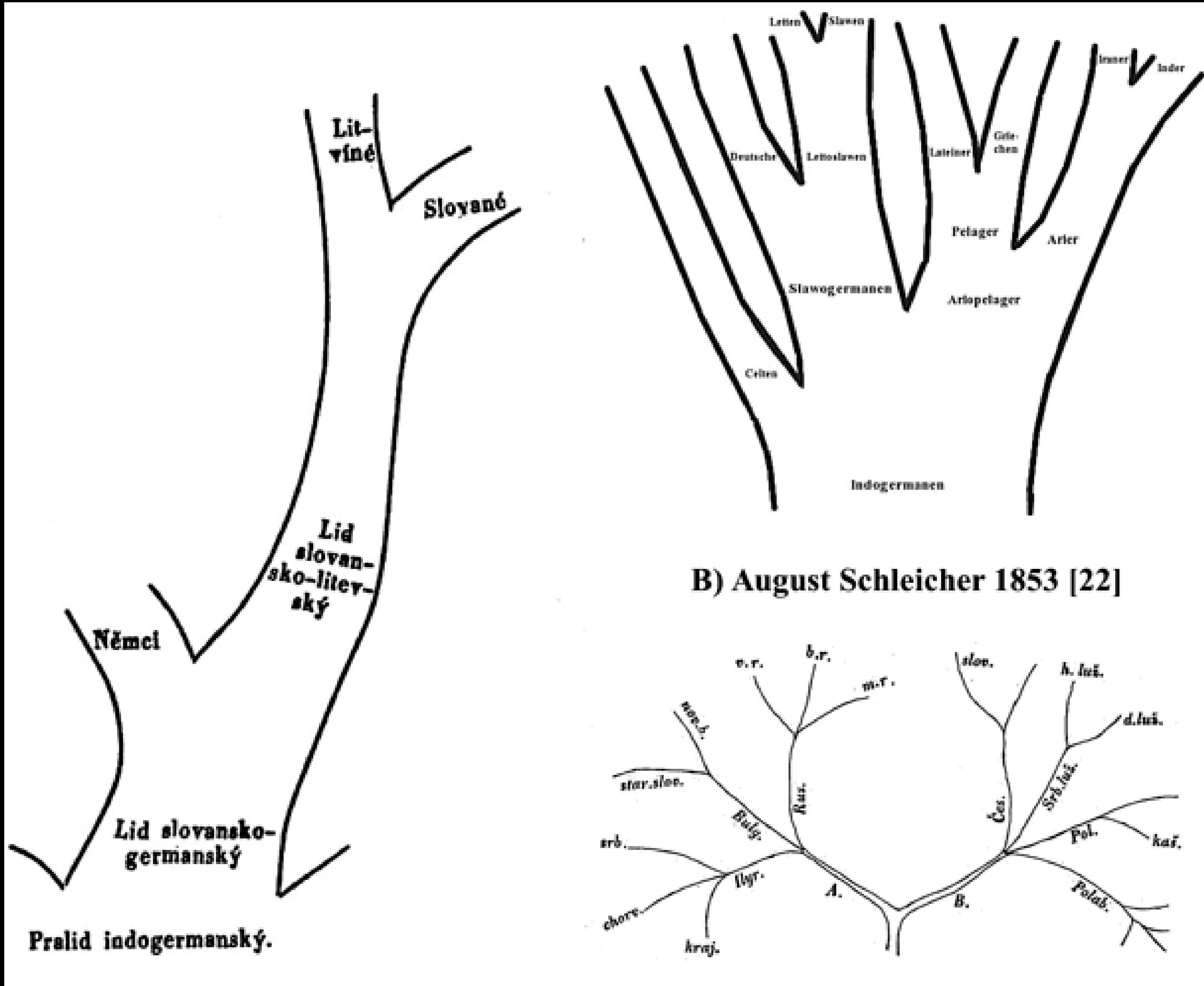
1. 音法則　比較言語学の伝統的手法
2. 基礎語彙　言語年代学 (1950年代~)
   Bayes統計の応用 (2000年代~)
3. 類型論　系統論への応用は少ない
   日本語系統論の最後の希望

今日のおはなし

1. 音法則　比較言語学の伝統的手法
2. 基礎語彙
3. 類型論

音法則は安心と信頼の枯れた技術

- サンスクリットとラテン語、ギリシア語の共通起源説と比較言語学のはじまり [Jones, 1786]
- インド･ヨーロッパ語族の系統樹 [Schleicher, 1853]
- Grimmの法則 [Grimm+, 1822] とVernerの法則 [Verner, 1875]
- 青年文法学派 (19世紀後半)
  – 音法則の無例外性
    (die Ausnahmslosigkeit der Lautgesetze)
- 喉音理論 [Saussure, 1879] のヒッタイト語による実証 [Kuryłowicz, 1927]



A) August Schleicher, 1853 [21]

B) August Schleicher 1853 [22]

C) František Ladislav Čelakovský 1853 [24]

Source: [List+, 2013]

音対応と音変化

伊

西

ある言語ペアが系統関係にある

音対応と音変化

伊	latte /latte/	otto /ɔtto/	stretto /stretto/
西	leche /letʃe/	ocho /otʃo/	estrecho /estɾetʃo/

ある言語ペアが系統関係にある

⇔ 語彙が偶然や借用で説明できないほど類似

cf. 名前とname, 骨とbone

音対応と音変化

伊	latte /latte/	otto /ɔtto/	stretto /stretto/
西	leche /letʃe/	ocho /otʃo/	estrecho /estretʃo/

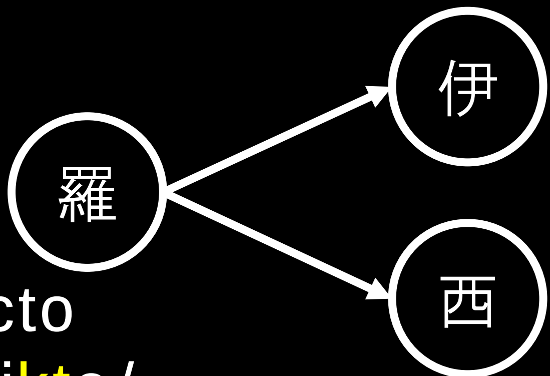
ある言語ペアが系統関係にある

⇔ 語彙が偶然や借用で説明できないほど類似

cf. 名前とname, 骨とbone

⇔ 規則的な音対応が存在

音対応と音変化



latte　otto　stretto
/latte/ /ɔtto/ /stretto/

leche　ocho　estrecho
/letʃe/ /otʃo/ /estretʃo/

lacte　octo　stricto
/lakte/ /okto/ /strikto/

ある言語ペアが系統関係にある

⇔ 語彙が偶然や借用で説明できないほど類似

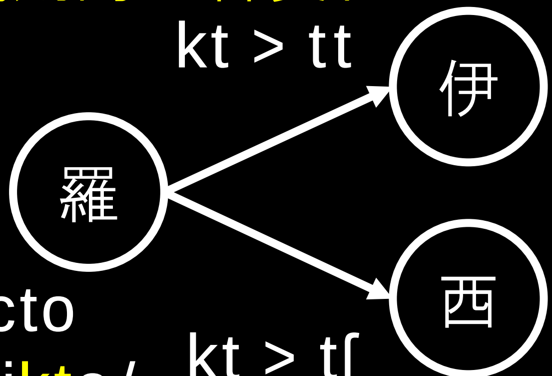
cf. 名前とname, 骨とbone

⇔ 規則的な音対応が存在

音対応と音変化

規則的な音変化

kt > tt



latte　otto　stretto
/latte/ /ɔtto/ /stretto/

leche　ocho　estrecho
/letʃe/ /otʃo/ /estretʃo/

lacte　octo　stricto
/lakte/ /okto/ /strikto/

kt > tʃ

ある言語ペアが系統関係にある

⇔ 語彙が偶然や借用で説明できないほど類似

cf. 名前とname, 骨とbone

⇔ 規則的な音対応が存在

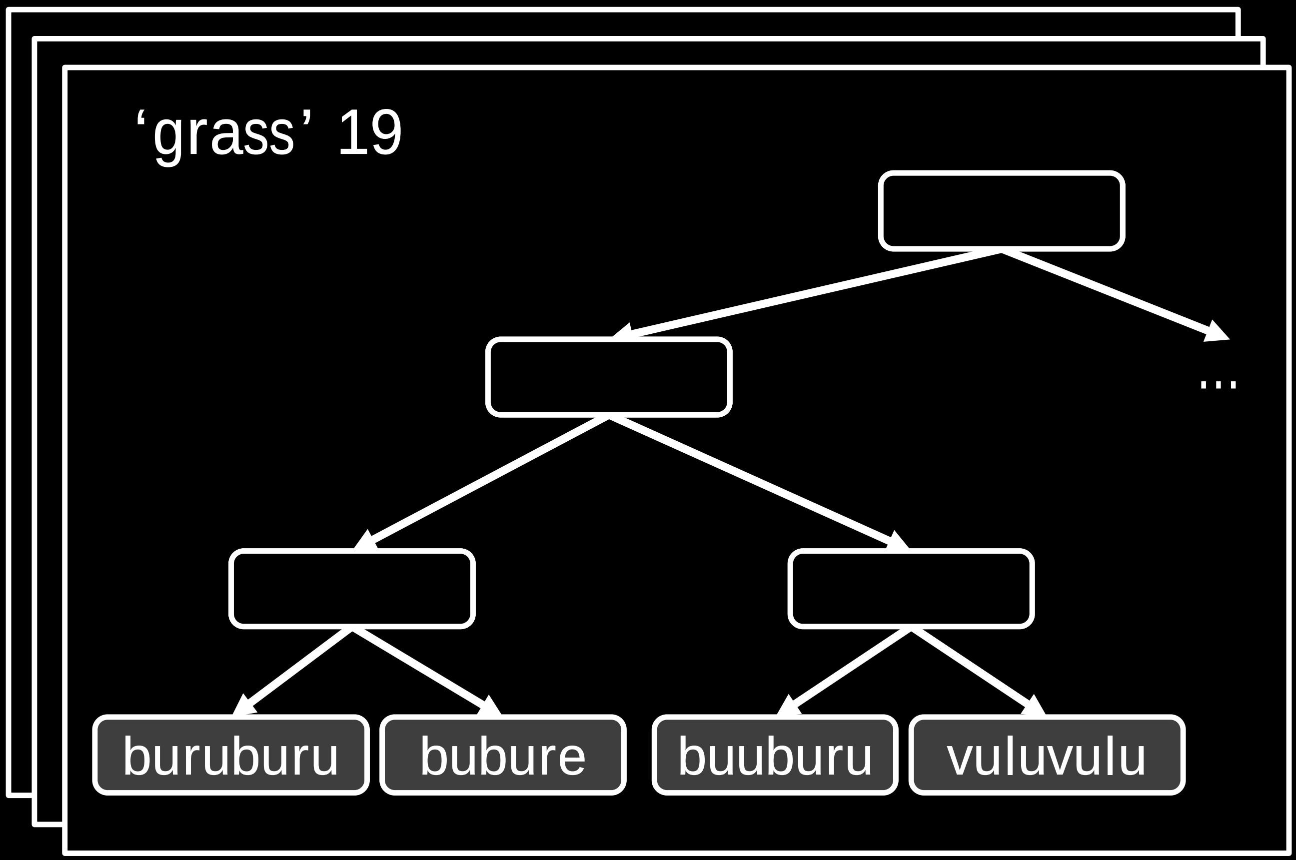
祖語の再構はアート

1. 通言語的傾向
   - 起こりやすい変化
     - 不変化: *X* > *X*
     - 弱化: p > ϕ > h > *Φ* (zero), s > h
     - 有声音間の有声化: p > b / vocalic _ vocalic
   - 起こりにくい変化
     - k > a, a > k
     - 弱化の反対: h > p
2. 体系の自然さ
   - 5母音体系なら/a/の出現頻度は30~40%が普通で、あまりに少ないと不自然
3. 内的再構
   - 交替現象: k ~ g, s ~ z, t ~ d ⇒ *p ~ b

祖語の再構手続き

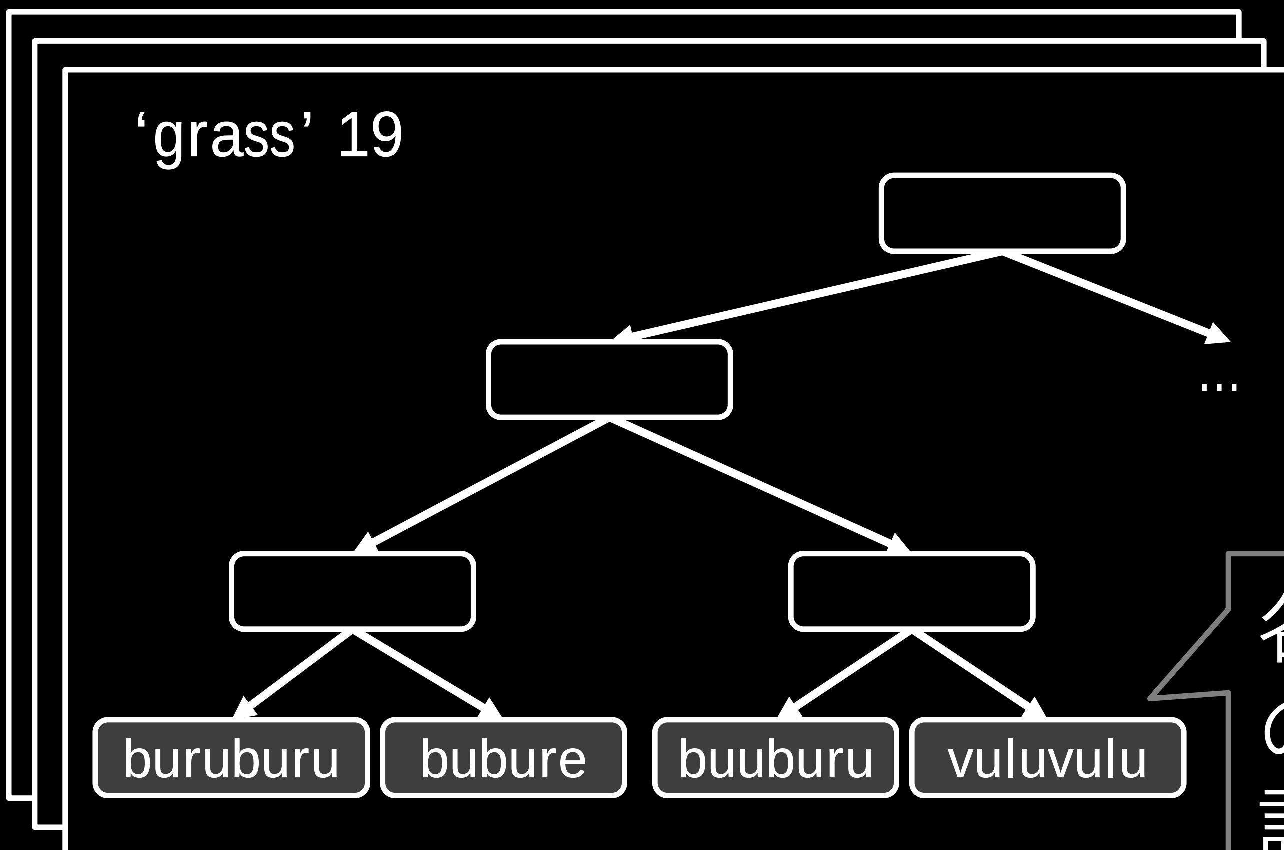
1. 対象の言語群について、語源を共有する語 (同源語, cognate) の候補を収集
   – 借用や偶然の一致を排除
2. 規則的な音対応を確立
   – 例外を個別に説明
3. 共通祖語を再構

確率モデルによる祖語再構 1/2



[Bouchard-Côté+, PNAS 2013]

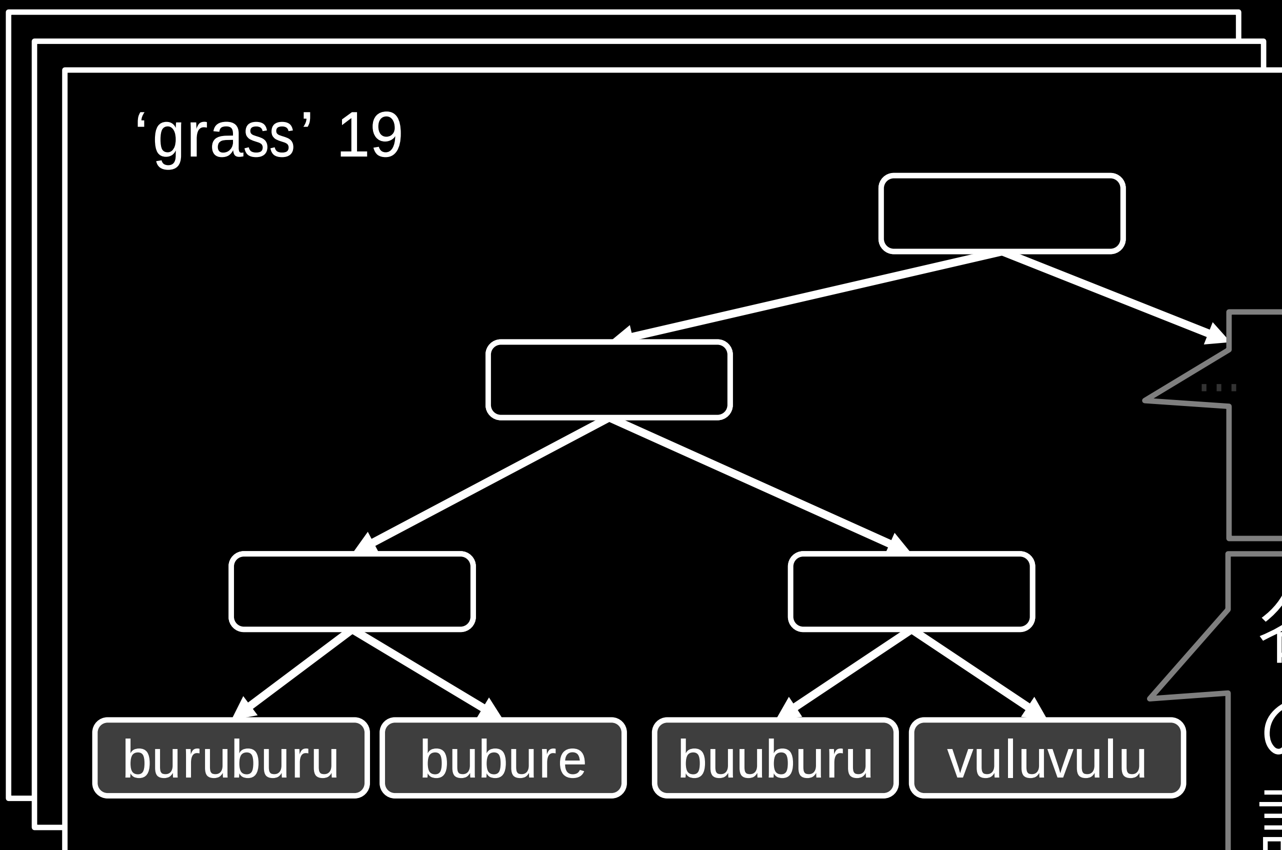
確率モデルによる祖語再構 1/2



各同源語群の現代語の語形は所与

[Bouchard-Côté+, PNAS 2013]

確率モデルによる祖語再構 1/2

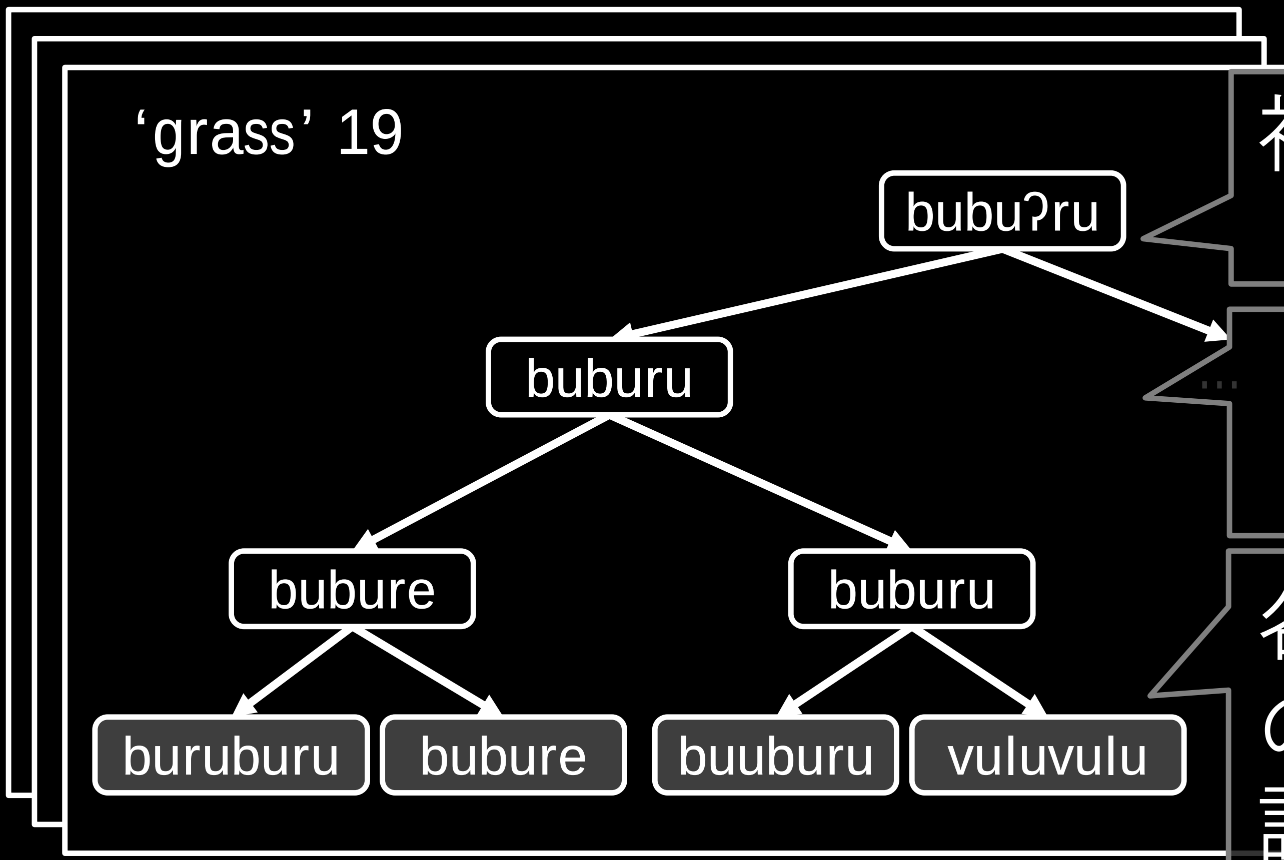


系統樹も
所与!

各同源語群
の現代語の
語形は所与

[Bouchard-Côté+, PNAS 2013]

確率モデルによる祖語再構 1/2



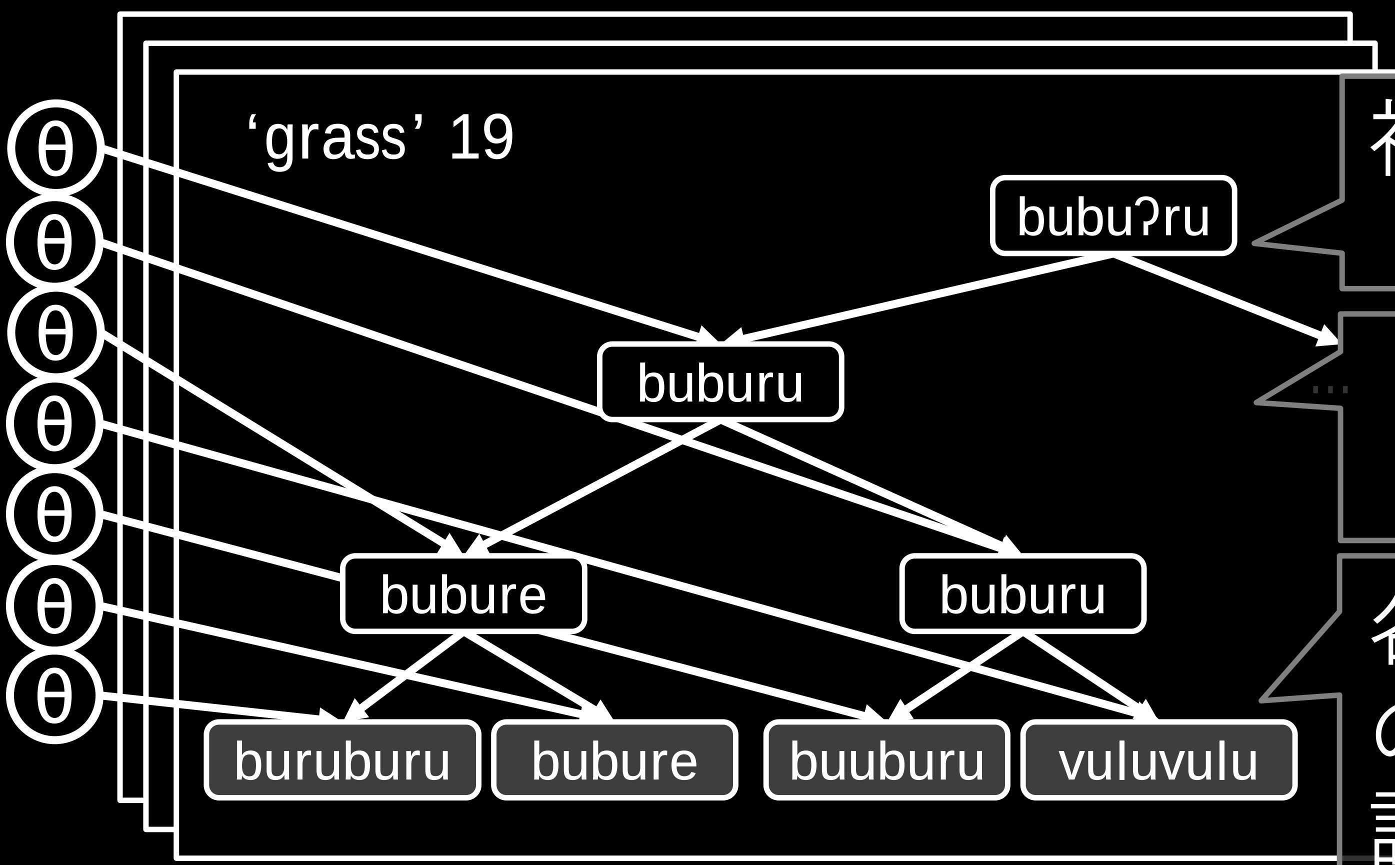
祖語の語形を推定

系統樹も所与!

各同源語群の現代語の語形は所与

[Bouchard-Côté+, PNAS 2013]

確率モデルによる祖語再構 1/2



祖語の語形を推定

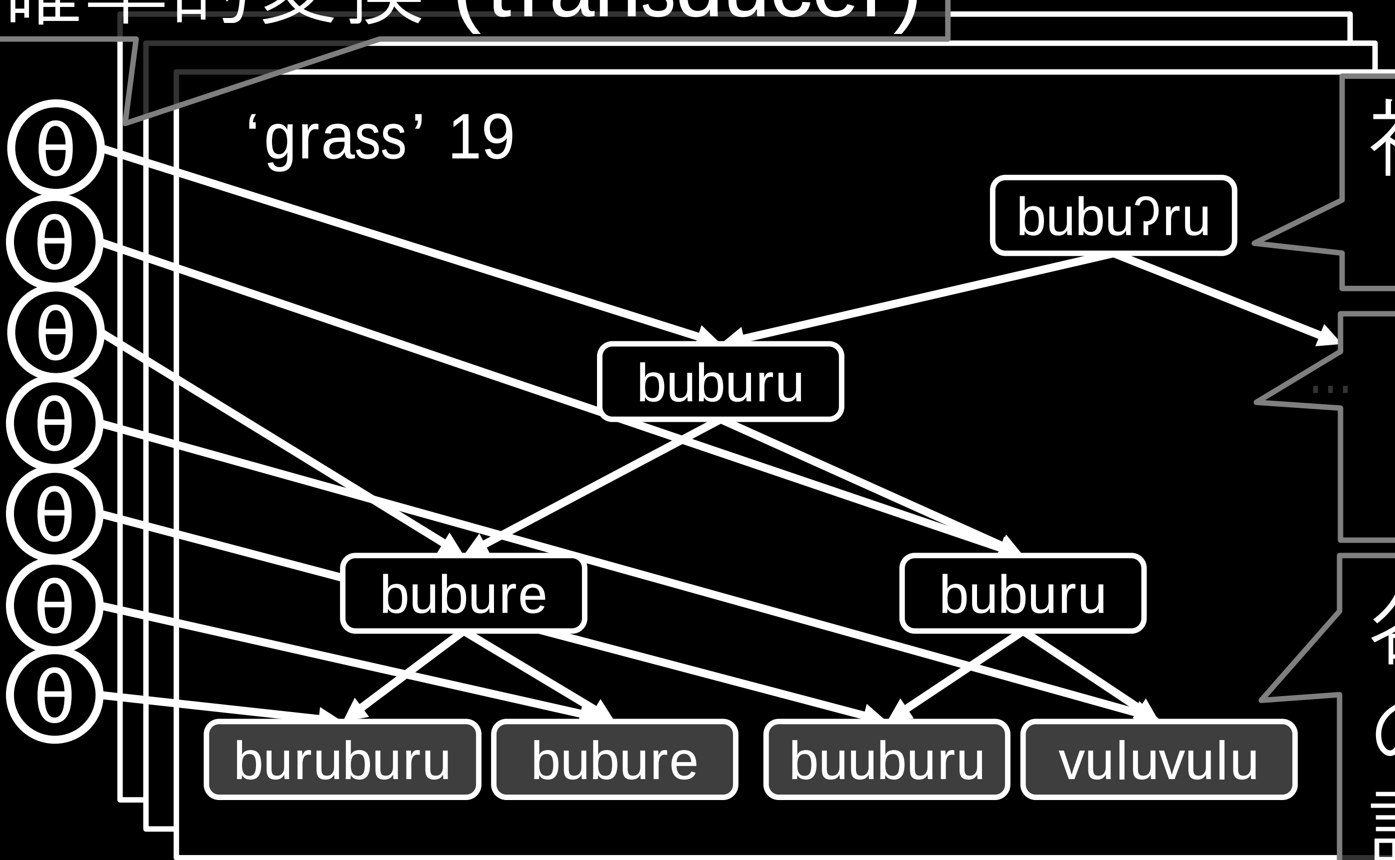
系統樹も所与!

各同源語群の現代語の語形は所与

[Bouchard-Côté+, PNAS 2013]

確率モデルによる祖語再構 1/2

文字列から文字列への
確率的変換 (transducer)



祖語の語形
を推定

系統樹も
所与!

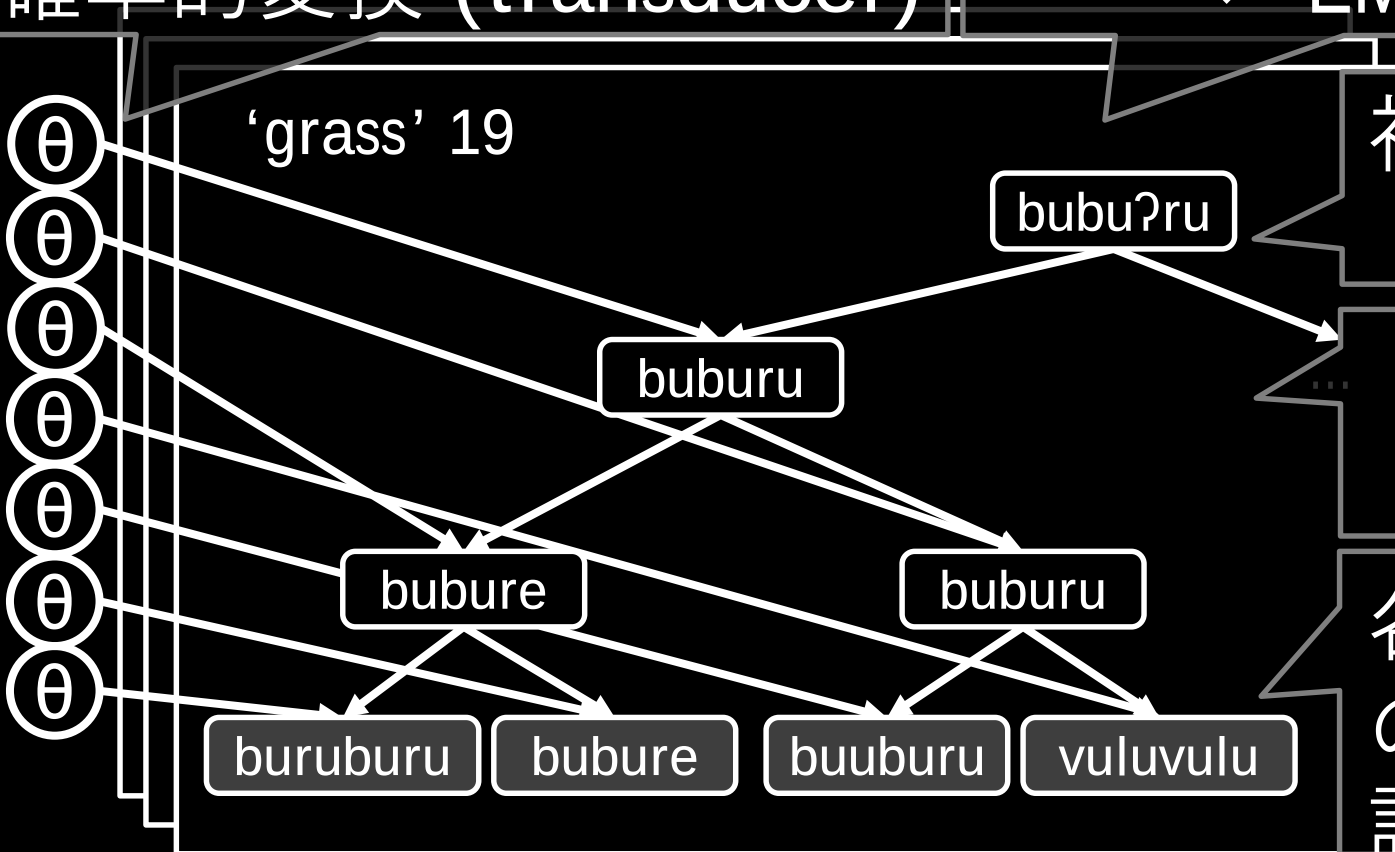
各同源語群
の現代語の
語形は所与

[Bouchard-Côté+, PNAS 2013]

確率モデルによる祖語再構 1/2

文字列から文字列への確率的変換 (transducer)

潜在変数とパラメータ ⇒ EMで推定



祖語の語形を推定

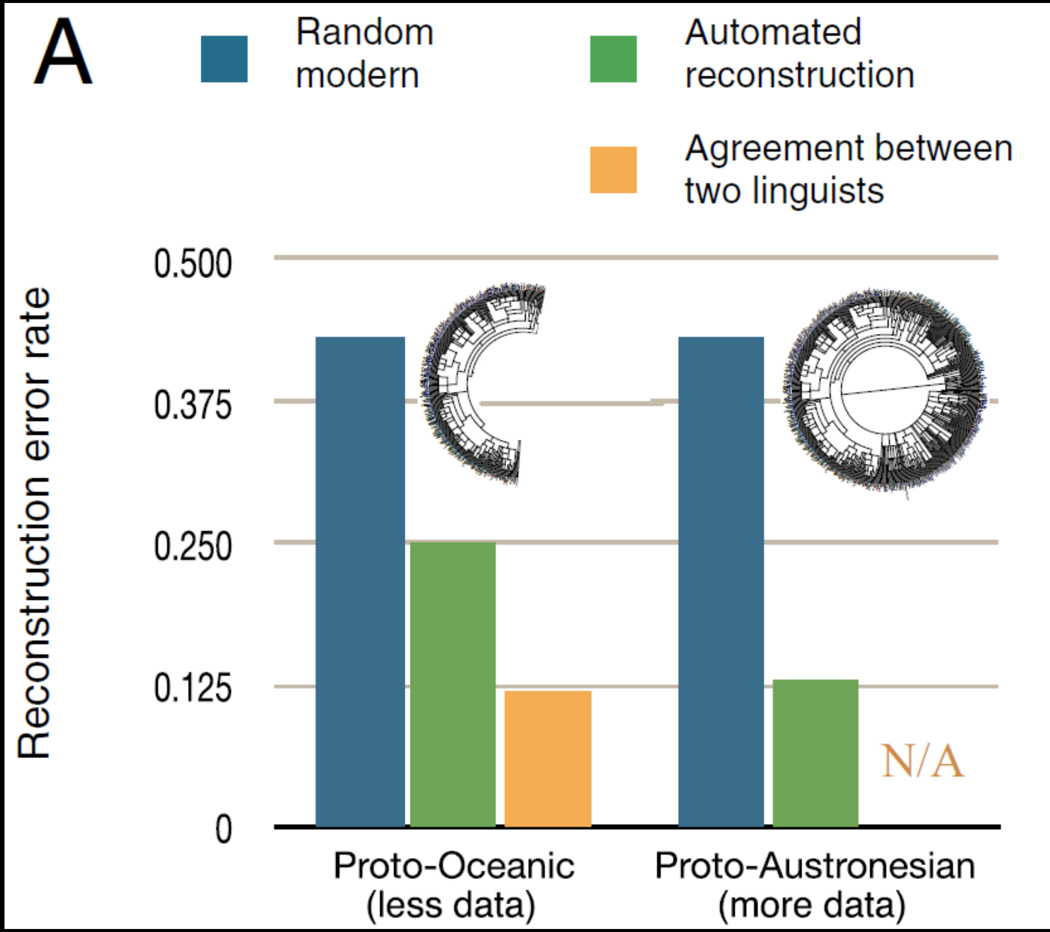
系統樹も所与!

各同源語群の現代語の語形は所与

[Bouchard-Côté+, PNAS 2013]

確率モデルによる祖語再構 2/2

- オーストロネシア語族 (659言語)
- 言語学者の再構した語形にかなり近い
- 言語に関する新たな知見は?



[Bouchard-Côté+, PNAS 2013]

音法則の限界

- そもそも同源語候補を大量に集めないと始まらない
  - 日本語と朝鮮語でそれができる可能性は低い
- さかのぼるのは6,000~7,000年が限度?
- 祖語の年代、位置のような連続値は直接推定できない

日本語と朝鮮語の間で同源語を確立するのは望み薄

- 候補が依然として347ペア [Whitman, 1985] しかないので量より質で勝負するしかない
  - cf. オーストロネシア語族: 同源語群7,708セット、142,661語形 [Bouchard-Côté+, PNAS 2013]
- そのほとんどが同源語と認めがたい [Vovin, 2010]

(28) (R) MK *pyé* ‘rice plant, grain of rice’[51] ~ OJ *po* ‘ear of rice’ < PJK *pye (Whitman 1985: 212).[52] It seems that Whitman assumes that OJ *po* is *pö* with the *otsu-rui* vowel /ö/, at least that would follow from his list of vocalic correspondences (1985: 129). Unfortunately, however, it is clearly WOJ *pô*, with *kō-rui* /ô/, because in the *Kojiki* the character 穗 ‘ear of rice’ is used as the *kungana* for the word *pô* ‘top’ in the phrase *NAMÎ-NÖ pô* ‘top of the waves’ (Igarashi 1969: 106, 127). Since WOJ *pô* ‘top’ is also spelled with the *ongana* 本 /pô/ in other cases, e.g., *ma-pô* (麻本) ‘pinnacle = true top’ (*KK* 30) and *pô-tu ye* (本都延) ‘top branches’ (*KK* 43), there is no doubt that ‘ear of rice’ was *pô* in Western Old Japanese. Thus, the vowels in Korean and Japonic really do not match, so I reject this etymology.

今日のおはなし

1. 音法則
2. 基礎語彙 言語年代学 (1950年代~)
   Bayes統計の応用 (2000年代~)
3. 類型論

語彙の2値ベクトル化

	水	大きい
英語	water	big
ドイツ語	Wasser	gross
ロシア語	вода	большой великий
フランス語	eau	grand
イタリア語	acqua	grande

語彙の2値ベクトル化

同源語群	水	大きい	
英語	water	big	
ドイツ語	Wasser	gross	
ロシア語	вода	большой	великий
フランス語	eau	grand	
イタリア語	acqua	grande	

語彙の2値ベクトル化

同源語群	水	大きい		
英語	1 water	3 big		{1, 3}
ドイツ語	Wasser	4 gross		{1, 4}
ロシア語	вода	5 большой	6 великий	{1, 5, 6}
フランス語	2 eau	7 grand		{2, 7}
イタリア語	acqua	grande		{2, 7}

語彙の2値ベクトル化

同源語群	水	大きい		
英語	1 water	3 big		1010000
ドイツ語	Wasser	4 gross		1001000
ロシア語	вода	5 большой	6 великий	1000110
フランス語	2 eau	7 grand		0100001
イタリア語	acqua	grande		0100001

語彙の2値ベクトル化

言語の近さが
簡単に計算できる

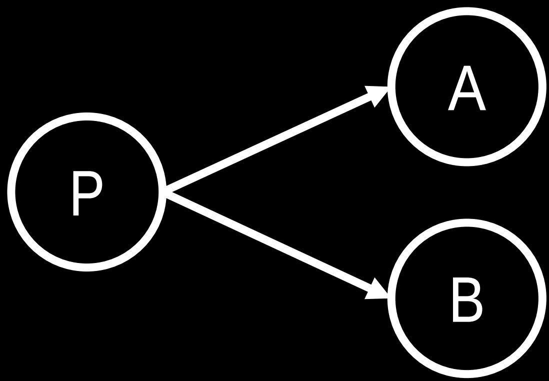
同源語群	水	大きい		
	1	3		
英語	water	big		1010000
		4		
ドイツ語	Wasser	gross		1001000
		5	6	
ロシア語	вода	большой	великий	1000110
	2	7		
フランス語	eau	grand		0100001
イタリア語	acqua	grande		0100001

言語年代学 (glottochronology) 1/2

- 人類学･言語学のアメリカ･インディアン分類問題 [Sapir, 1921]
  - 音法則によらず語彙の異同で分類したい
- 考古学での放射性炭素年代測定 [Libby, 1946]
- 祖語の年代推定への応用 [Swadesh, 1948,1951]
  - インド･ヨーロッパ語族から基礎語彙の残存率を求め、アメリカ･インディアンの言語に適用
- 日本語方言、アイヌ語方言への適用 [服部, 1954][服部+, 1960]
- 系統論の消極的手がかり: 日本語と朝鮮語が仮に同系だとしても、祖語は6,000年以上前 [服部, 1999[1956]]
- 生物学の分子時計仮説 [Zuckerkandl+, 1965]

言語年代学 (glottochronology) 2/2

$$t = \frac{\log c}{2 \log r}$$



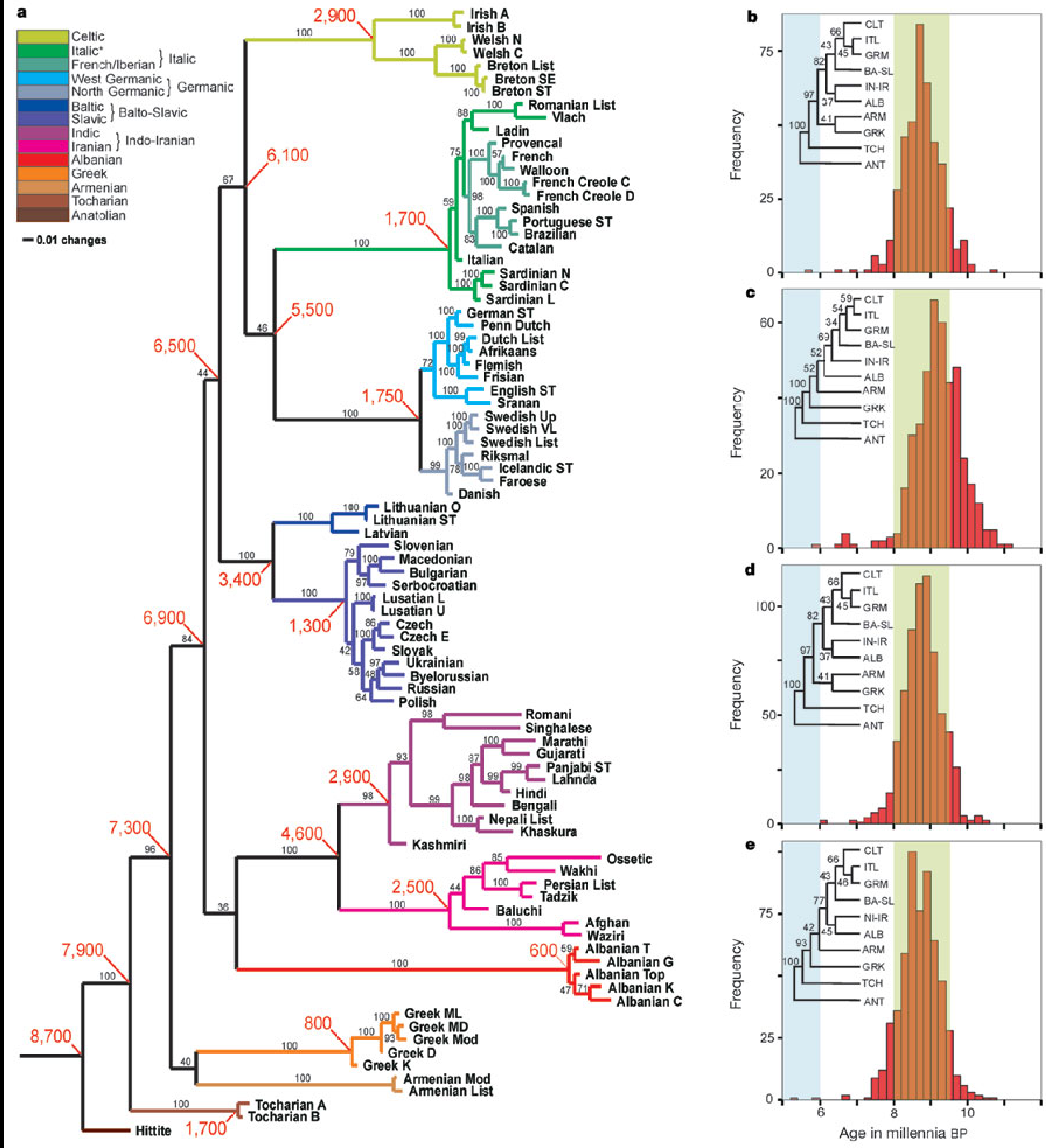
- t: 祖語Pの年代 (単位: 千年)
- c: A, Bの基礎語彙共有率
- r: 基礎語彙の残存率 (200項目で0.81)

年代	1K	2K	3K	4K	5K	6K	7K	8K	9K
共有率 (r=.81)	.66	.43	.28	.19	.12	.08	.05	.03	.02

言語年代学への批判

- 基礎語彙の残存率が一定という仮定がなりたたない
  - 古ノルド語からアイスランド語への残存率は>0.95 [Bergsland+, 1962]
- 同系言語からの借用は区別が難しい
- 基礎語彙の中でも語によって安定性が異なるのでは?

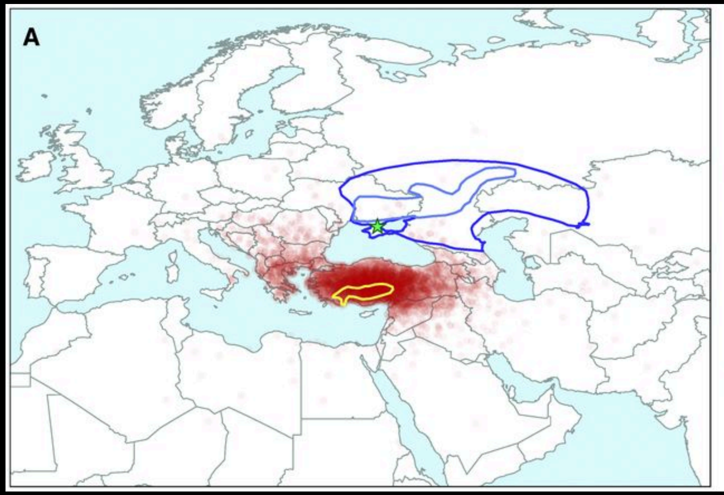
Bayes
系統推定
モデル



[Gray+,
Nature 2003]

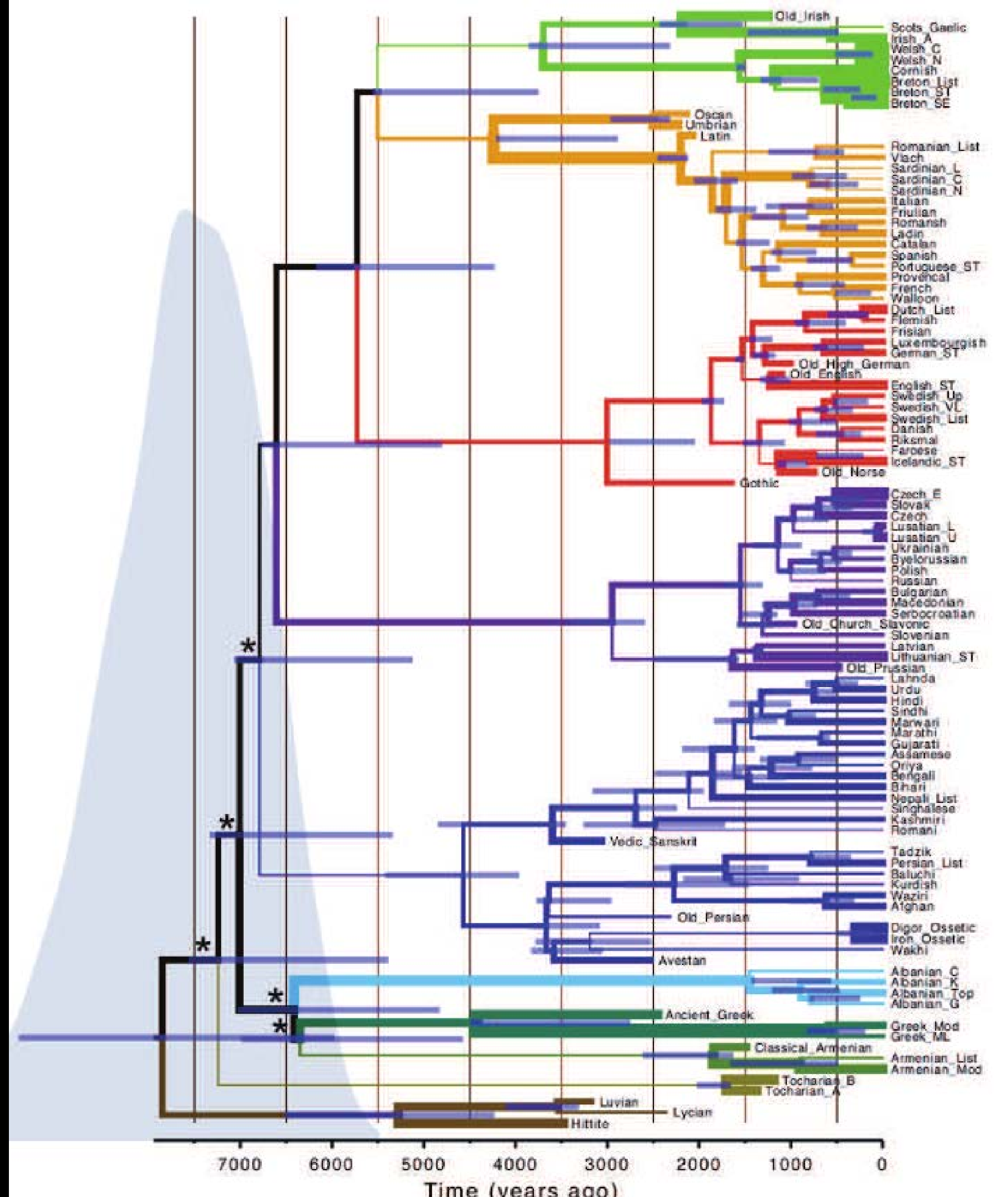
インド・ヨーロッパ祖語の年代と故地 (Urheimat)

1. クルガン仮説
   – 5,000-6,000年前
   – 黒海周辺のステップ
   – 遊牧民の軍事的征服
   – 言語学者の広い支持
2. アナトリア仮説 (Gray+はこちら)
   – 8,000-9,500年前
   – アナトリア
   – 農耕とともに拡大
   – 考古学者 (Renfrew) の農耕･言語同時伝播モデル
   – 批判: 印欧語アナトリア語派は祖語からかけ離れすぎ

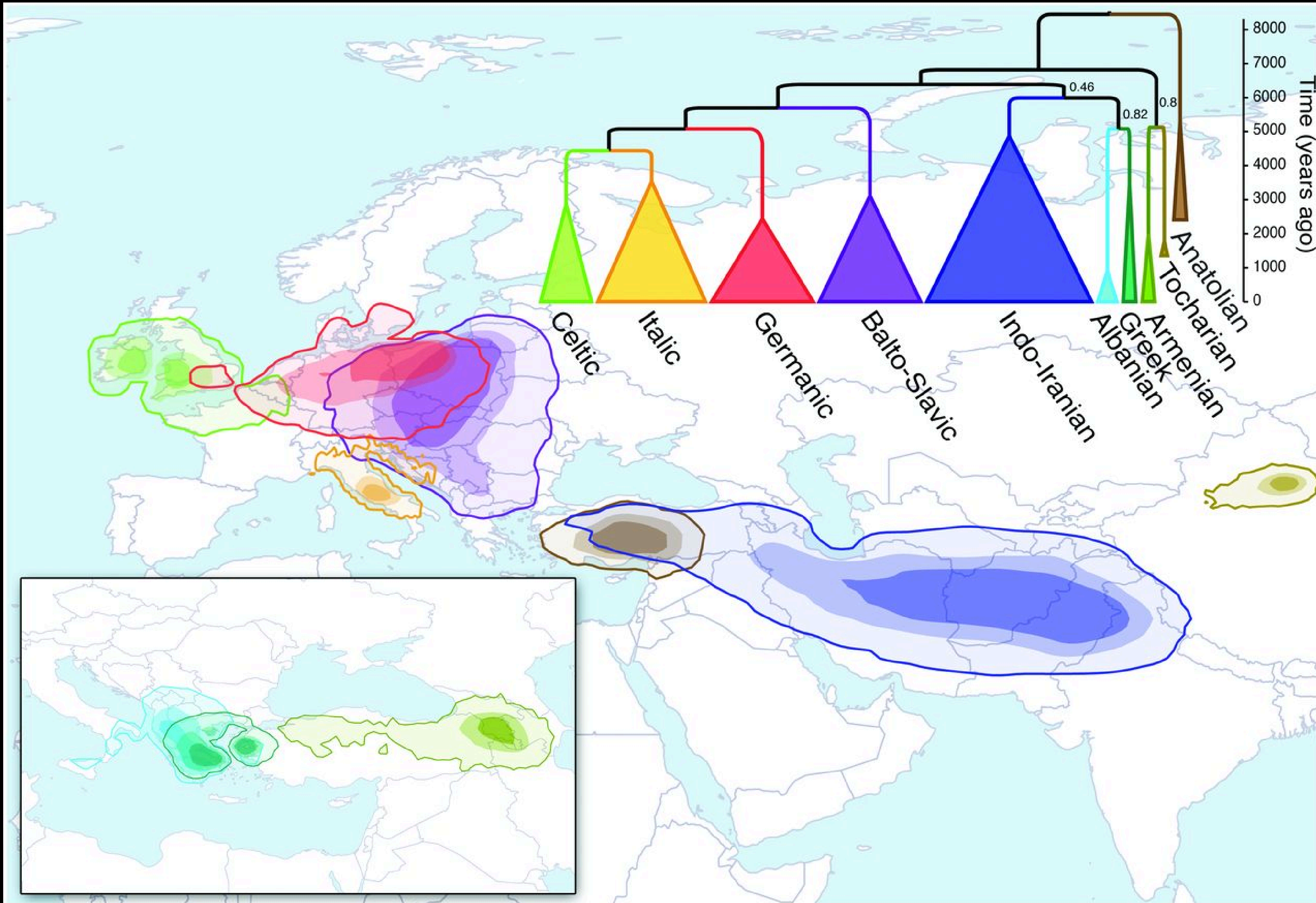


[Bouckaert+, Science 2012]

Bayes
系統推定
モデル

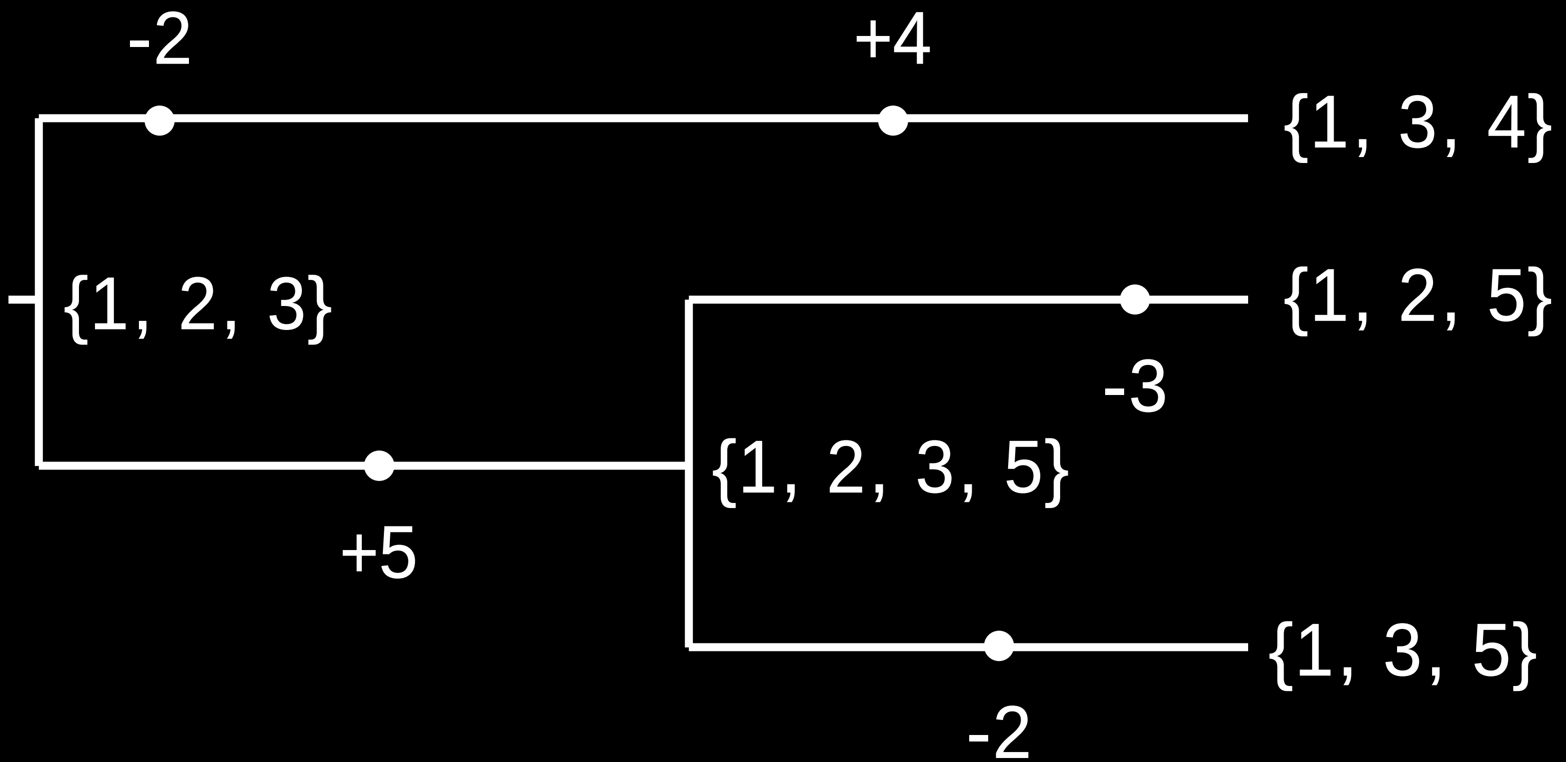


[Bouckaert+, Science 2012]



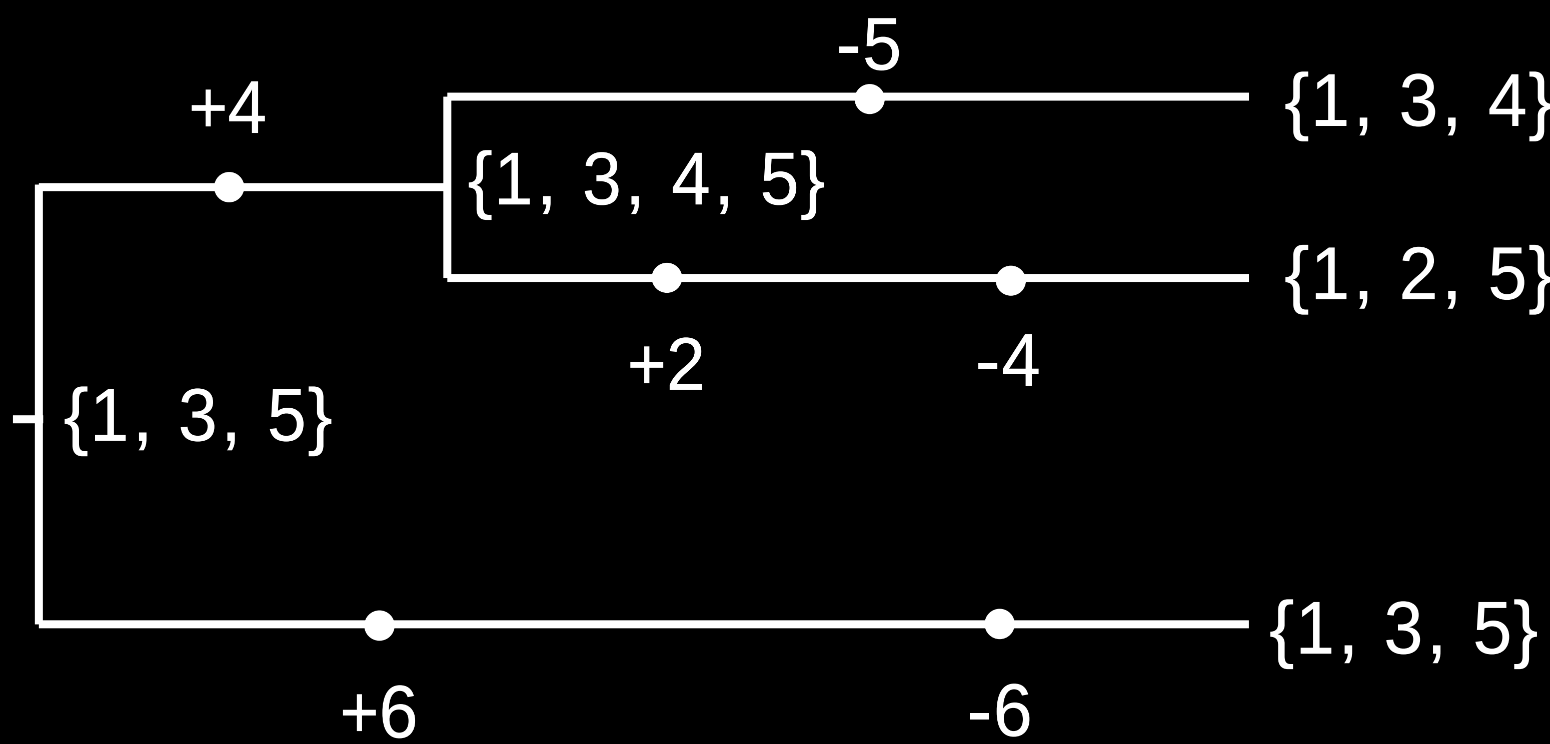
[Bouckaert+, Science 2012]

Bayes系統推定 1/3



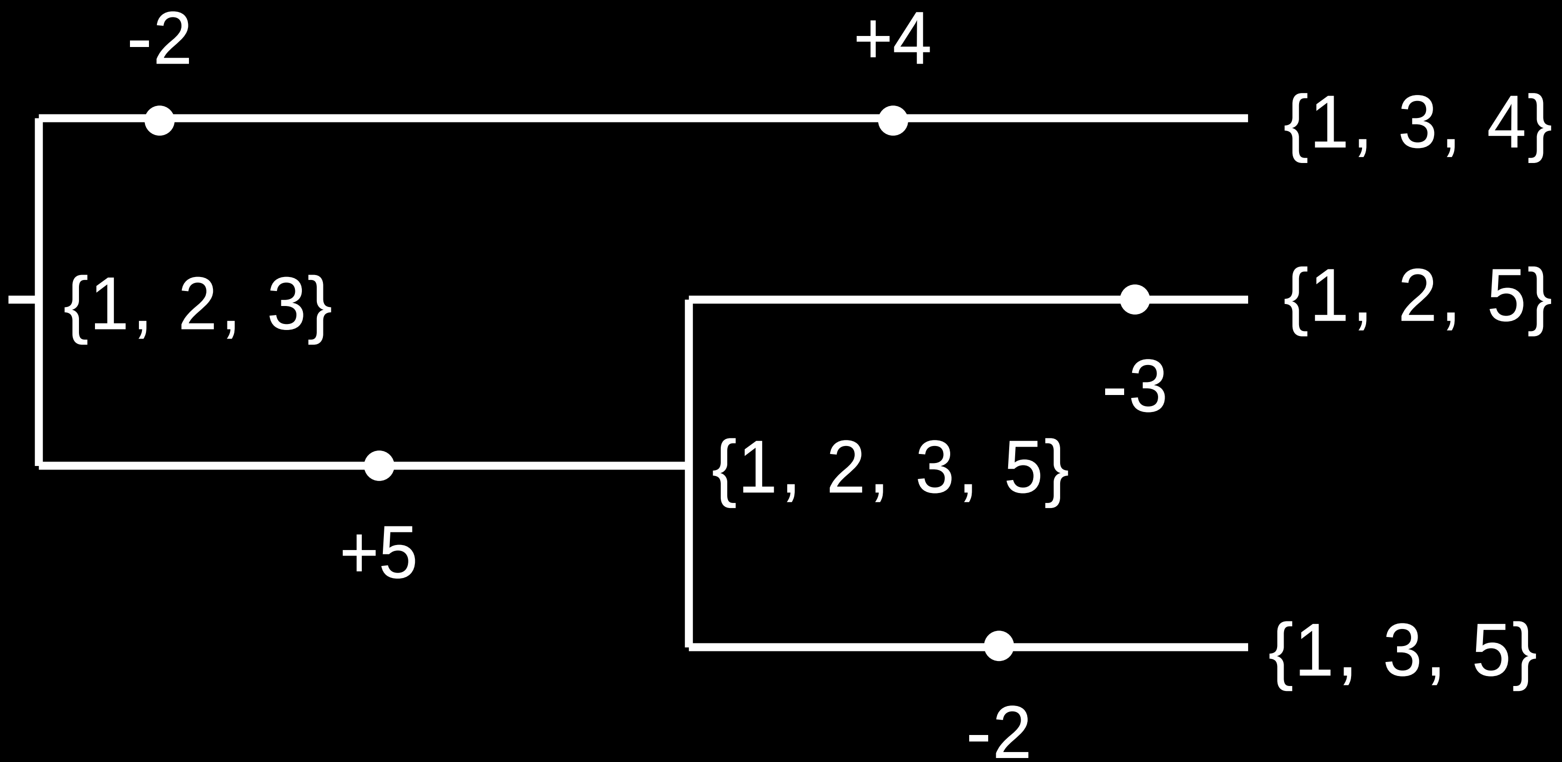
- 二分木を推定 (cf. 言語ペアの比較)
- 同源語の生死を直接モデル化 (cf. 共有率)

Bayes系統推定 1/3



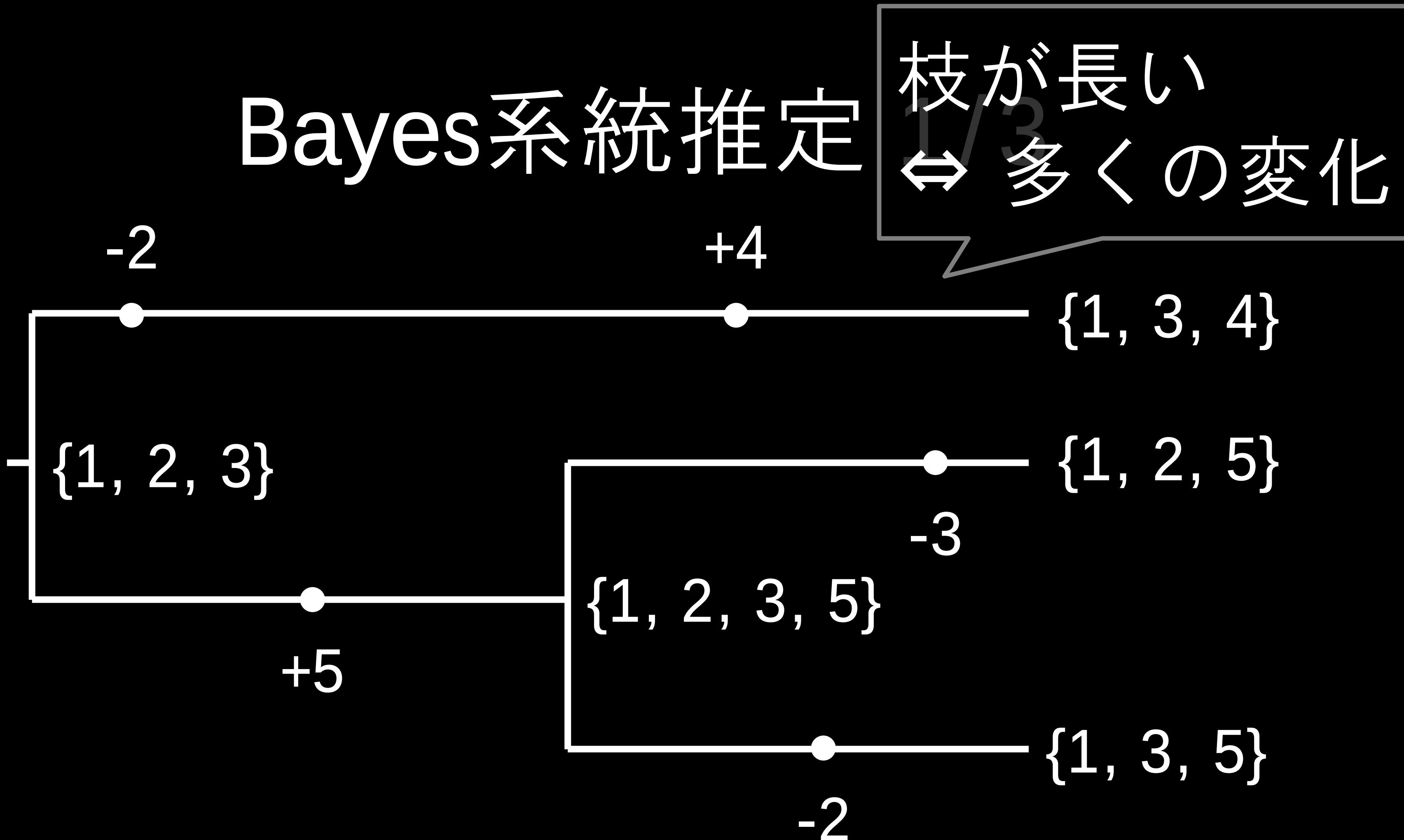
- 二分木を推定 (cf. 言語ペアの比較)
- 同源語の生死を直接モデル化 (cf. 共有率)

Bayes系統推定 1/3



- 二分木を推定 (cf. 言語ペアの比較)
- 同源語の生死を直接モデル化 (cf. 共有率)

Bayes系統推定 1/3



- 二分木を推定 (cf. 言語ペアの比較)
- 同源語の生死を直接モデル化 (cf. 共有率)

Bayes系統推定による拡張 2/3

- 語の誕生 (0→1)、死亡 (1→0) のモデル
  - cf. DNAの突然変異 (ACGTの置換モデル)

遷移行列: $Q = \begin{pmatrix} -\alpha & \alpha \\ \beta & -\beta \end{pmatrix}$

言語変化の速度を制御

遷移確率: $P(x = j | \pi(x) = i, t) = \exp(tQ)_{i,j}$

- 緩和時計モデル

遷移率を一定にせず、事前分布を置く

$\alpha_i \sim \mathrm{Exp}(\lambda)$ or $\alpha_i \sim \mathrm{LogNormal}(\mu, \sigma^2)$

Bayes系統推定による拡張 3/3

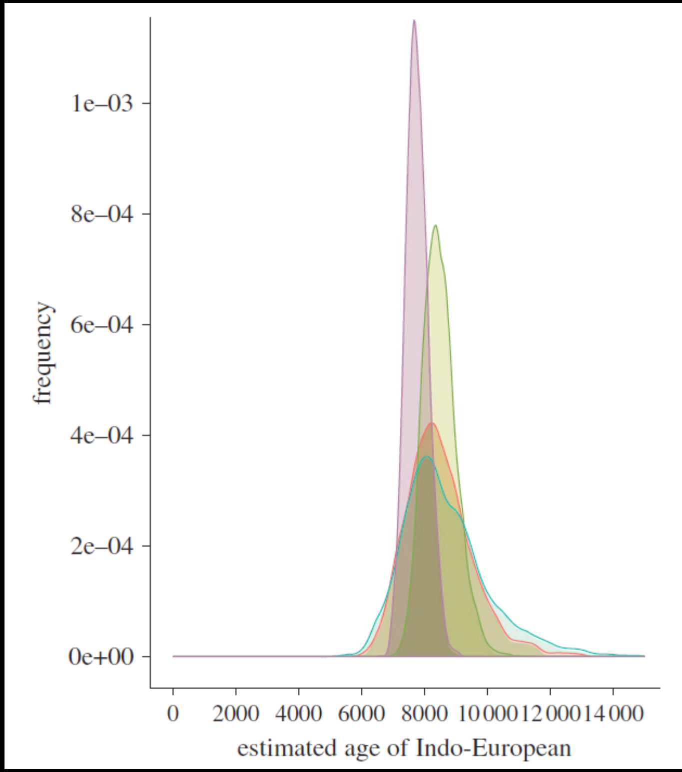
- 年代較正 (calibration)
  - 古代語の年代に事前分布をおく

    $\mathrm{age}(\mathrm{Latin}) \sim \mathcal{N}(\mu = 2050.0\ \mathrm{BP}, \sigma = 75.0)$
  - 間接的に遷移率を制御
- ベイズ推定
  - パラメータθを点推定するのではなく、事前分布をおいて潜在変数扱いする

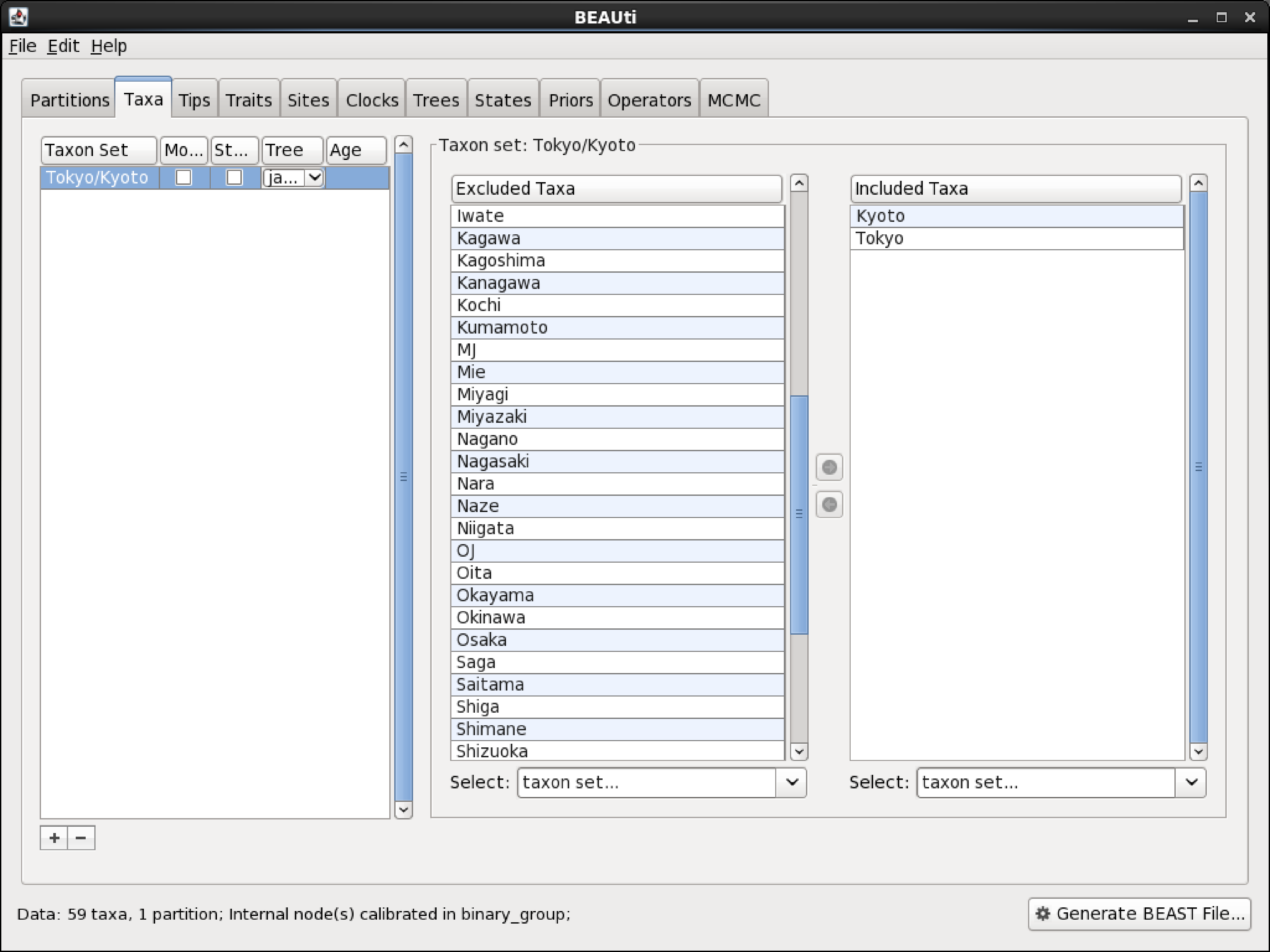
    P(潜在変数|観測変数, ハイパーパラメータ)
  - サンプリング (≒ランダム探索) で近似的に求解

結局、Bayes系統推定は言語年代学と何が違うのか

- 変化の速度 (遷移率) が一定という制約を緩和
- 代わりにいくつかの古代語の絶対年代をソフトな制約として与えて年代較正
- サンプリングにより、祖語の年代を点ではなく、分布として推定

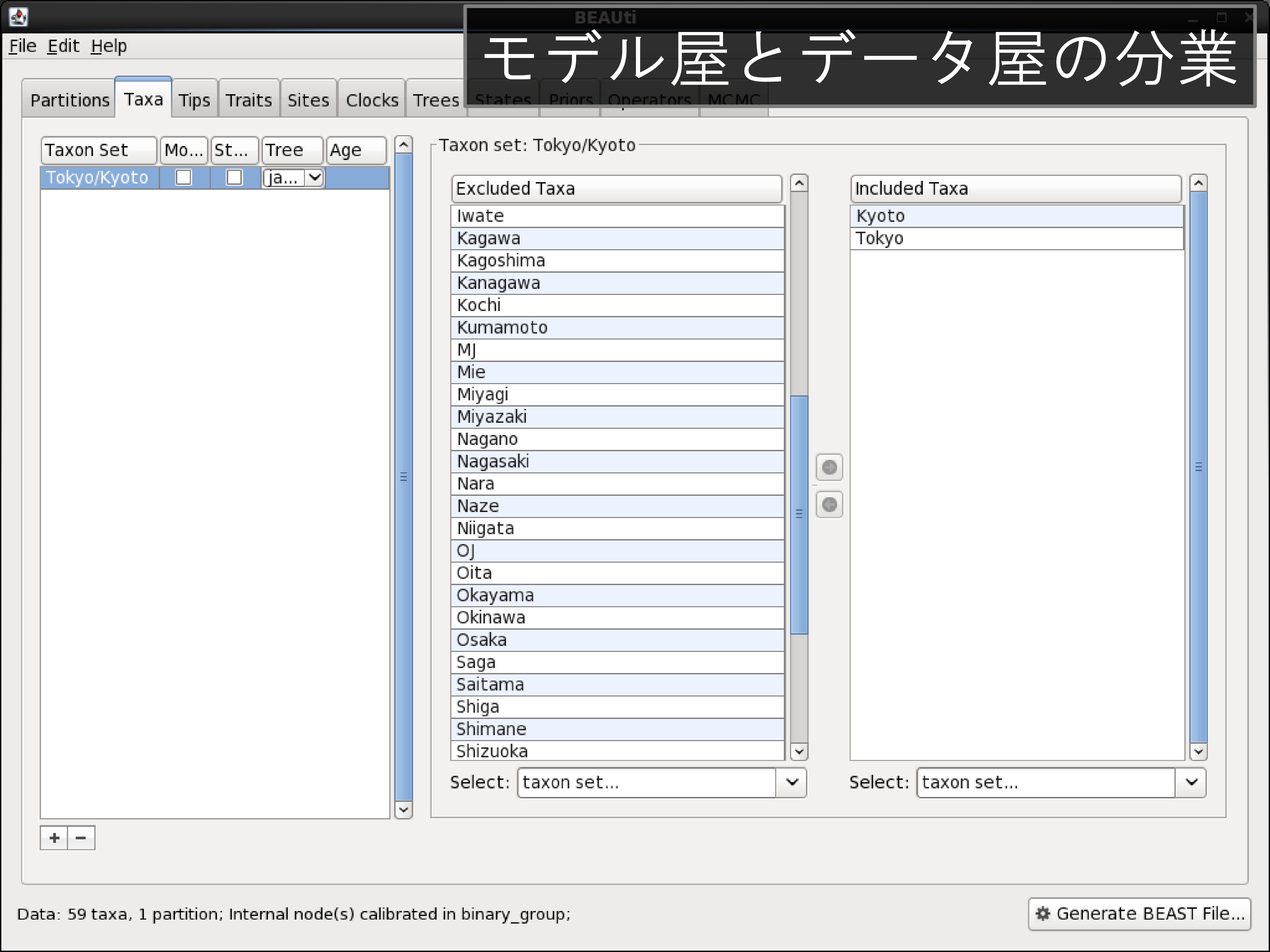


[Gray+, 2011]

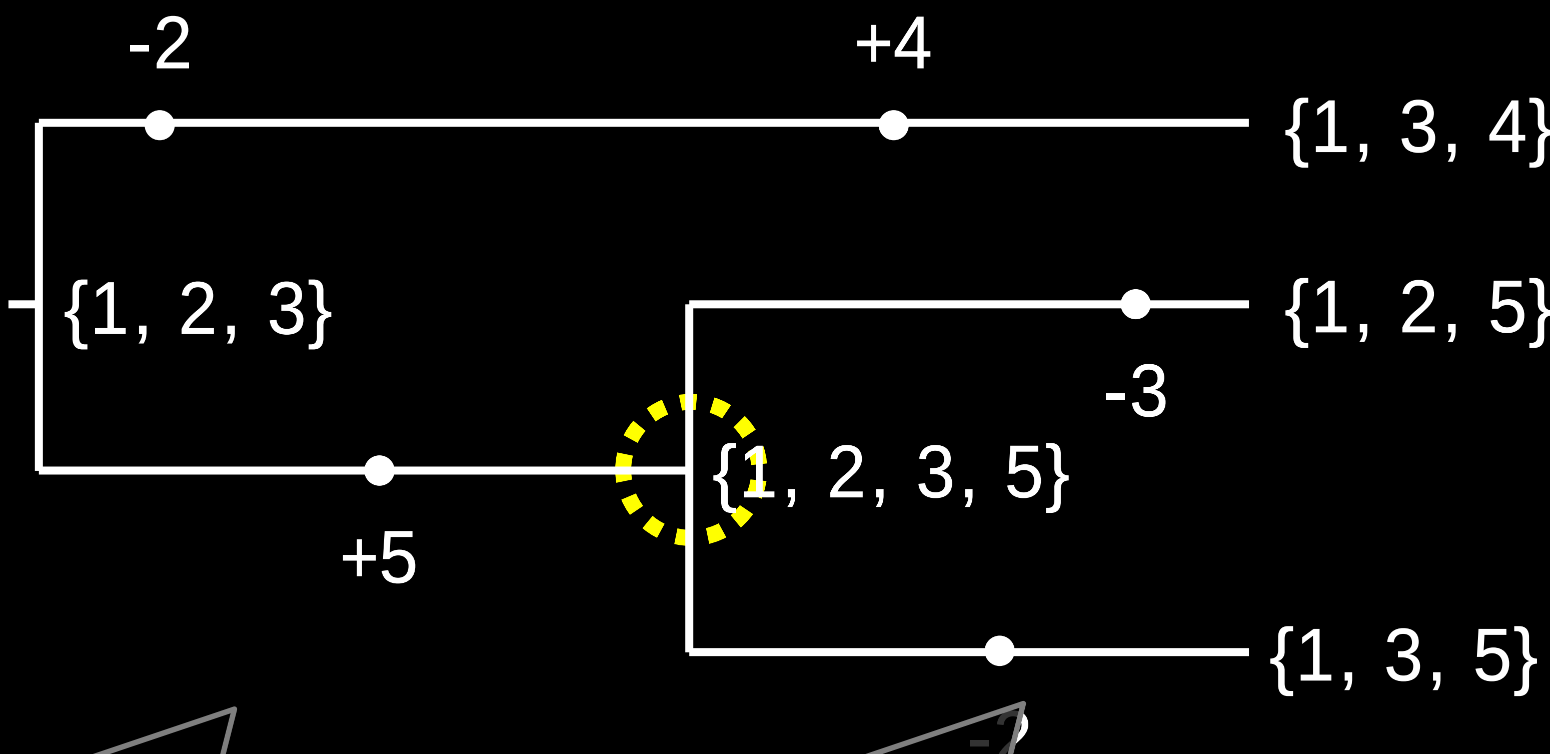


BEAUti
File Edit Help
Partitions
Taxa
Tips
Traits
Sites
Clocks
Trees
States
Priors
Operators
MCMC
Taxon Set
Mo...
St...
Tree
Age
Tokyo/Kyoto
ja...
Taxon set: Tokyo/Kyoto
Excluded Taxa
Iwate
Kagawa
Kagoshima
Kanagawa
Kochi
Kumamoto
MJ
Mie
Miyagi
Miyazaki
Nagano
Nagasaki
Nara
Naze
Niigata
OJ
Oita
Okayama
Okinawa
Osaka
Saga
Saitama
Shiga
Shimane
Shizuoka
Select: taxon set...
Included Taxa
Kyoto
Tokyo
Select: taxon set...
Data: 59 taxa, 1 partition; Internal node(s) calibrated in binary_group;
Generate BEAST File...

モデル屋とデータ屋の分業



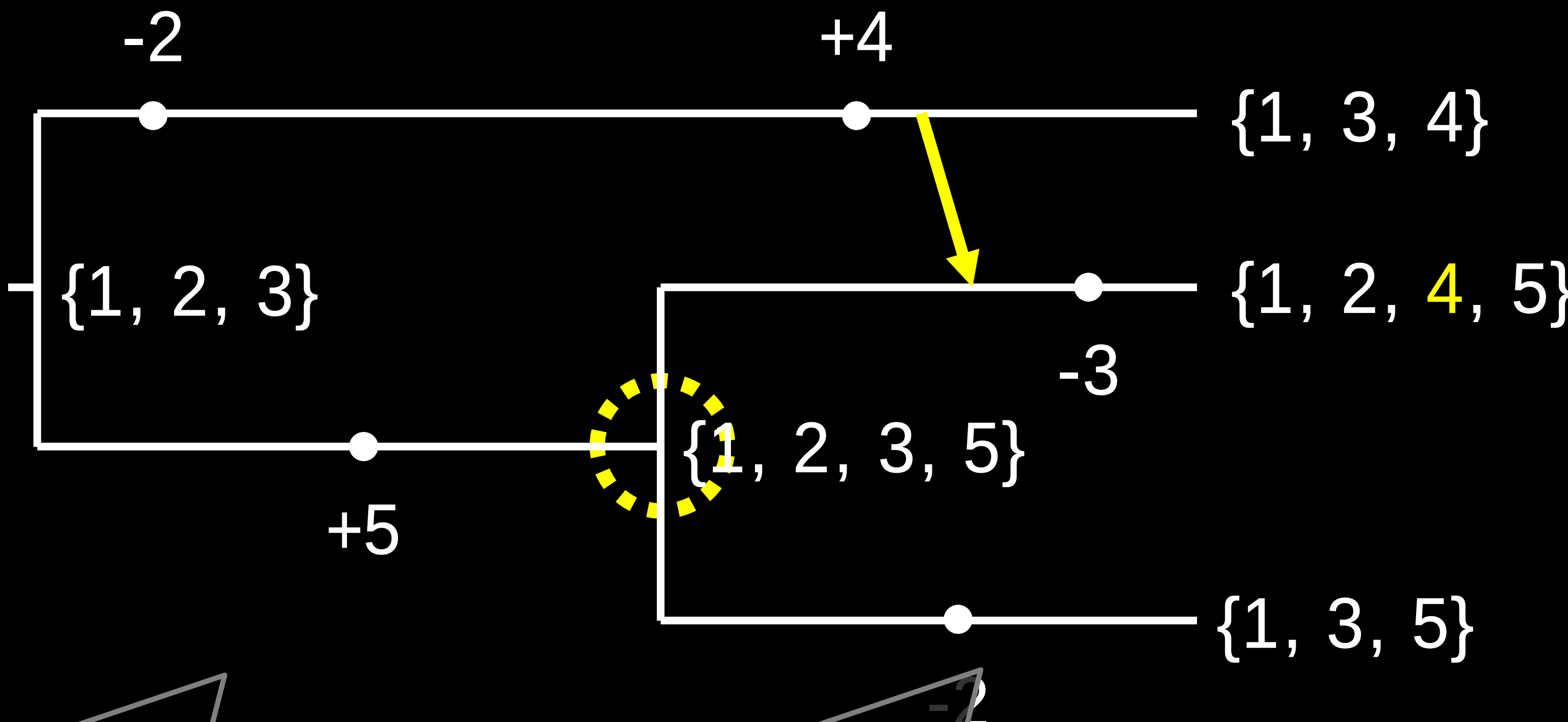
疑問: 本当に木構造が適切か



分岐前は
同一の状態

分岐後は
別々に変化

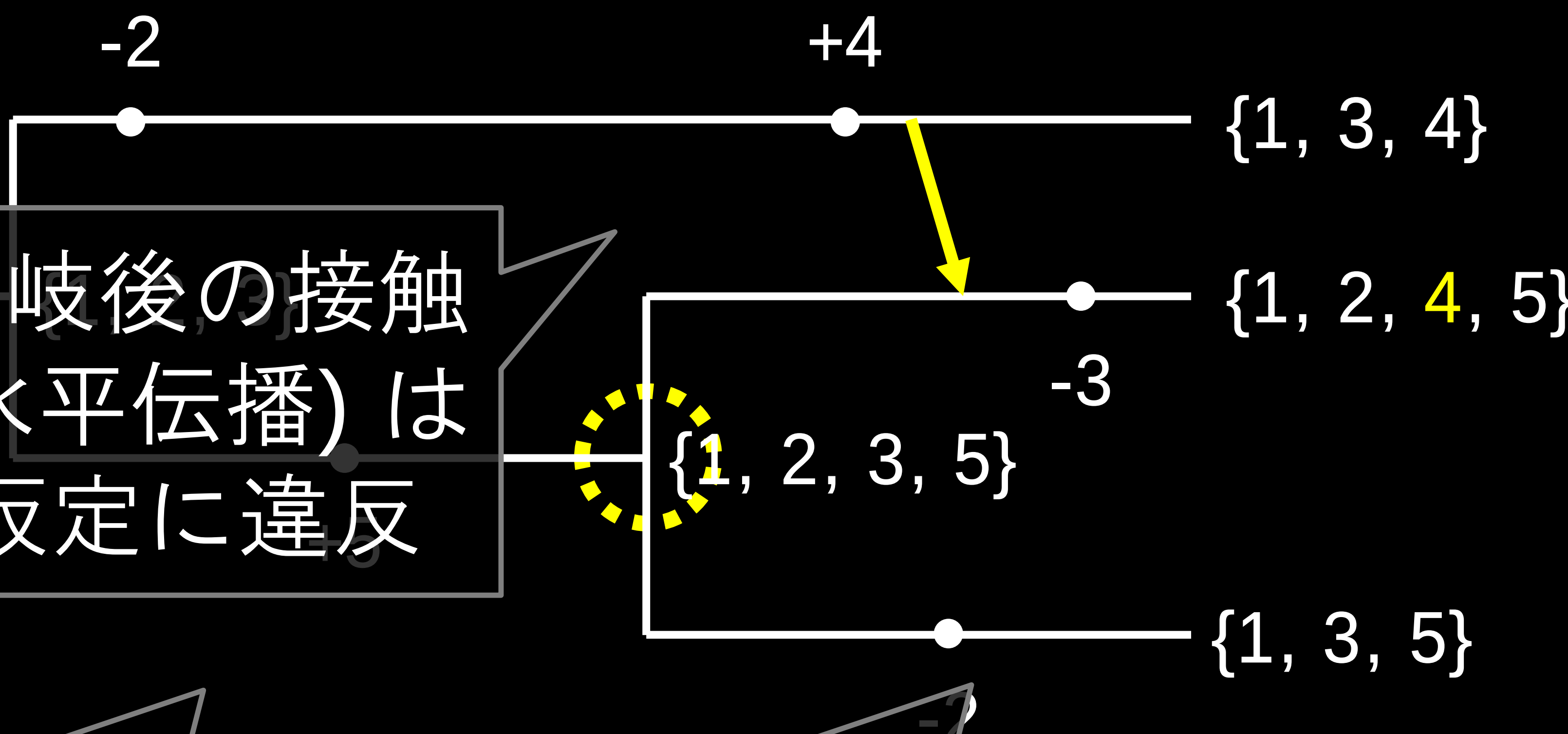
疑問: 本当に木構造が適切か



分岐前は
同一の状態

分岐後は
別々に変化

疑問: 本当に木構造が適切か



分岐後の接触
(水平伝播) は
仮定に違反

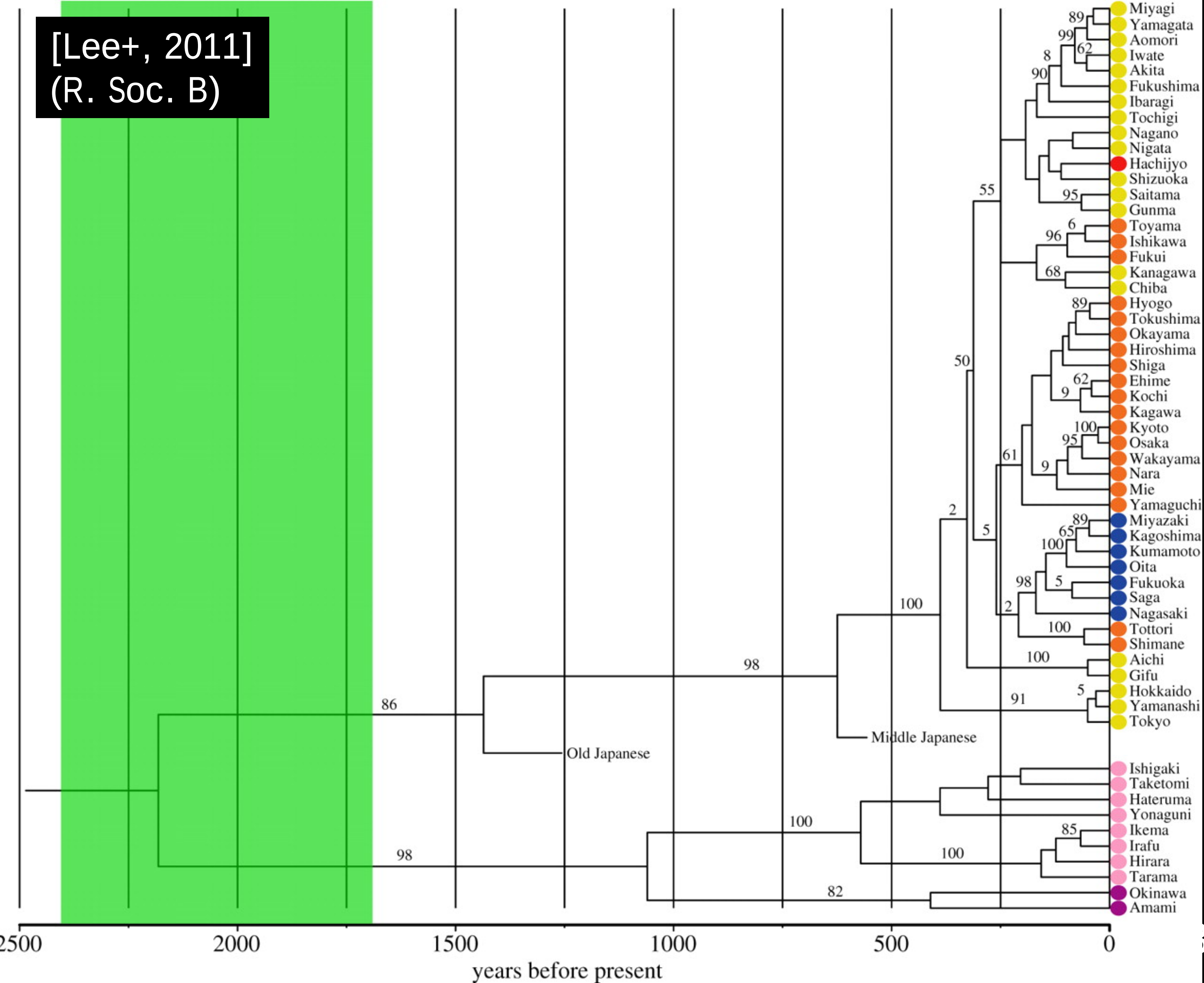
分岐前は
同一の状態

分岐後は
別々に変化

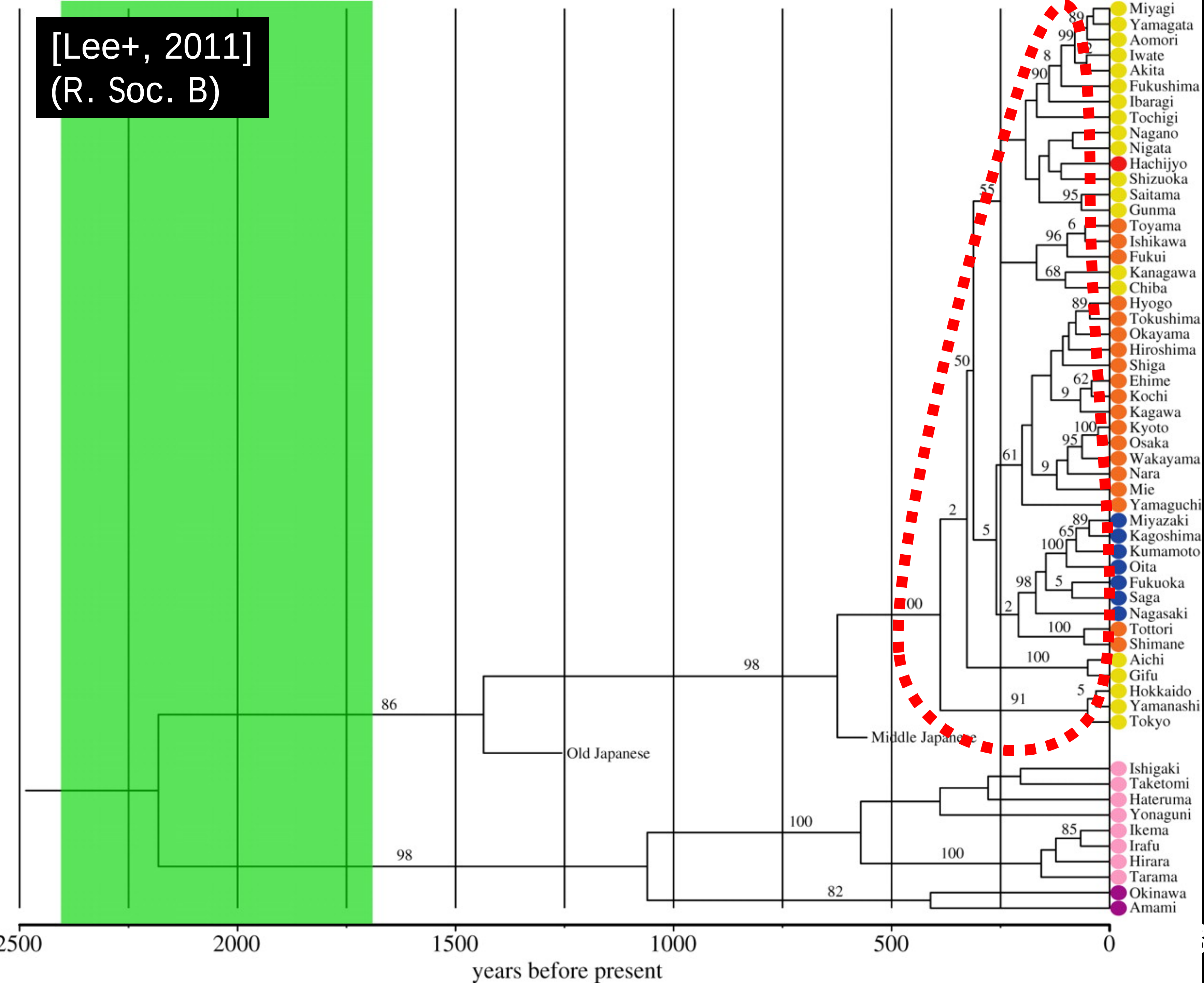
日本語方言への適用

[Lee+, 2011]
(R. Soc. B)

[Lee+, 2011]
(R. Soc. B)

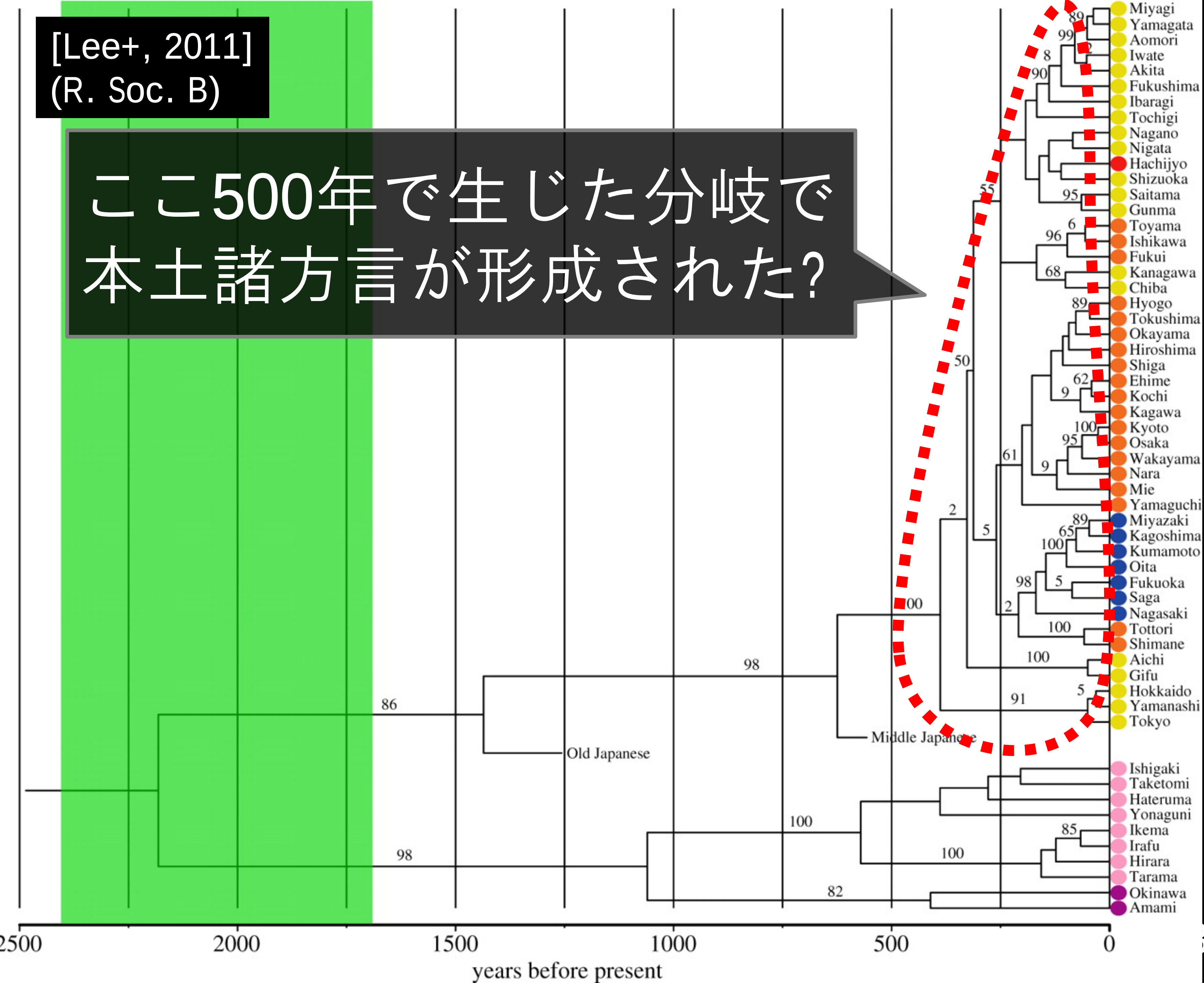


[Lee+, 2011]
(R. Soc. B)



[Lee+, 2011]
(R. Soc. B)

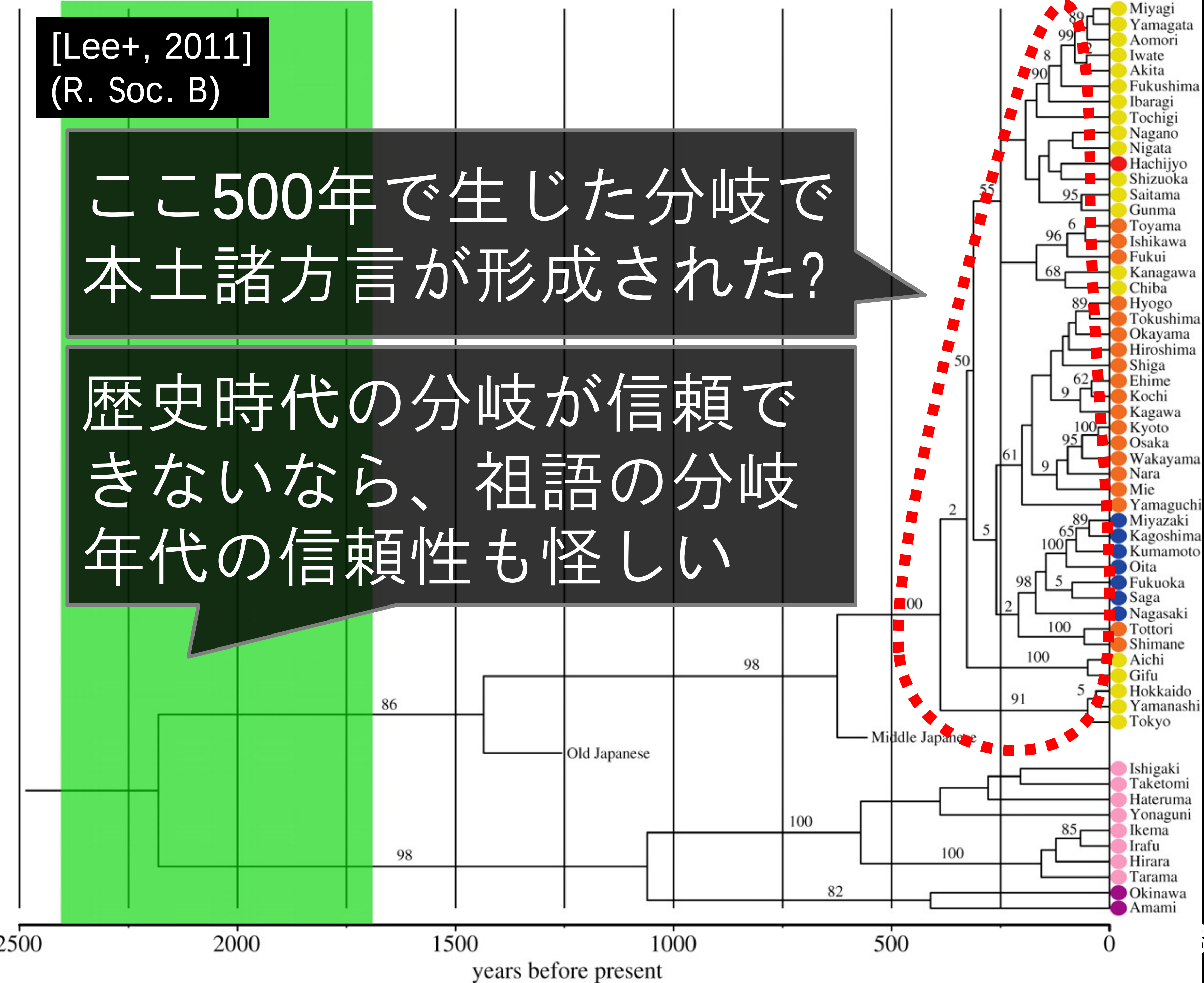
ここ500年で生じた分岐で
本土諸方言が形成された?



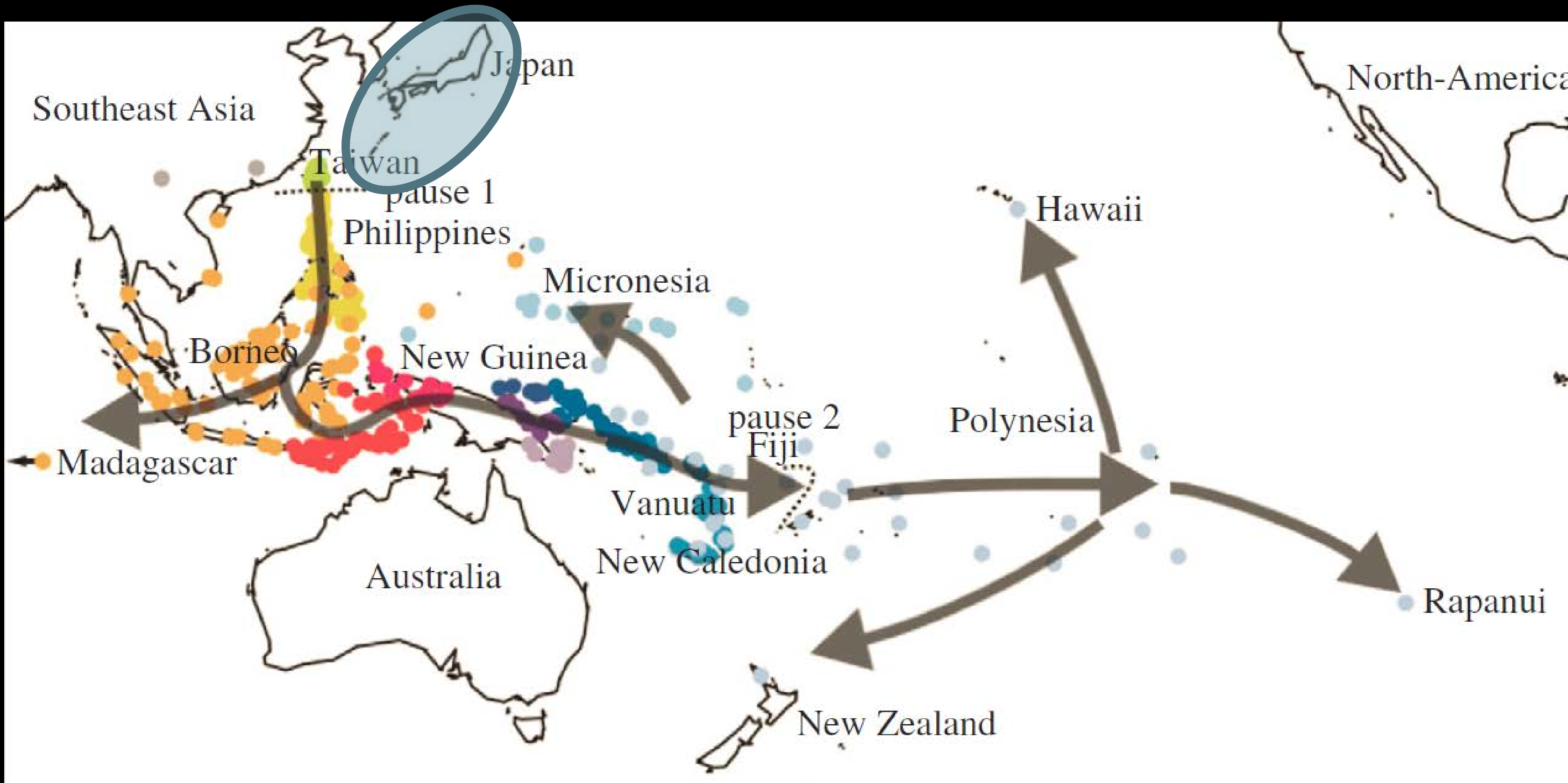
[Lee+, 2011]
(R. Soc. B)

ここ500年で生じた分岐で本土諸方言が形成された?

歴史時代の分岐が信頼できないなら、祖語の分岐年代の信頼性も怪しい

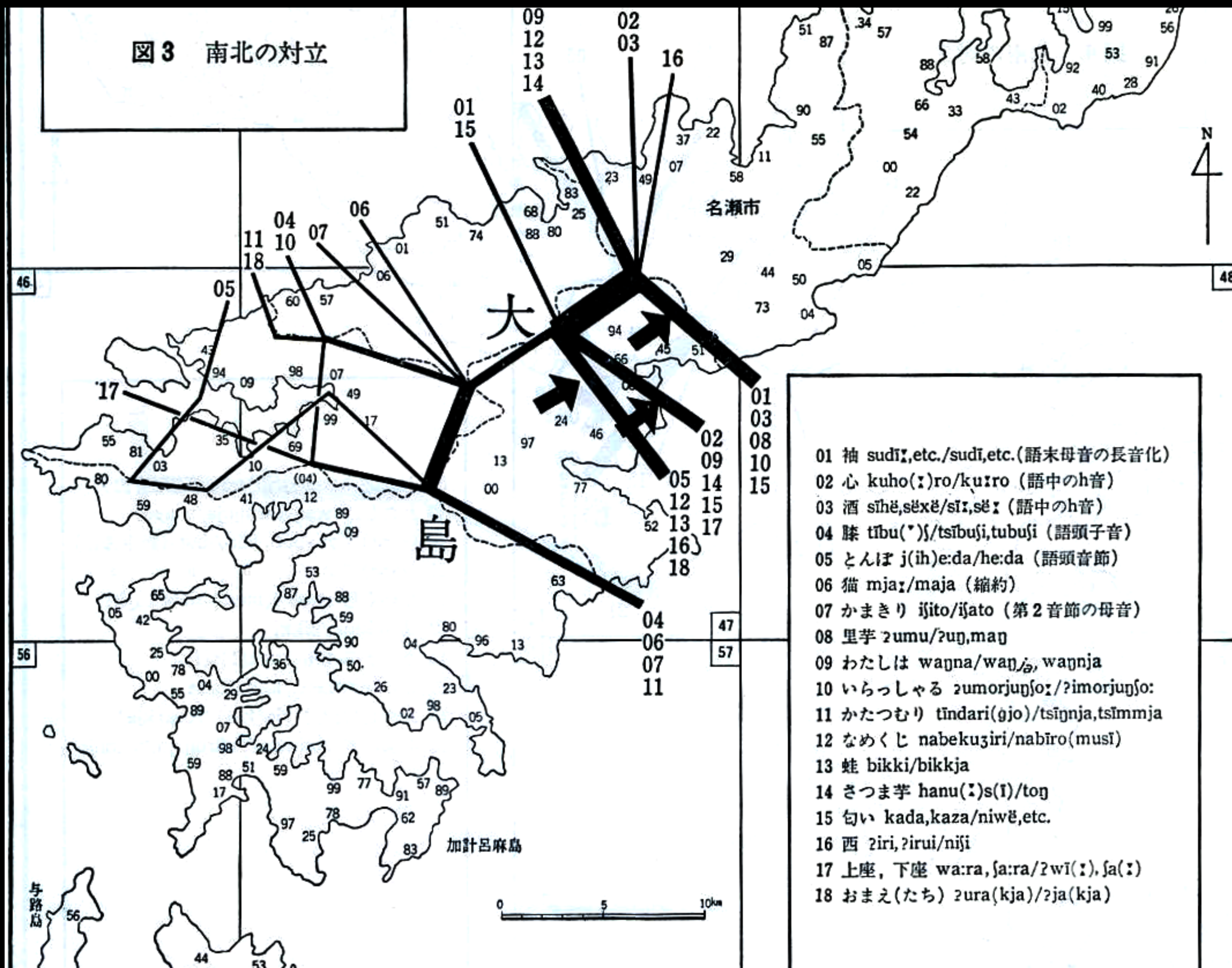


水平伝播を無視するには規模が小さすぎるのでは?



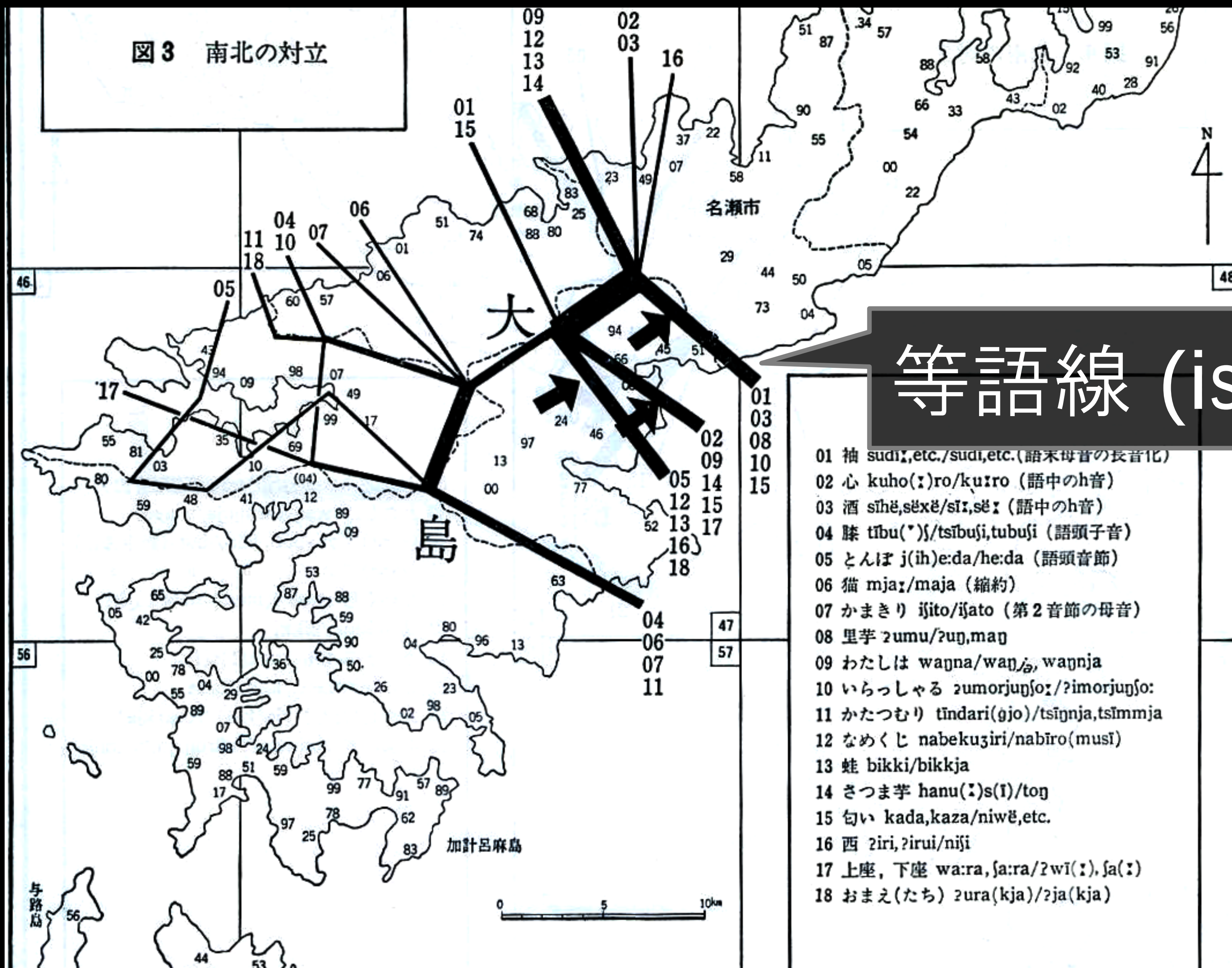
[Gray+, 2011]

方言地理学は空間的説明



[柴田, 1982]

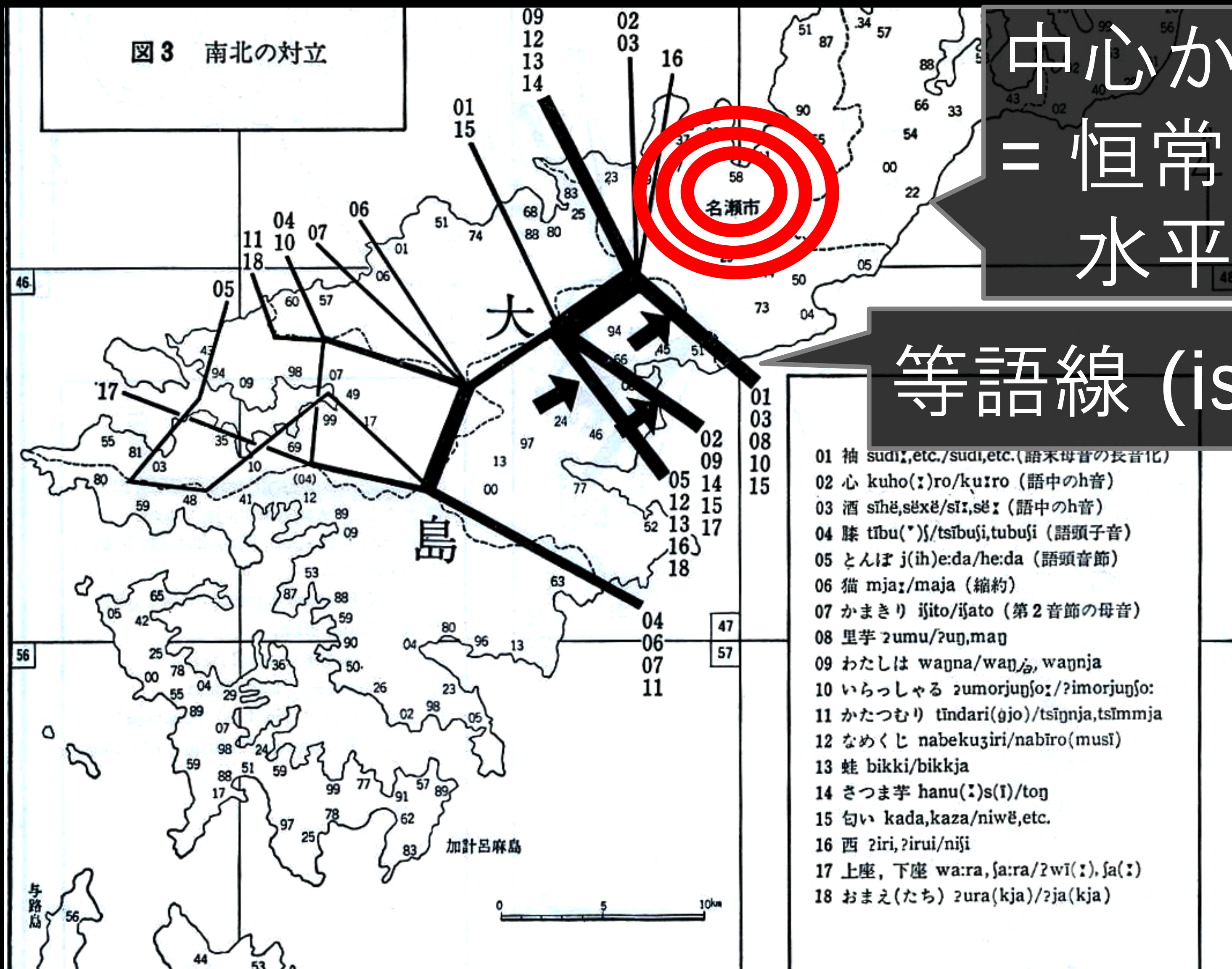
方言地理学は空間的説明



等語線 (isogloss)

[柴田, 1982]

方言地理学は空間的説明



中心からの波
= 恒常的な
水平伝播

等語線 (isogloss)

[柴田, 1982]

時間構造 v. 空間構造

- 系統樹は時間構造のモデル
  - 言語間の類似は祖語の共有による
  - 水平伝播はモデルの仮定に違反
- 方言地理学は空間構造のモデル
  - 言語のネットワーク
  - 恒常的な水平伝播

時間構造 v. 空間構造

- 系統樹は時間構造のモデル
  - 言語間の類似は祖語の共有による
  - 水平伝播はモデルの仮定に違反
- 方言地理学は空間構造のモデル
  - 言語のネットワーク
  - 恒常的な水平伝播

ネットワークは木よりも自由度が高く、推定が大変

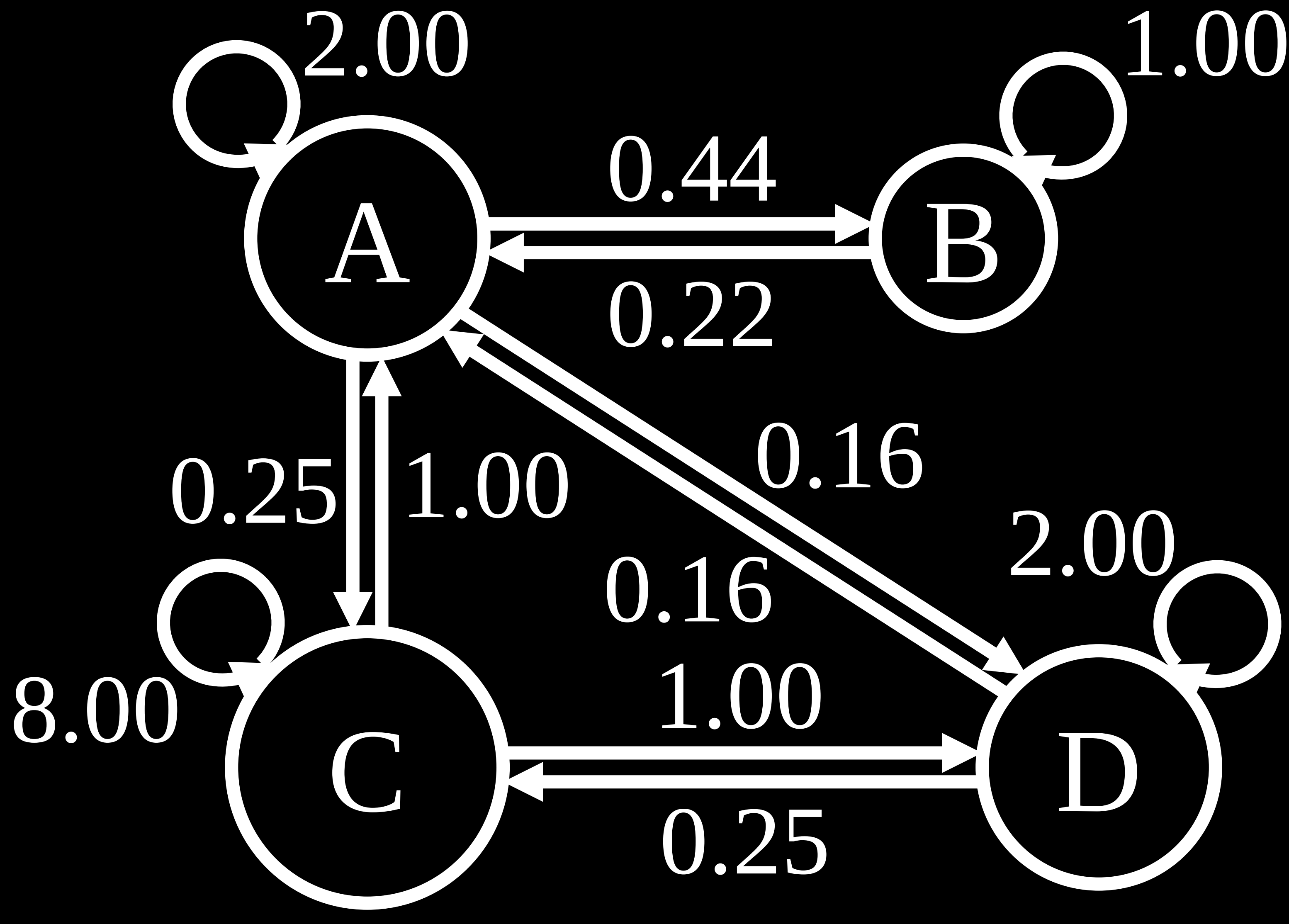
時間構造 v. 空間構造

- 系統樹は時間構造のモデル
  - 言語間の類似は祖語の共有による
  - 水平伝播はモデルの仮定に違反
- 方言地理学は空間構造のモデル
  - 言語のネットワーク
  - 恒常的な水平伝播

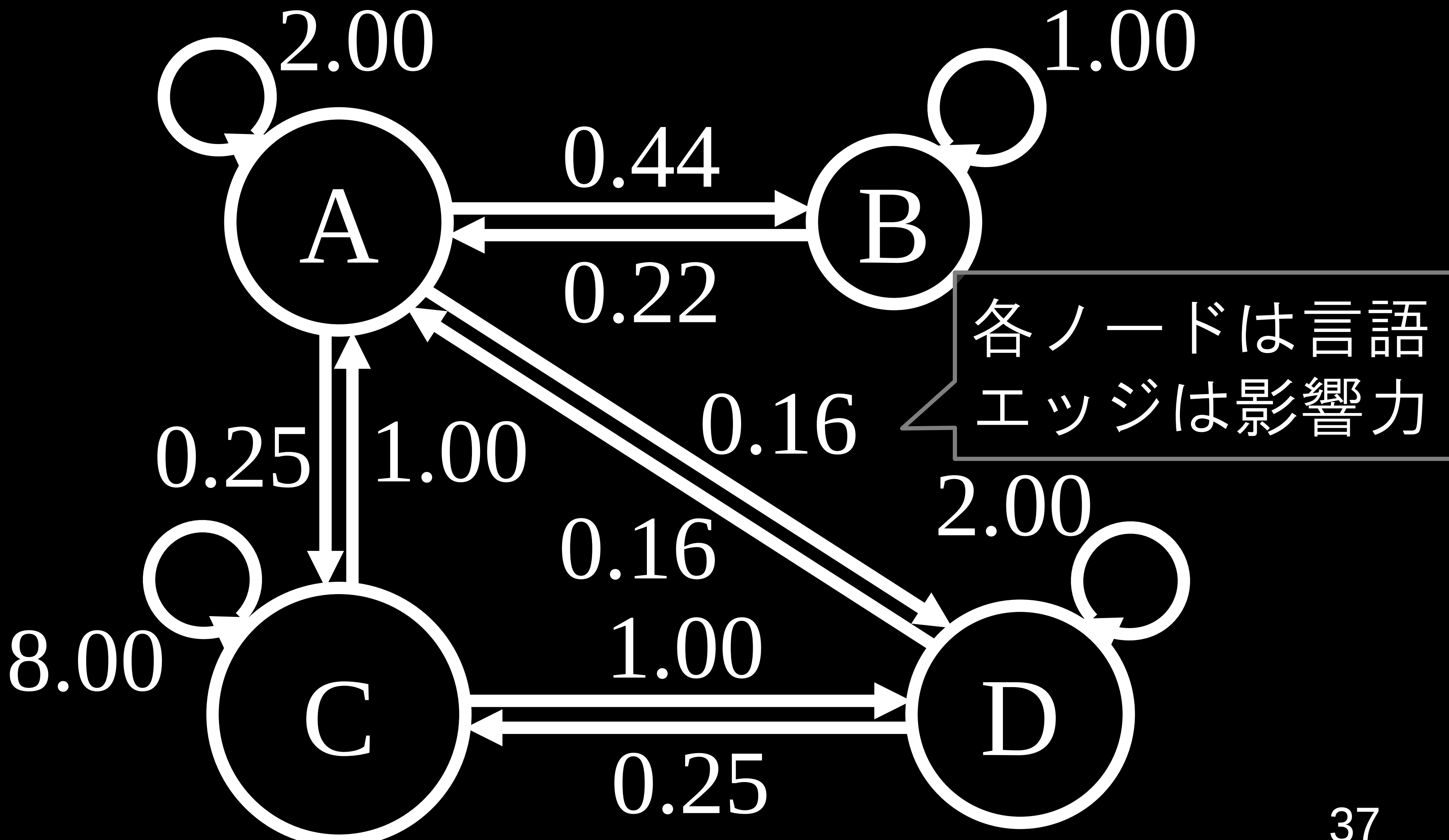
ネットワークは木よりも自由度が高く、推定が大変

とりあえずシミュレーション。結果を系統モデルに与えて実データを再現する

空間構造モデル 1/2

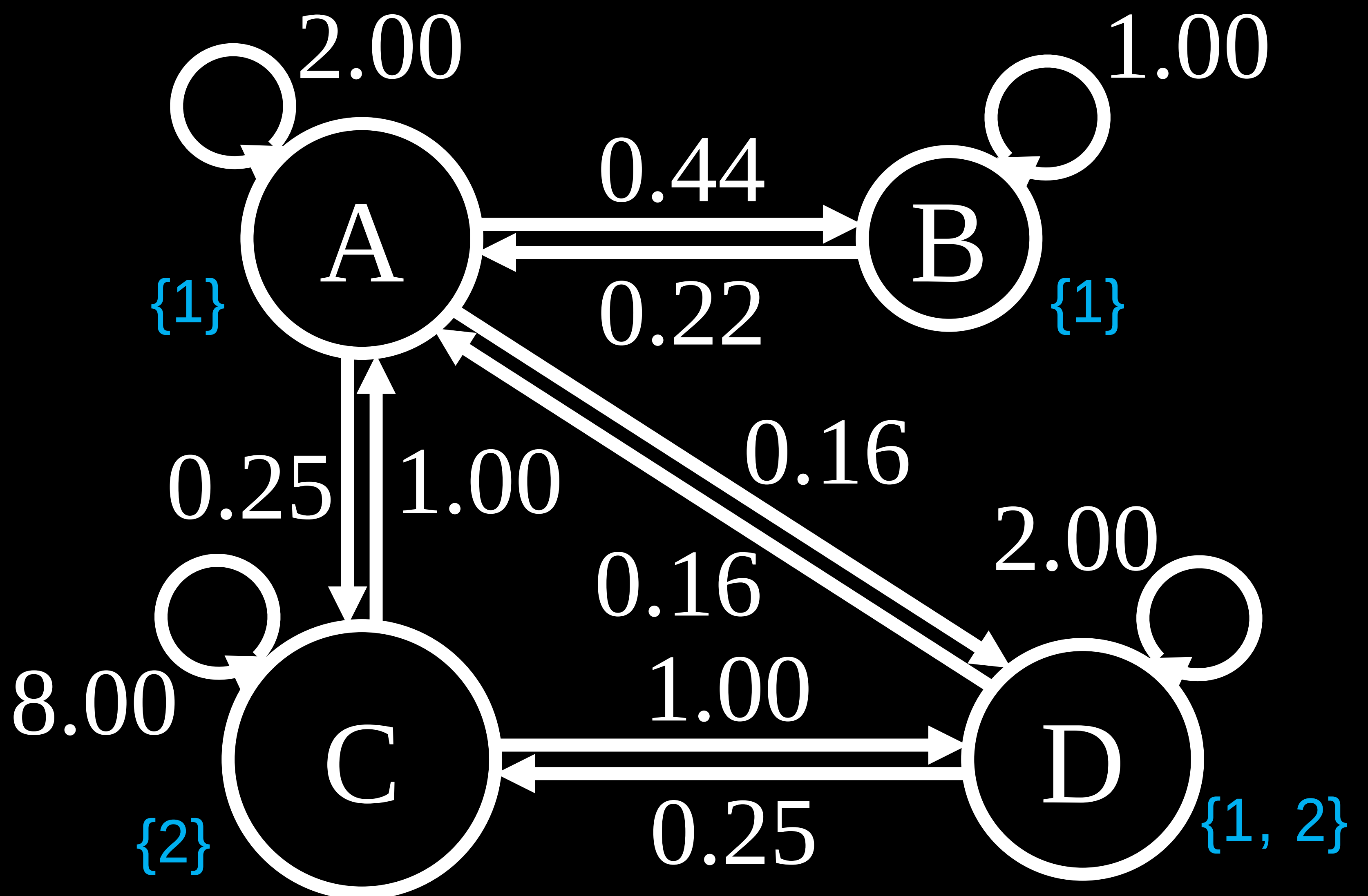


空間構造モデル 1/2

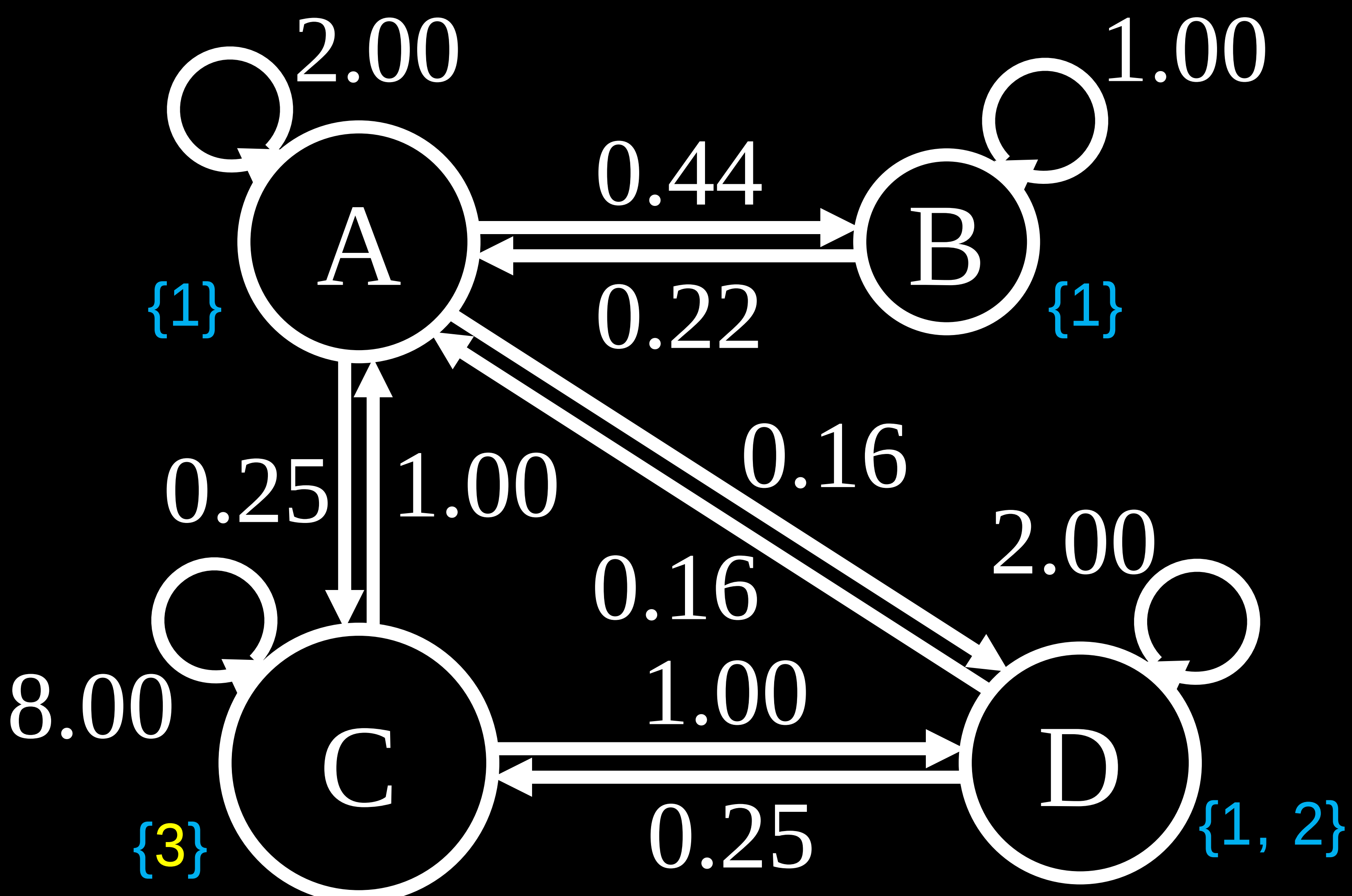

2.00
1.00
0.44
A
B
0.22
各ノードは言語
エッジは影響力
0.25
1.00
0.16
2.00
0.16
8.00
C
1.00
D
0.25

時刻 t

空間構造モデル 2/2

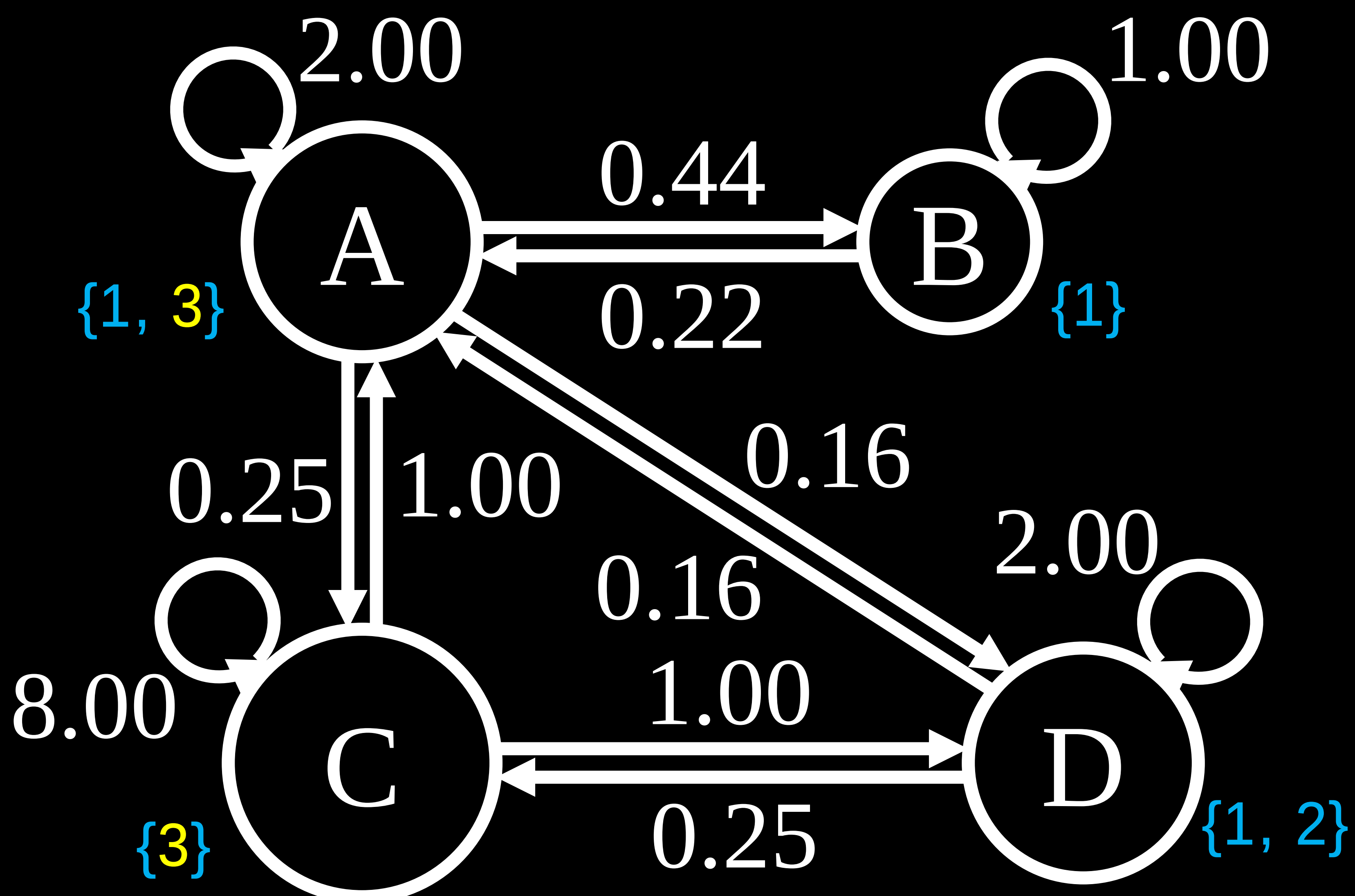


時刻 $t+1$ 空間構造モデル 2/2



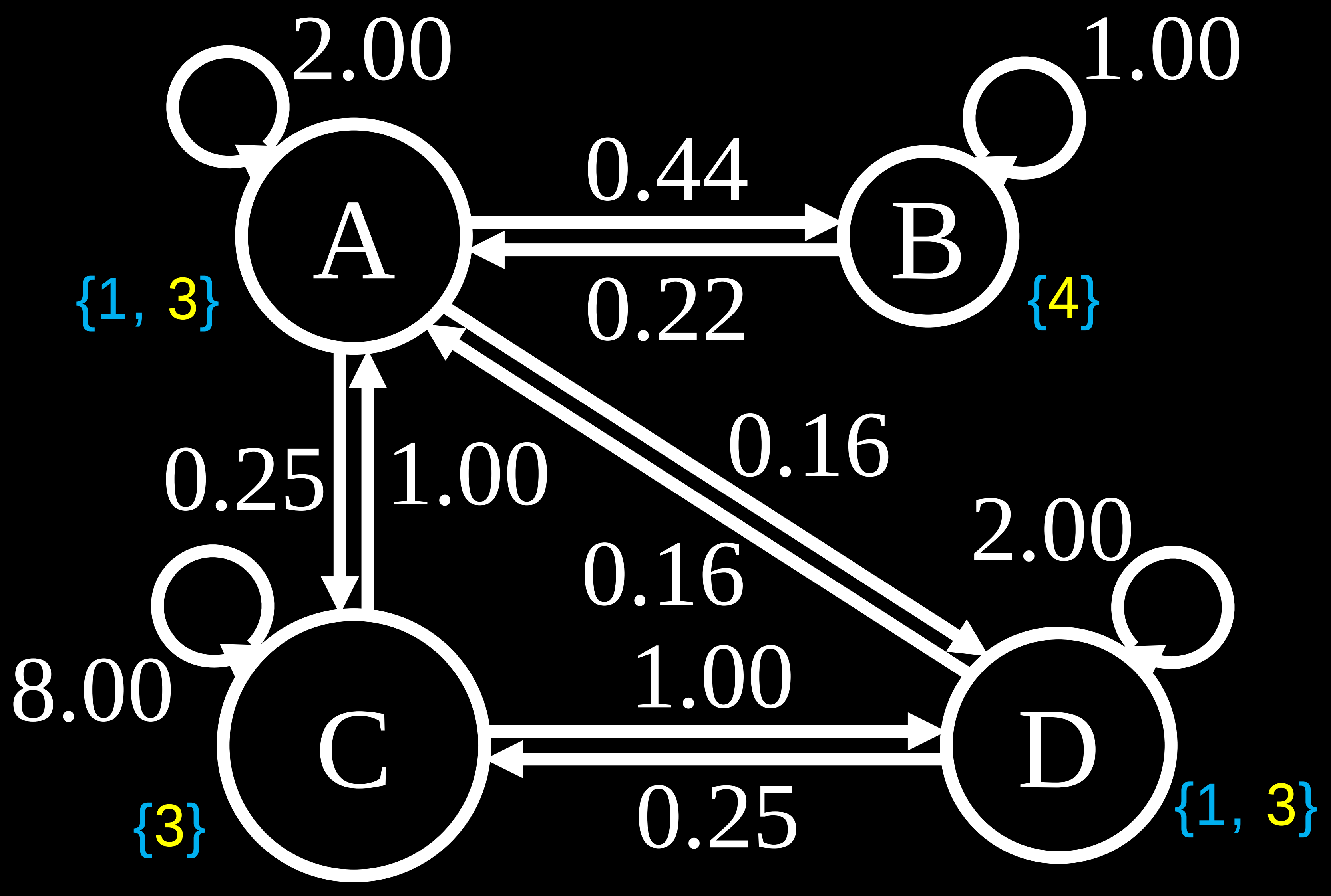
時刻 $t+2$

空間構造モデル 2/2

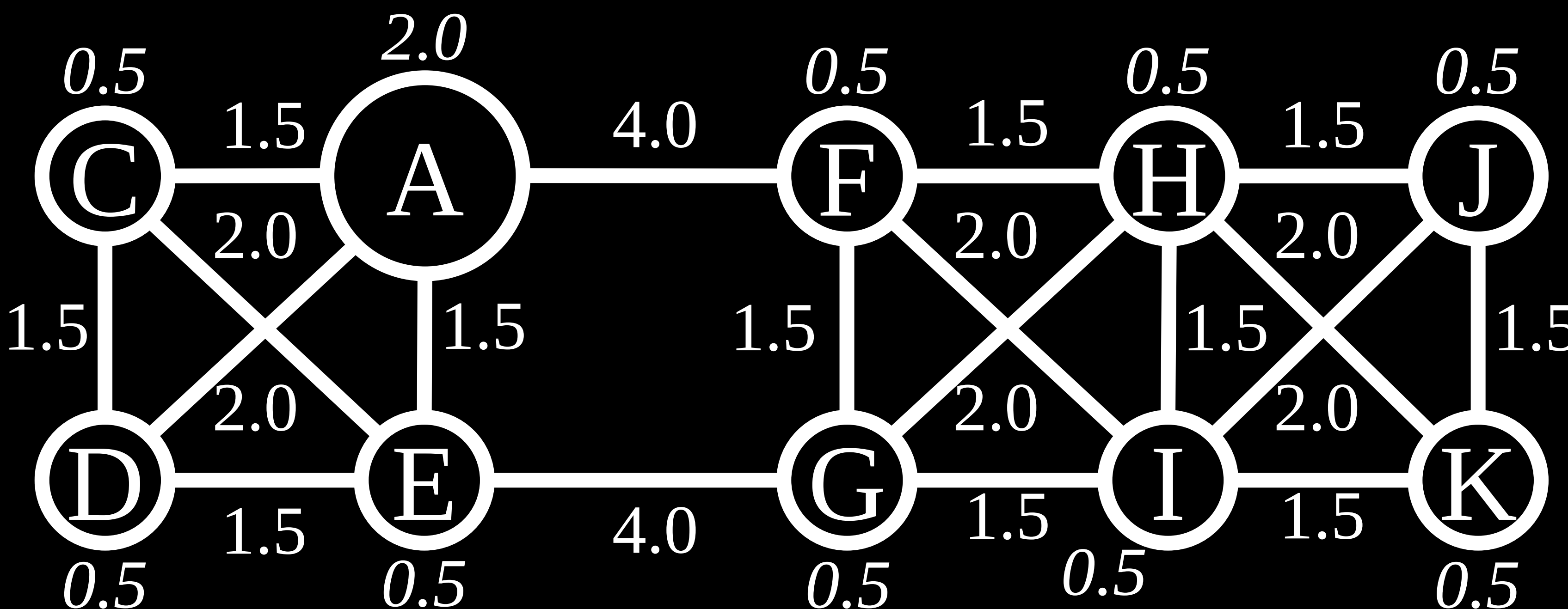


時刻 $t+3$

空間構造モデル 2/2

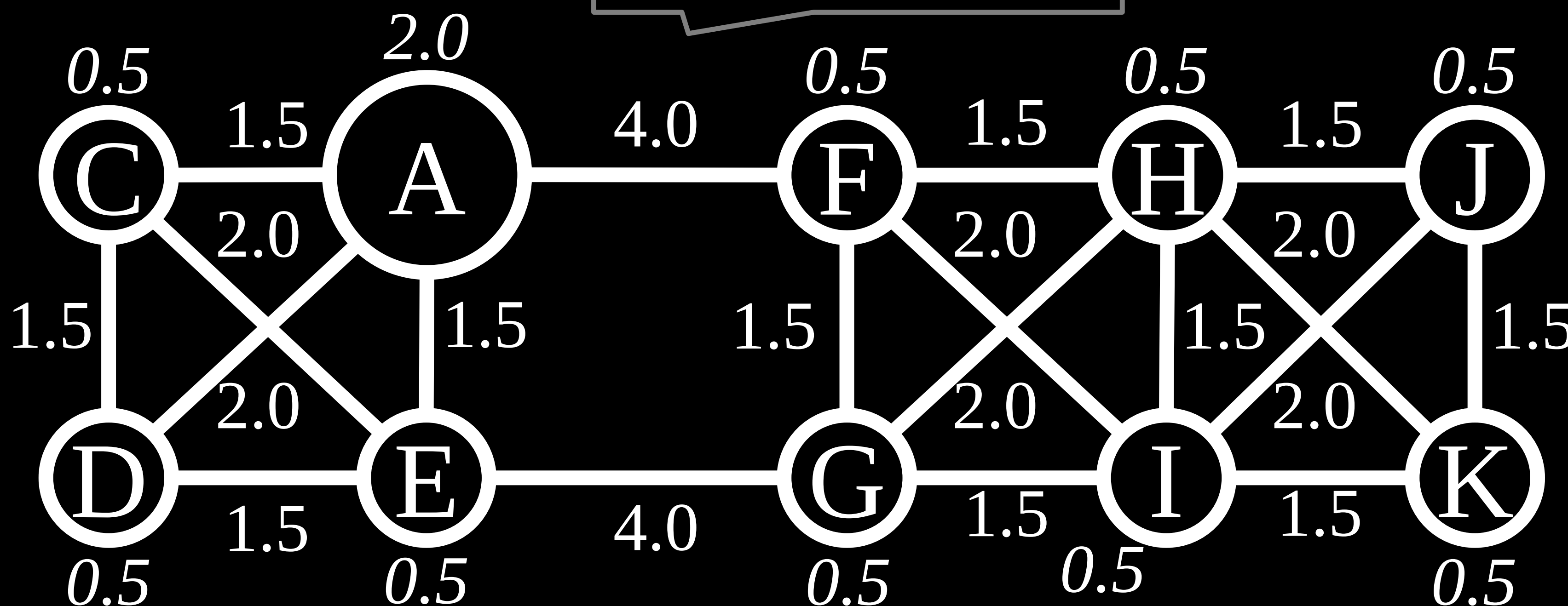


シミュレーション：
殖民型のトポロジー



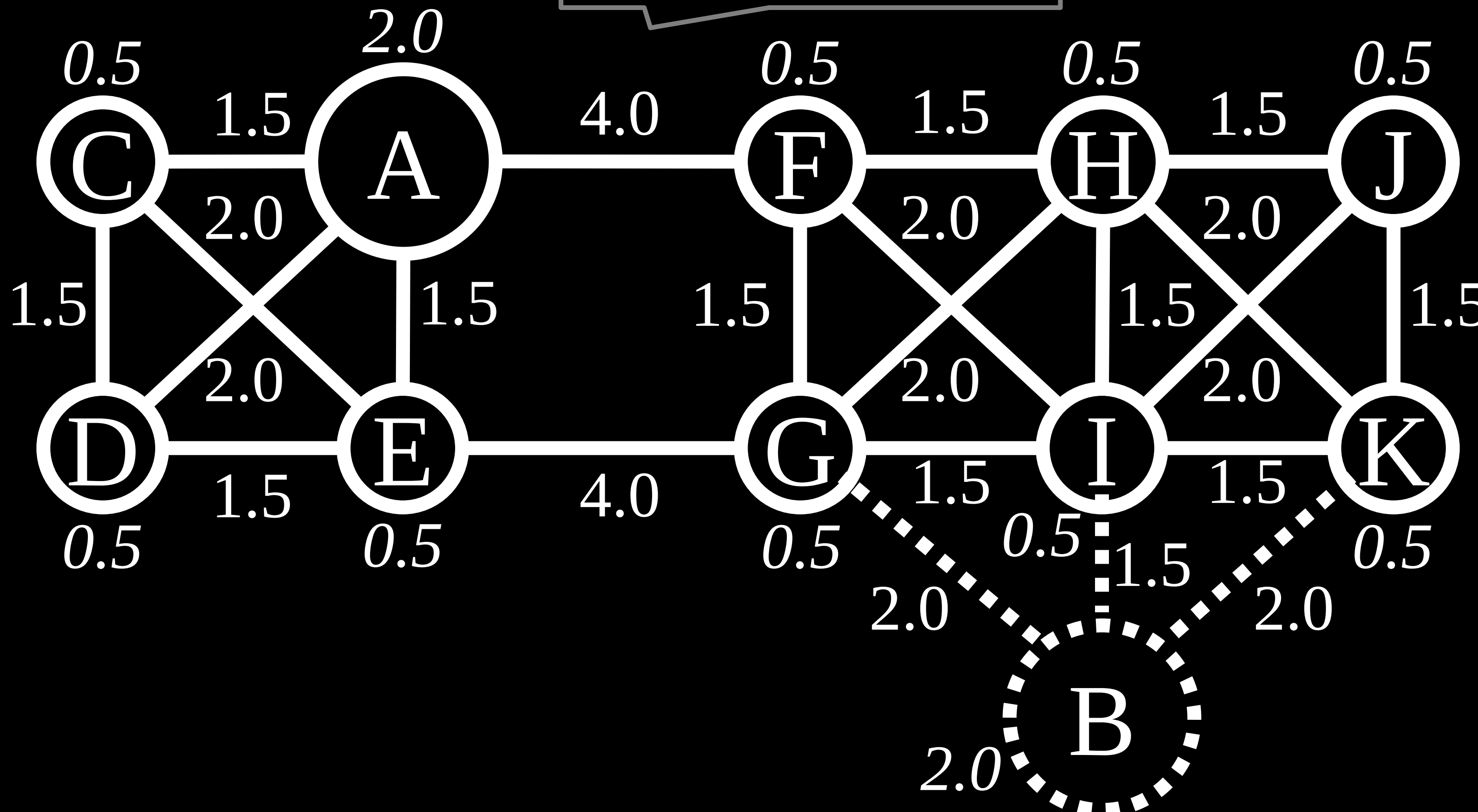
シミュレーション：
殖民型のトポロジー

東西間に隘路



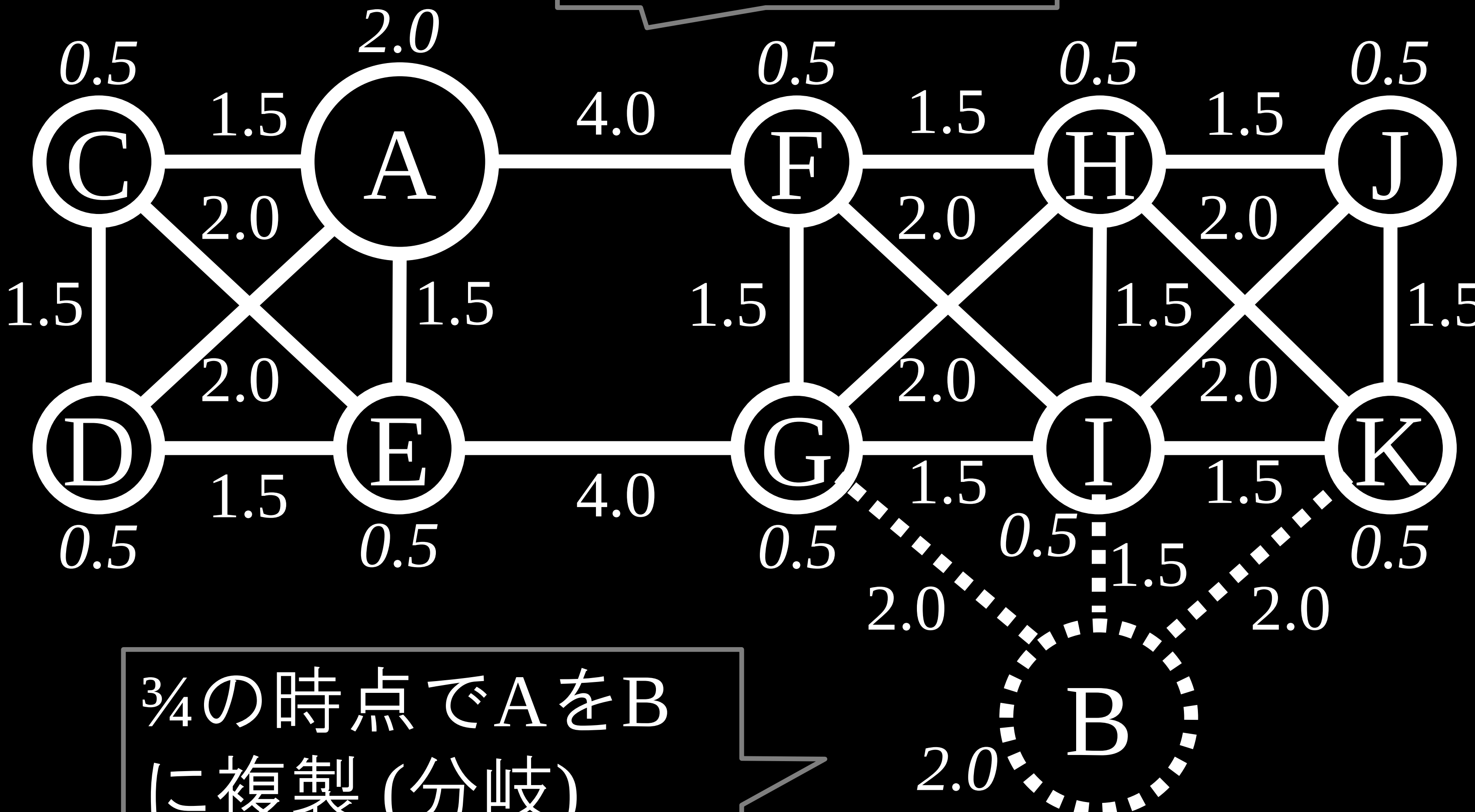
シミュレーション：殖民型のトポロジー

東西間に隘路



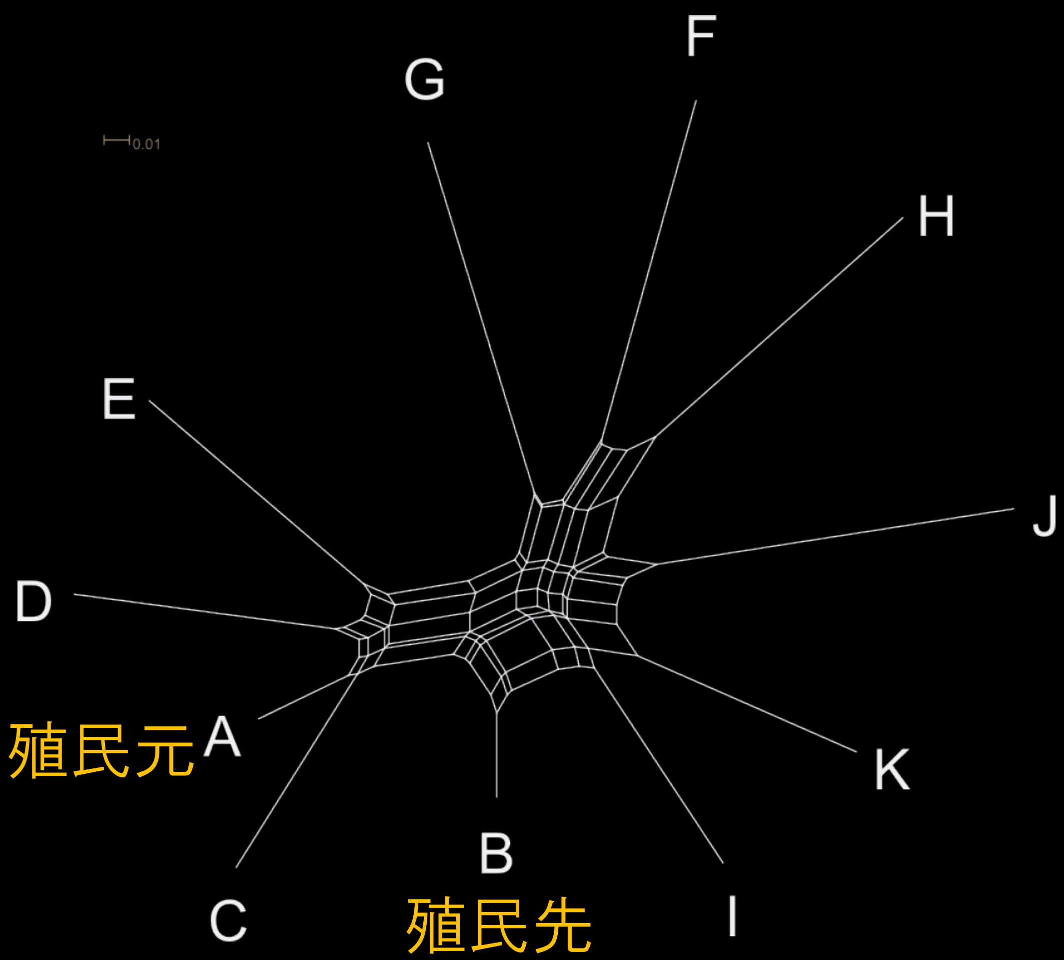
シミュレーション：
殖民型のトポロジー

東西間に隘路



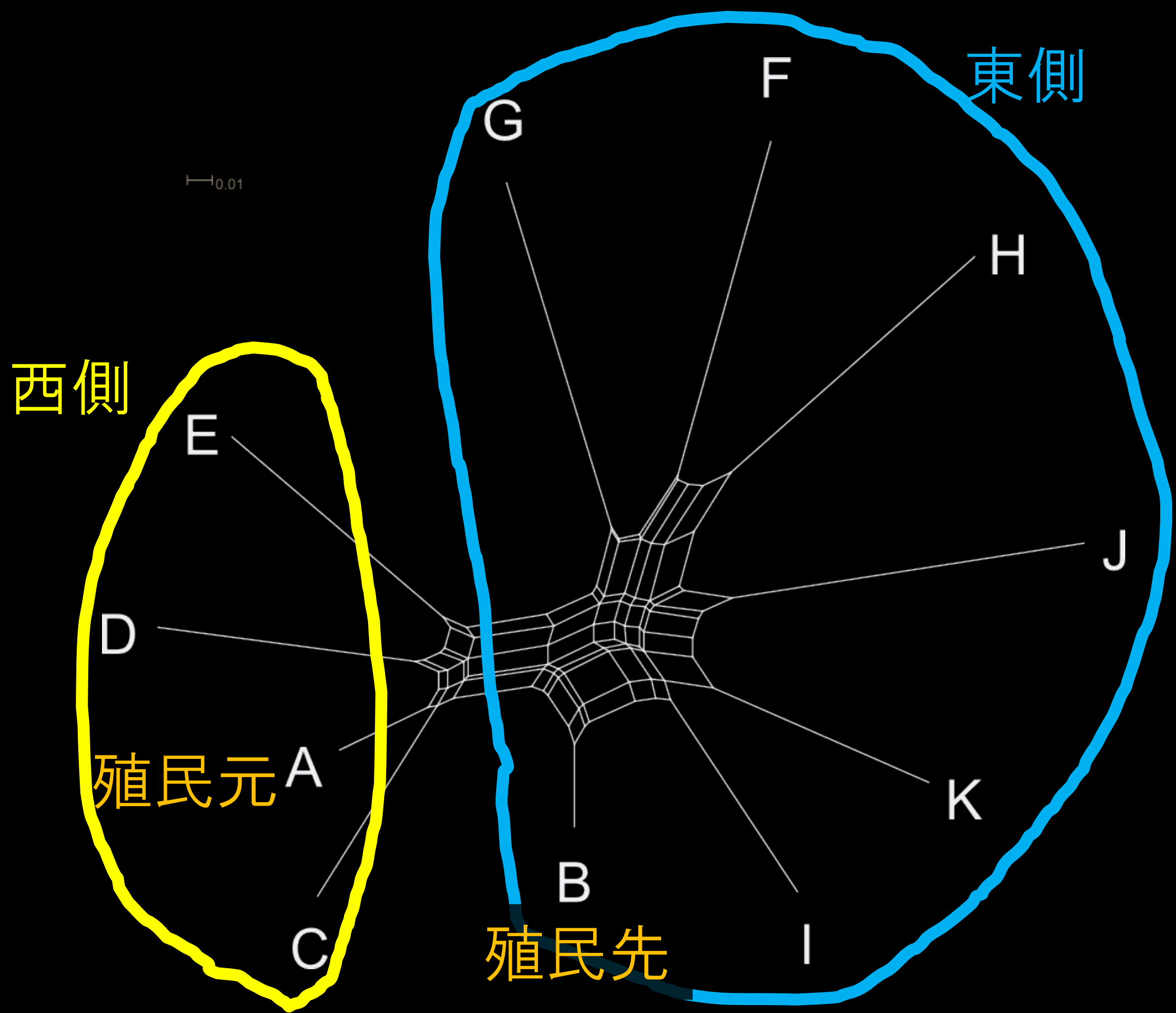
¾の時点でAをB
に複製 (分岐)

殖民型のNeighborNet



読み方:
- 距離に基づくクラスタリング
- 網 (平行四辺形)が矛盾を示す
- 2辺の長さ比が1に近いほど木らしくない

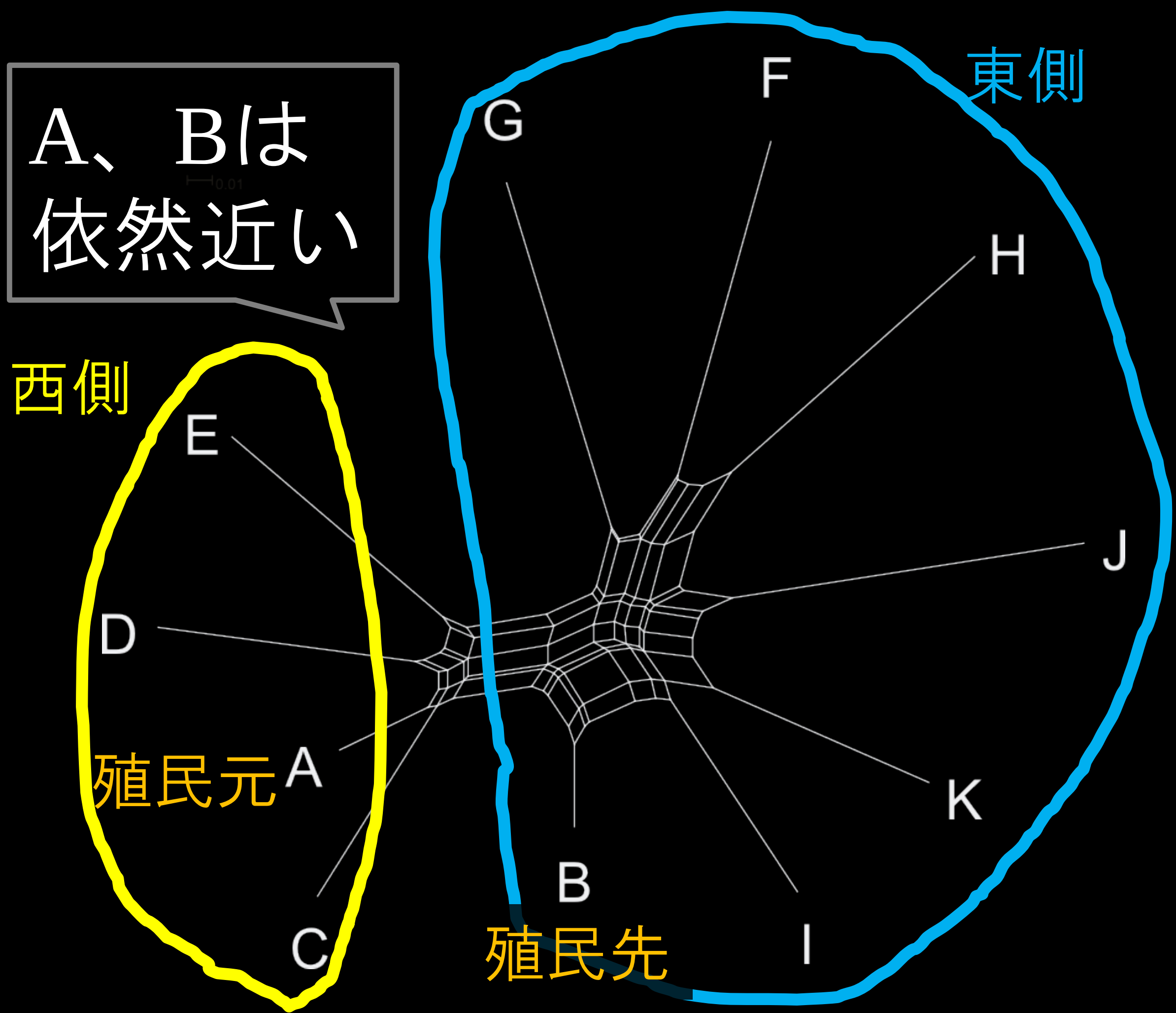
殖民型のNeighborNet



読み方:

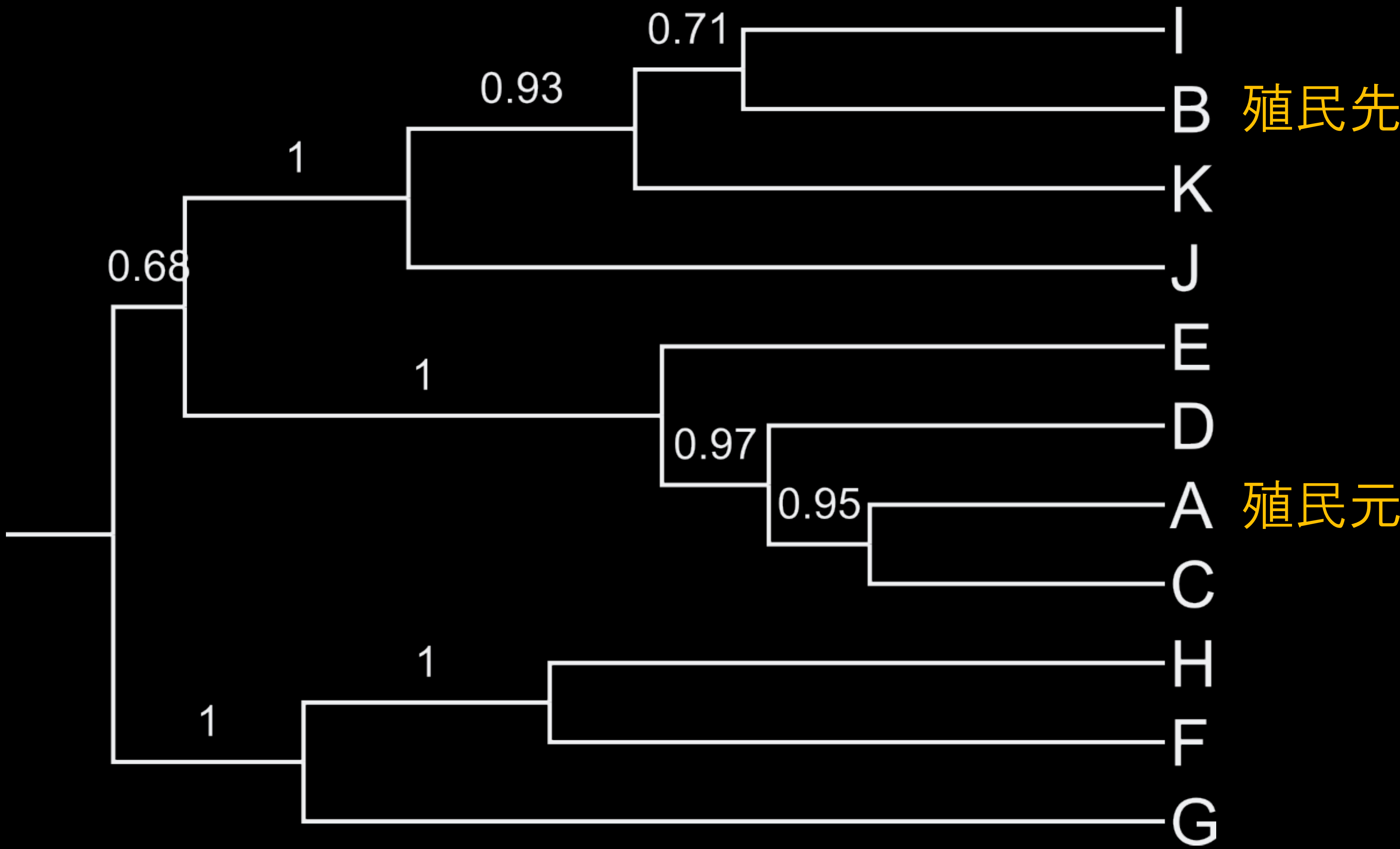
- 距離に基づくクラスタリング
- 網 (平行四辺形)が矛盾を示す
- 2辺の長さ比が1に近いほど木らしくない

殖民型のNeighborNet

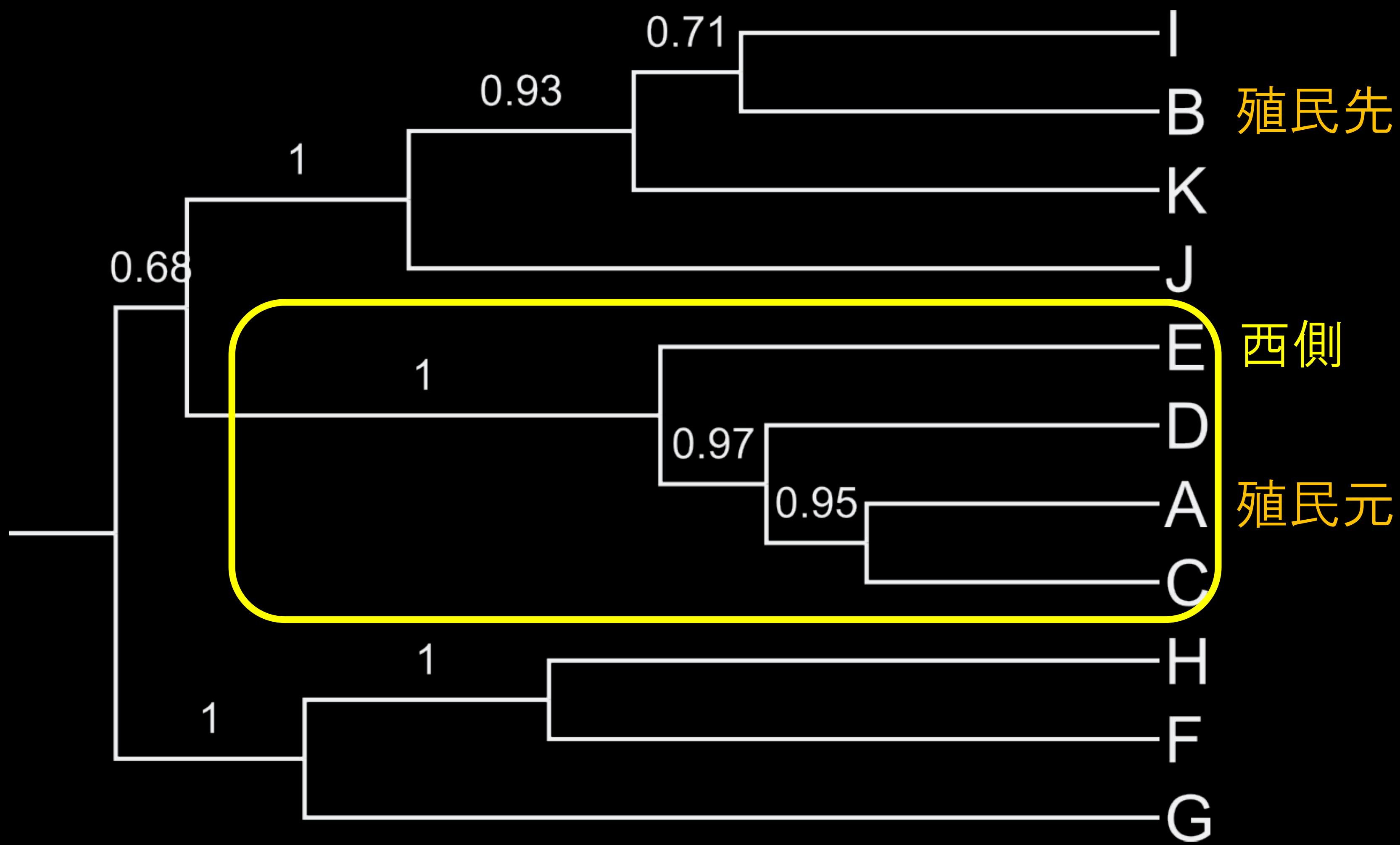


読み方:
- 距離に基づくクラスタリング
- 網 (平行四辺形) が矛盾を示す
- 2辺の長さ比が1に近いほど木らしくない

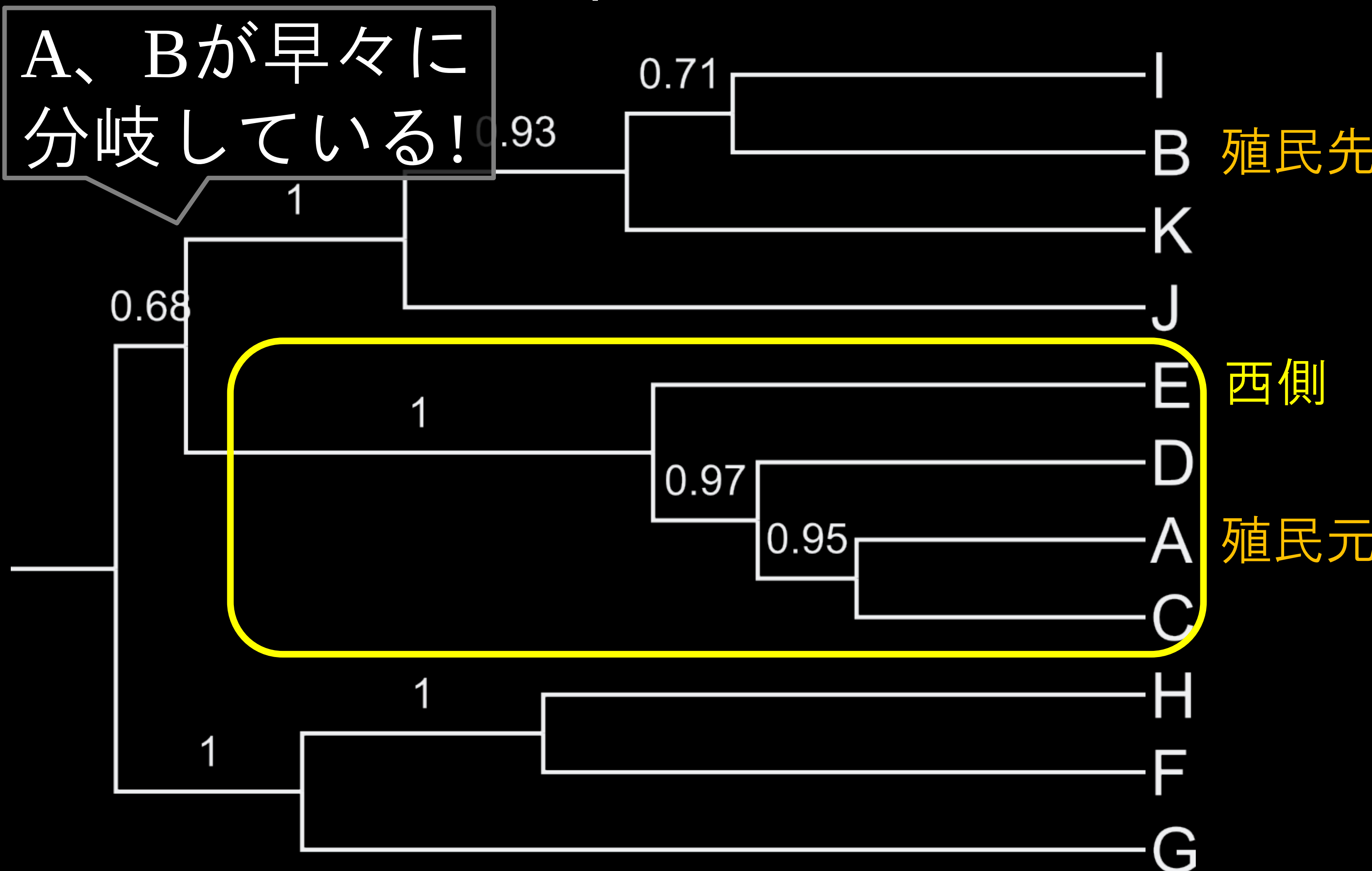
殖民型の推定された系統樹



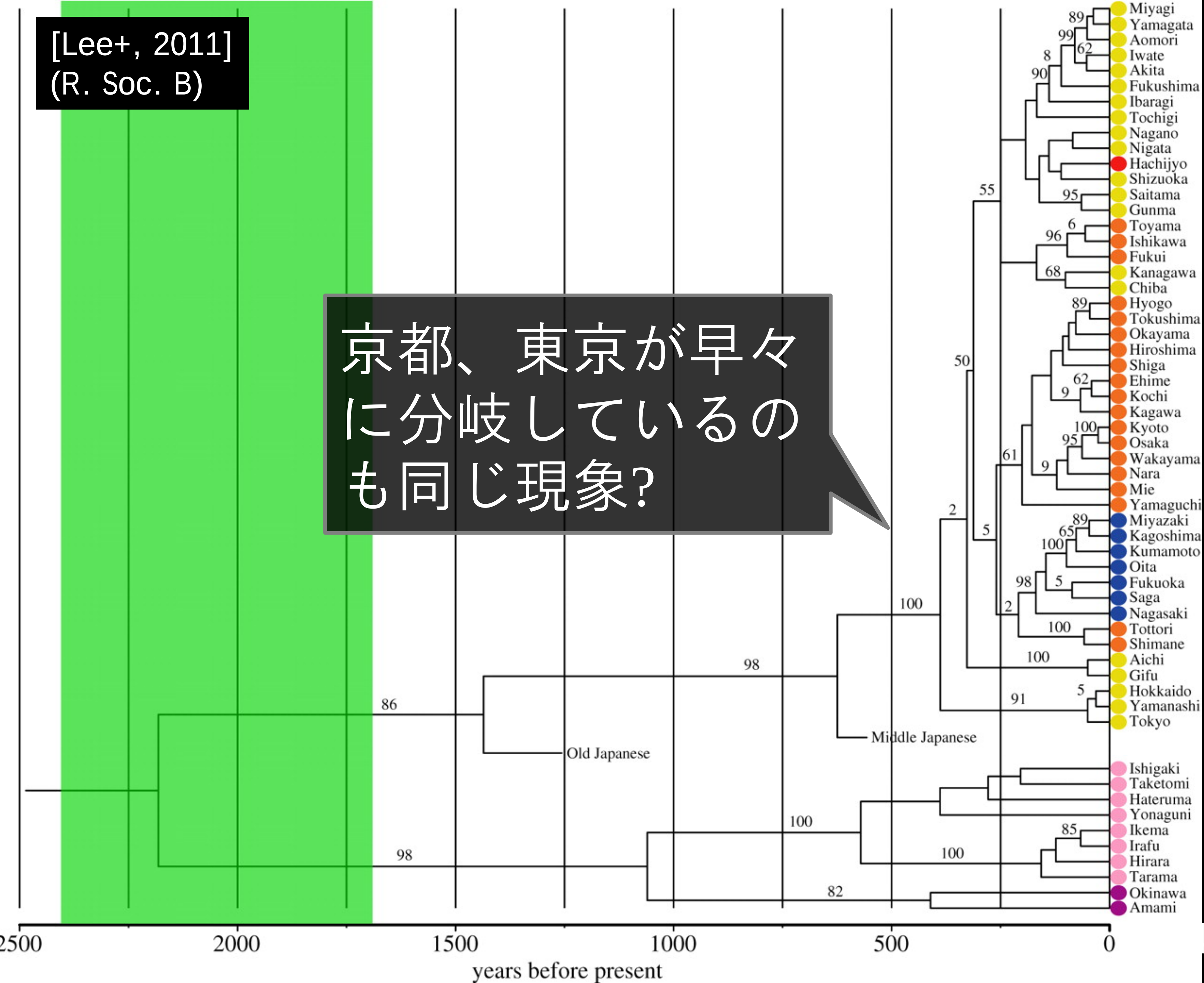
殖民型の推定された系統樹



殖民型の推定された系統樹



[Lee+, 2011]
(R. Soc. B)



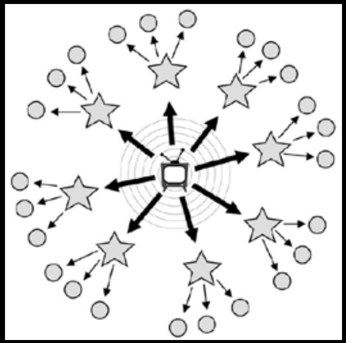
京都、東京が早々に分岐しているのも同じ現象?

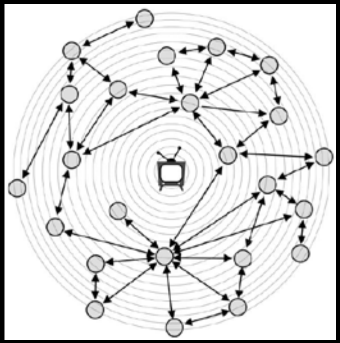
京都

東京

関連研究

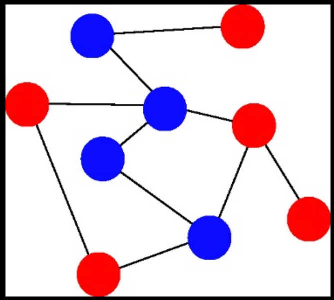
- Word of mouth: 世論形成、マーケティング、技術革新の拡散





[Watts, 2007]

- 進化グラフ



ゲームの利得行列:

$$\begin{array}{ccc} & A & B \\ A & \multicolumn{2}{c}{\multirow{2}{*}{$\begin{pmatrix} a & b \\ c & d \end{pmatrix}$}} \\ B & & \end{array}$$

[Nowak+, 2010]

今日のおはなし

1. 音法則

2. 基礎語彙

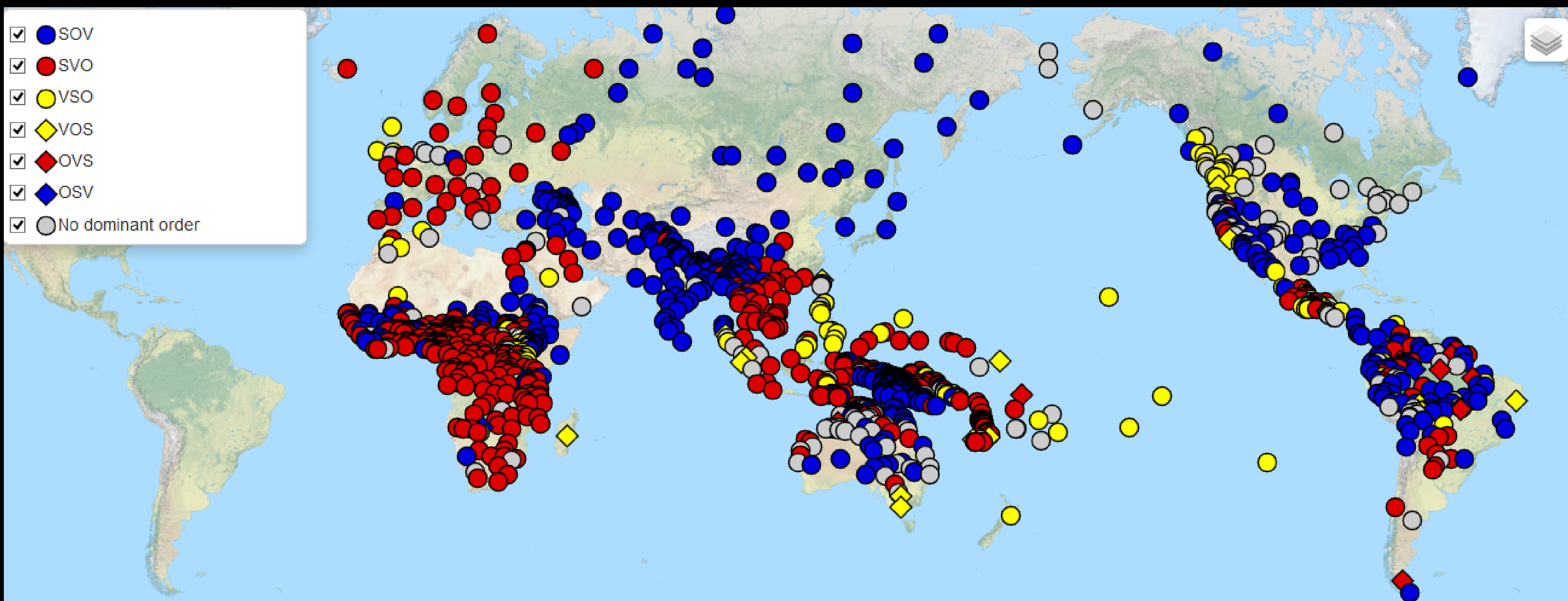
3. 類型論

系統論への応用は少ない
日本語系統論の最後の希望

言語類型論 (Linguistic Typology)

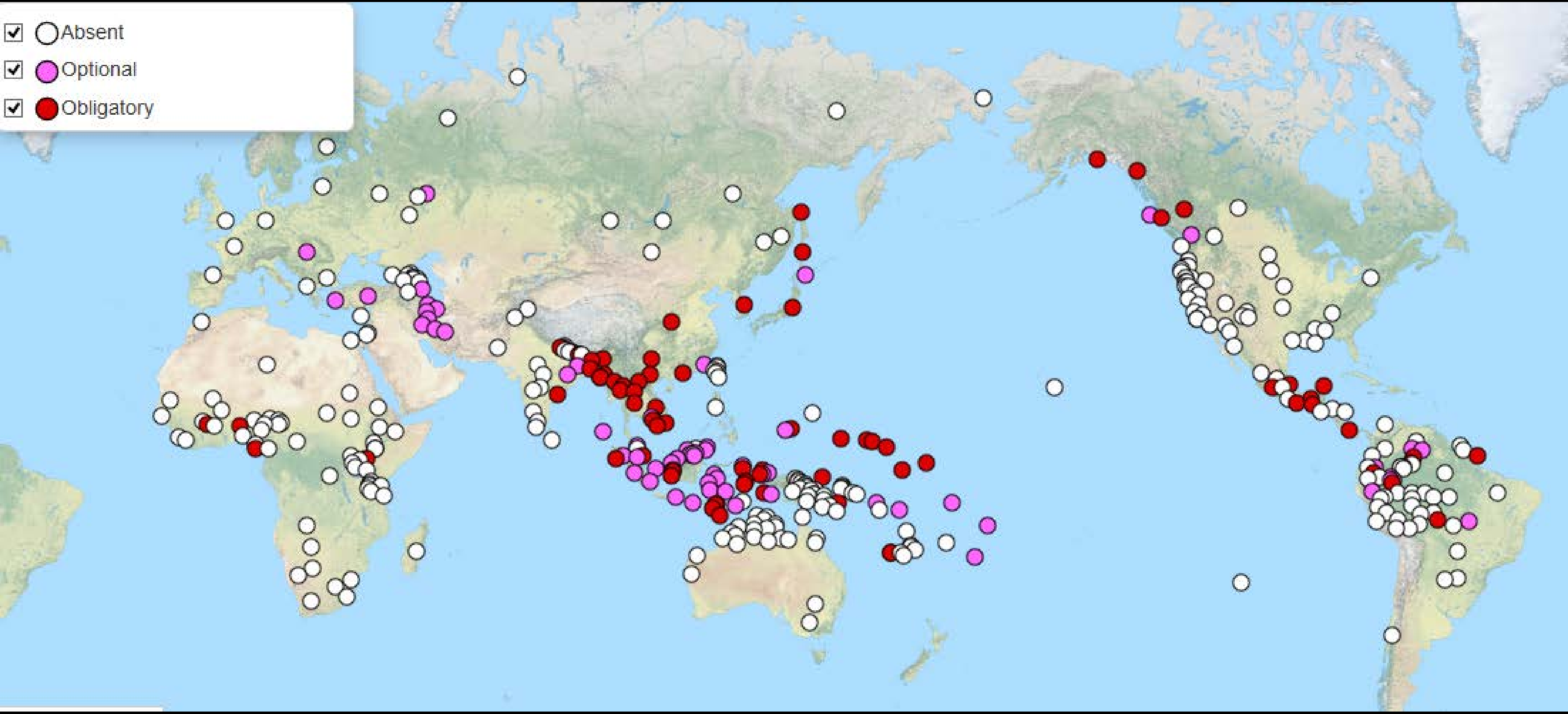
- 世界の言語を類型によって分類
  - 語順、助数詞の有無、声調の有無, etc
- The World Atlas of Language Structures (WALS)
  - 2,679言語
  - 192種類の特徴量
  - ただし言語・特徴量ペアの被覆率は<15%
    - 欠損値推定問題はとりあえず忘れてください

Feature 81A:
Order of Subject, Object and Verb



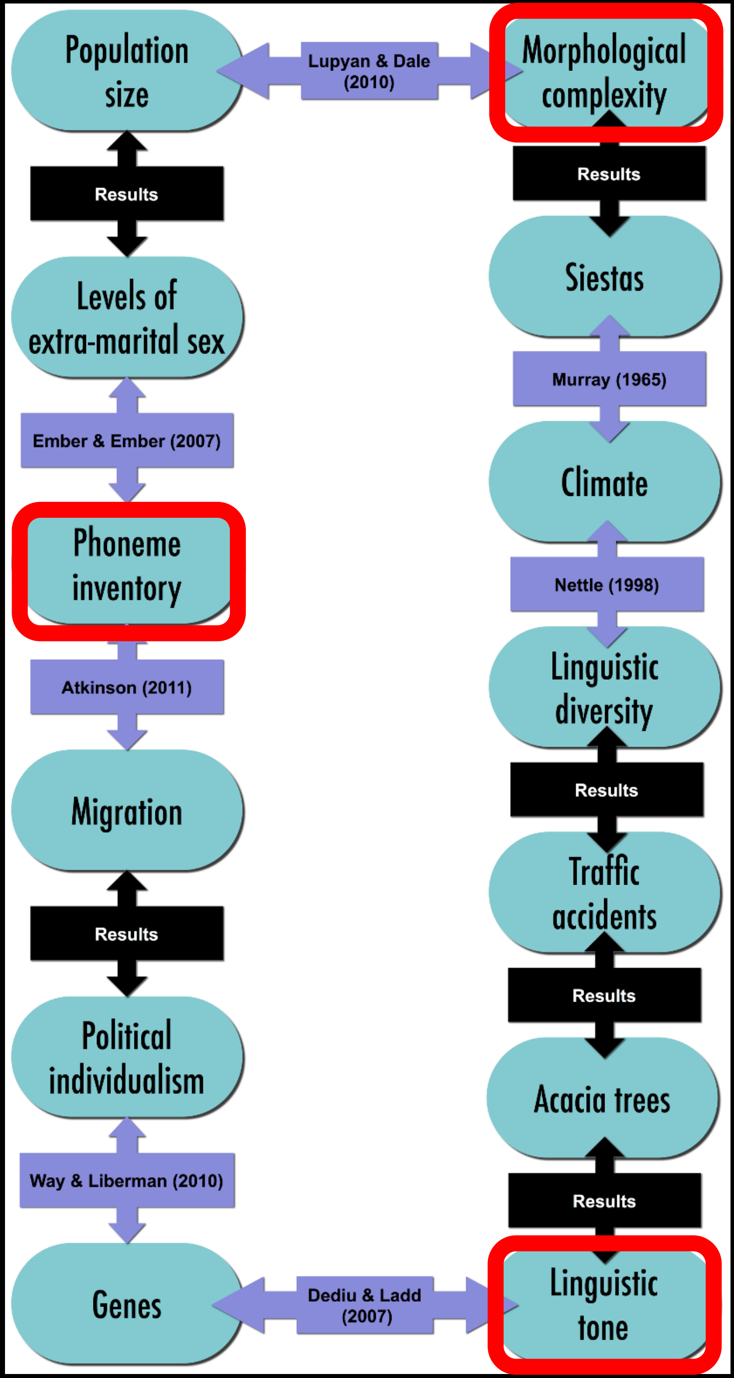
http://wals.info/feature/81A

Feature 55A: Numeral Classifiers (助数詞を使うか)



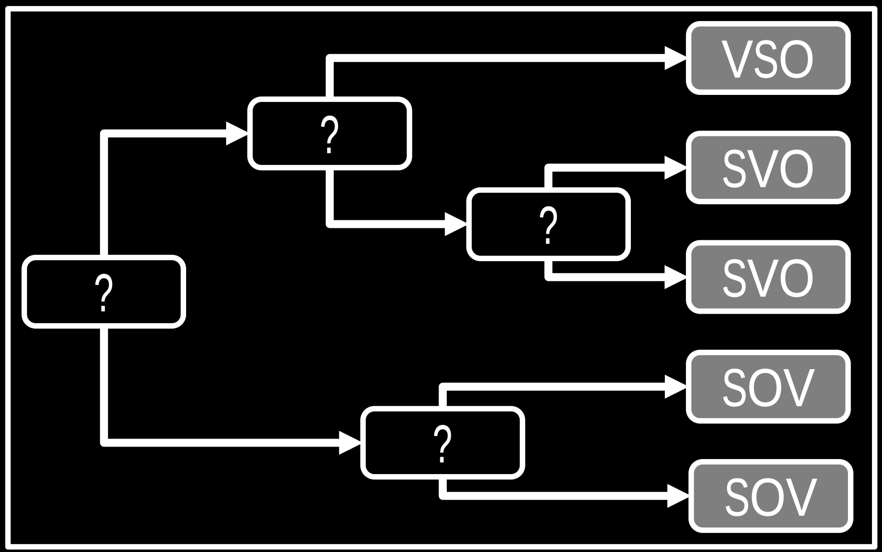
http://wals.info/feature/55A

WALSの応用例: 言語と社会の相関



[Roberts, 2013]

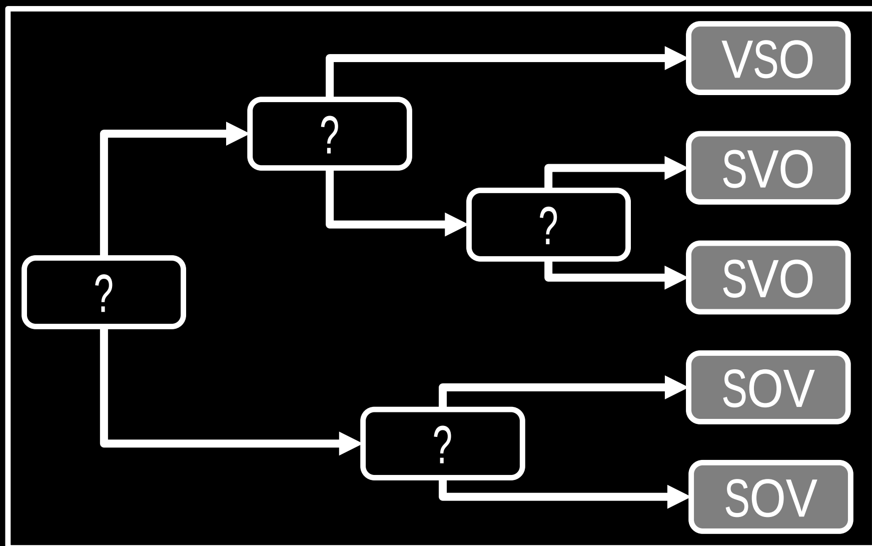
語順変化確率の推定 [Maurits+, PNAS 2014]

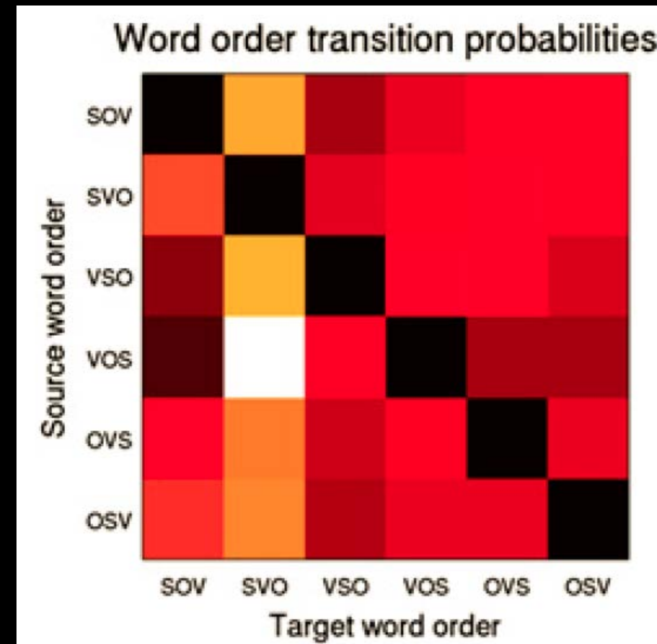


＋

$$Q = \begin{array}{c} \text{SOV} \\ \text{SVO} \\ \text{VSO} \\ \text{VOS} \\ \text{OVS} \\ \text{OSV} \\ \end{array} \begin{array}{c} \begin{pmatrix} -\lambda_1 & \lambda_1\alpha_{12} & \lambda_1\alpha_{13} & \lambda_1\alpha_{14} & \lambda_1\alpha_{15} & \lambda_1\alpha_{16} \\ \lambda_2\alpha_{21} & -\lambda_2 & \lambda_2\alpha_{23} & \lambda_2\alpha_{24} & \lambda_2\alpha_{25} & \lambda_2\alpha_{26} \\ \lambda_3\alpha_{31} & \lambda_3\alpha_{32} & -\lambda_3 & \lambda_3\alpha_{34} & \lambda_3\alpha_{35} & \lambda_3\alpha_{36} \\ \lambda_4\alpha_{41} & \lambda_4\alpha_{42} & \lambda_4\alpha_{43} & -\lambda_4 & \lambda_4\alpha_{45} & \lambda_4\alpha_{46} \\ \lambda_5\alpha_{51} & \lambda_5\alpha_{52} & \lambda_5\alpha_{53} & \lambda_5\alpha_{54} & -\lambda_5 & \lambda_5\alpha_{56} \\ \lambda_6\alpha_{61} & \lambda_6\alpha_{62} & \lambda_6\alpha_{63} & \lambda_6\alpha_{64} & \lambda_6\alpha_{65} & -\lambda_6 \end{pmatrix} \\ \begin{array}{cccccc} \text{SOV} & \text{SVO} & \text{VSO} & \text{VOS} & \text{OVS} & \text{OSV} \end{array} \end{array}$$

語順変化確率の推定 [Maurits+, PNAS 2014]





＋

$$
Q = \begin{array}{c}
\text{SOV}\\ \text{SVO}\\ \text{VSO}\\ \text{VOS}\\ \text{OVS}\\ \text{OSV}\\ \\
\end{array}
\begin{array}{c}
\begin{pmatrix}
-\lambda_1 & \lambda_1\alpha_{12} & \lambda_1\alpha_{13} & \lambda_1\alpha_{14} & \lambda_1\alpha_{15} & \lambda_1\alpha_{16} \\
\lambda_2\alpha_{21} & -\lambda_2 & \lambda_2\alpha_{23} & \lambda_2\alpha_{24} & \lambda_2\alpha_{25} & \lambda_2\alpha_{26} \\
\lambda_3\alpha_{31} & \lambda_3\alpha_{32} & -\lambda_3 & \lambda_3\alpha_{34} & \lambda_3\alpha_{35} & \lambda_3\alpha_{36} \\
\lambda_4\alpha_{41} & \lambda_4\alpha_{42} & \lambda_4\alpha_{43} & -\lambda_4 & \lambda_4\alpha_{45} & \lambda_4\alpha_{46} \\
\lambda_5\alpha_{51} & \lambda_5\alpha_{52} & \lambda_5\alpha_{53} & \lambda_5\alpha_{54} & -\lambda_5 & \lambda_5\alpha_{56} \\
\lambda_6\alpha_{61} & \lambda_6\alpha_{62} & \lambda_6\alpha_{63} & \lambda_6\alpha_{64} & \lambda_6\alpha_{65} & -\lambda_6
\end{pmatrix} \\
\begin{array}{cccccc}
\text{SOV} & \text{SVO} & \text{VSO} & \text{VOS} & \text{OVS} & \text{OSV}
\end{array}
\end{array}
$$

SOVからSVOへの変化の方がその反対より起こりやすい

類型論は日本語系統論最後の希望

[Janhunen, 2003]

A Framework for the Study of Japanese Language Origins

Juha JANHUNEN

University of Helsinki

Keywords: Japonic, Korean, Ainu; Jomon, Yayoi, Paekche; linguistic expansions, areal contacts, typological change.

10. *Japonic belongs to the Altaic type*. In contrast to the failure of the genetic comparisons, it is impossible to deny the fact that Japonic is typologically linked with the continental languages traditionally identified as Altaic (or Ural-Altaic). The Altaic features of Japonic cover most aspects of the language, including segmental structure, morphology, morphosyntax, and syntax. In view of these features, Japanese may well be called an Altaic language, but only in the typological sense. The same is true of the other "Altaic" languages, which, in spite of their shared typology, seem to represent separate genetic lineages.

22. *Japonic had originally a non-Altaic typology*. The application of the method of internal reconstruction to Japonic linguistic material suggests that Pre-Proto-Japonic may originally have been characterized by a non-Altaic typology. The most important non-Altaic features of Pre-Proto-Japonic seem to have been its predominantly monosyllabic morpheme structure and the presence of tonal distinctions (Janhunen 1997).

25. *Japonic had once a Sinitic typology*. Moving further along these lines, it can be assumed that the non-Altaic features in the typology of Japonic, including both the tonal

類型論は日本語系統論最後の希望

[Janhunen, 2003]

A Framework for the Study of Japanese Language Origins

Juha JANHUNEN

University of Helsinki

Keywords：Japonic, Korean, Ainu; Jomon, Yayoi, Paekche; linguistic expansions, areal contacts, typological change.

10. *Japonic belongs to the Altaic type*. In contrast to the failure of the genetic comparisons, it is impossible to deny the fact that Japonic is typologically linked with the continental languages traditionally identified as Altaic (or Ural-Altaic). The Altaic features of Japonic cover most aspects of the language, including segmental structure, morphology, morphosyntax, and syntax. In view of these features, Japanese may well be called an Altaic language, but only in the typological sense. The same is true of the other "Altaic" languages, which, in spite of their shared typology, seem to represent separate genetic lineages.

22. *Japonic had originally a non-Altaic typology*. The application of the method of internal reconstruction to Japonic linguistic material suggests that Pre-Proto-Japonic may originally have been characterized by a non-Altaic typology. The most important non-Altaic features of Pre-Proto-Japonic seem to have been its predominantly monosyllabic morpheme structure and the presence of tonal distinctions (Janhunen 1997).

25. *Japonic had once a Sinitic typology*. Moving further along these lines, it can be assumed that the non-Altaic features in the typology of Japonic, including both the tonal

類型論は日本語系統論最後の希望

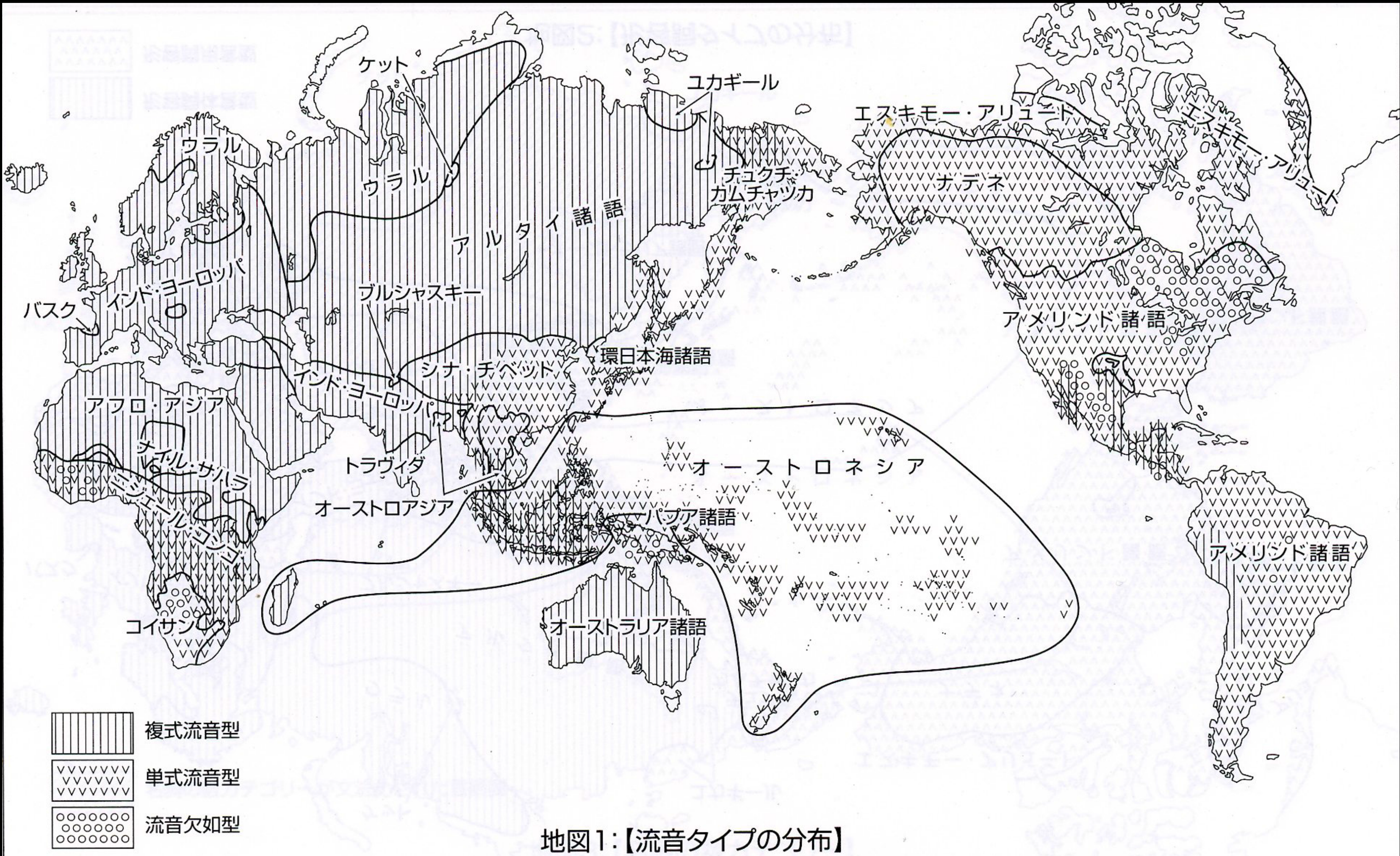
- 任意の言語が比較できる

日本語:	1	1	2	…	0	4
朝鮮語:	2	2	1	…	0	4
アイヌ語:	0	1	1	…	0	4

- 類型論の変化は長期的
  - 語順の変化は多くても2,000年に1回程度
    - cf. 基礎語彙の残存率は1,000年で~81%
  - ただし、不確実性が高く扱いが難しそう

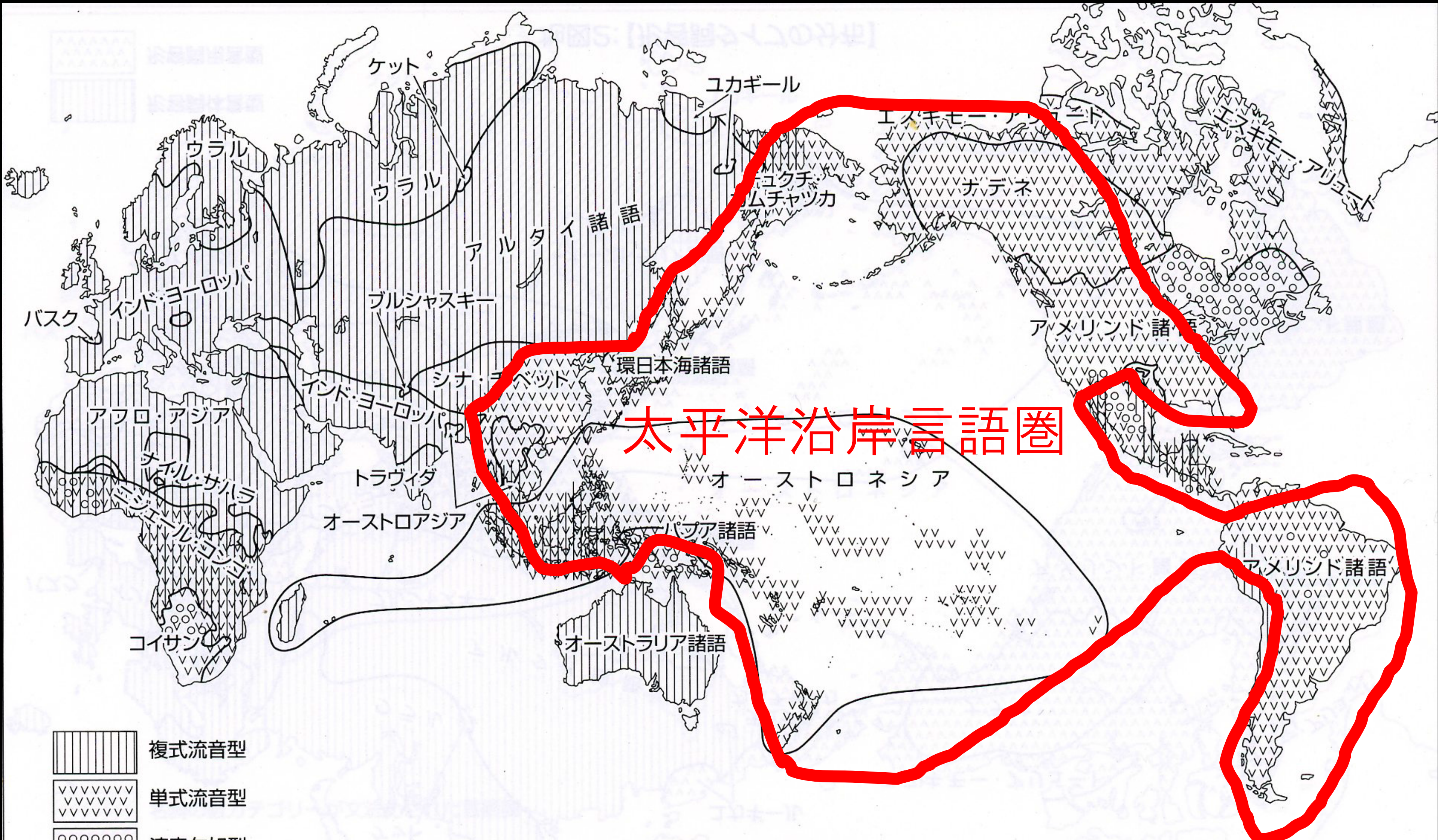
通時類型論 (Diachronic Typology) の従来研究

- 特徴量の変化を諸言語から調査
- 言語連合: 系統に反する特徴量の変化
  [Trubetzkoy, 1923][Aikhenvald+, 2001][Daumé III, NAACL 2009]
- 特定の特徴量の歴史的安定性の主張
  [Nichols, 1992][松本, 2007[2003]]
- 欠けている研究
  – 特徴量の (系統に沿った) 安定性の定量的議論
  – 特徴量の安定性を考慮した系統推定



地図1:【流音タイプの分布】

[松本, 2007[2003]]



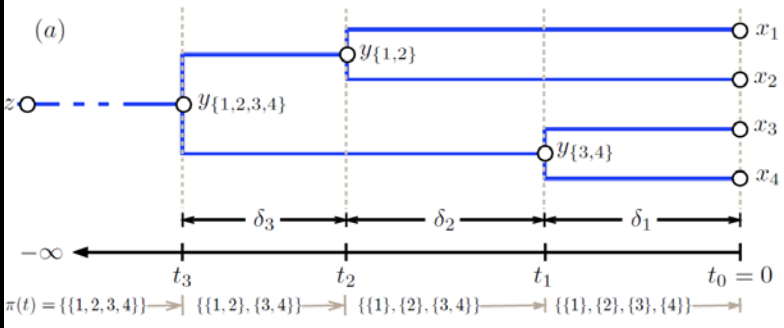
地図1:【流音タイプの分布】

[松本, 2007[2003]]

従来の計算的取り組みは
データの特性を無視

[Teh+, NIPS 2007]

木の確率モデル (提案手法)



＋ 遷移行列: $Q_i = \begin{pmatrix} -\alpha_i & \alpha_i \\ \beta_i & -\beta_i \end{pmatrix}$

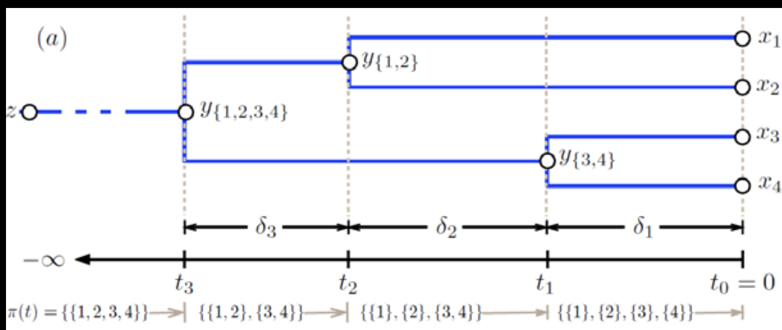
＋ 各言語の ベクトル

2	2	0	⋯	3	3

従来の計算的取り組みはデータの特性を無視

[Teh+, NIPS 2007]

木の確率モデル (提案手法)



＋ 遷移行列: $Q_i = \begin{pmatrix} -\alpha_i & \alpha_i \\ \beta_i & -\beta_i \end{pmatrix}$

＋ 各言語のベクトル

2	2	0	⋯	3	3

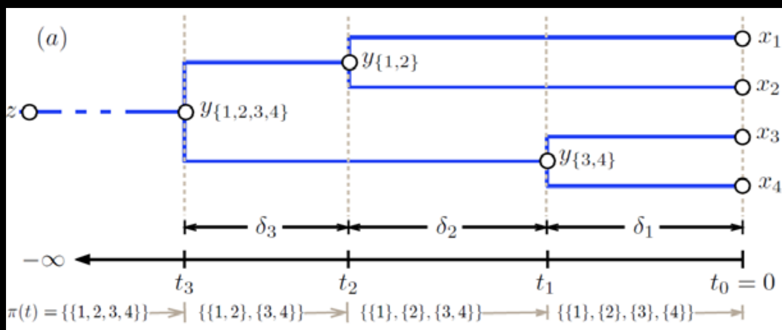
1-of-K変換によるバイナリ化

0	0	1	0	0	0	⋯	0	1

従来の計算的取り組みは
データの特性を無視

[Teh+, NIPS 2007]

木の確率モデル (提案手法)



＋ 遷移行列: $Q_i = \begin{pmatrix} -\alpha_i & \alpha_i \\ \beta_i & -\beta_i \end{pmatrix}$

＋ 各言語のベクトル

2	2	0	⋯	3	3

1-of-K変換によるバイナリ化

0	0	1	0	0	0	⋯	0	1

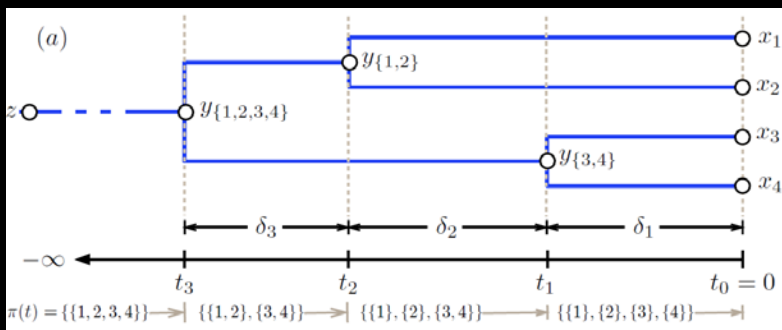
推定された祖語

0	1	1	0	0	1	⋯	1	0

従来の計算的取り組みは
データの特性を無視

[Teh+, NIPS 2007]

木の確率モデル (提案手法)



＋ 遷移行列: $Q_i = \begin{pmatrix} -\alpha_i & \alpha_i \\ \beta_i & -\beta_i \end{pmatrix}$

＋ 各言語のベクトル

2	2	0	⋯	3	3

1-of-K変換によるバイナリ化

0	0	1	0	0	0	⋯	0	1

推定された祖語

0	1	1	0	0	1	⋯	1	0

論理的にありえない!

類型論の目的の一つは普遍性の探究

- 示唆的普遍性 (implicational universal) [Greenberg, 1963][Daumé III, ACL 2007]
  - OV ⊃ 後置詞型，VO ⊃ 前置詞型
  - 後置詞型 ⊃ 属格節-名詞の語順
- 斉一性 (uniformitarianism) 仮説 [Jakobson, 1971[1957]]
  - 現代語から導かれる普遍性は古代語にもあてはまる

類型論の目的の一つは普遍性の探究

- 示唆的普遍性 (implicational universal) [Greenberg, 1963][Daumé III, ACL 2007]
  - OV ⊃ 後置詞型，VO ⊃ 前置詞型
  - 後置詞型 ⊃ 属格節-名詞の語順
- 斉一性 (uniformitarianism) 仮説 [Jakobson, 1971[1957]]
  - 現代語から導かれる普遍性は古代語にもあてはまる

- 特徴量間に相関があるなら、行列をかけて変換すれば良いのでは?
- 言語 (特徴量ベクトル) の自然さを現代語から教師なしで学習し、祖語に適用すれば良いのでは?

提案手法: 類型論の連続空間表現

h	0.21	0.84	⋯	0.03

⬆ encode $\mathrm{h} = \sigma(\mathrm{W_e x} + \mathrm{b_e})$

x	0	0	1	0	0	0	⋯	0

⬆ 多値の特徴量をバイナリ列へ変換 (1-of-K変換)

v	2	2	0	⋯	3	3

提案手法: 類型論の連続空間表現

v'					
2	2	0	⋯	3	3

⬆ バイナリ列から多値の特徴量へ変換

x''							
0	0	1	0	0	0	⋯	0

⬆ 特徴量の制約に基づきバイナリ列を復元

x'							
0.01	0.00	0.92	0.02	0.01	0.01	⋯	0.00

⬆ decode (誤差あり)

$$\mathrm{x}' = \sigma\left(\mathrm{W_d}\mathrm{h} + \mathrm{b_d}\right)$$

h			
0.21	0.84	⋯	0.03

⬆ encode

$$\mathrm{h} = \sigma(\mathrm{W_e}\mathrm{x} + \mathrm{b_e})$$

x							
0	0	1	0	0	0	⋯	0

⬆ 多値の特徴量をバイナリ列へ変換 (1-of-K変換)

v					
2	2	0	⋯	3	3

提案手法: 類型論の連続空間表現

v' | 2 | 2 | 0 | ⋯ | 3 | 3 |

バイナリ列から多値の特徴量へ変換

x'' | 0 | 0 | 1 | 0 | 0 | 0 | ⋯ | 0 |

特徴量の制約に基づきバイナリ列を復元

x' | 0.01 | 0.00 | 0.92 | 0.02 | 0.01 | 0.01 | ⋯ | 0.00 |

decode (誤差あり) $\mathrm{x}' = \sigma(\mathrm{W_d h} + \mathrm{b_d})$

h | 0.21 | 0.84 | ⋯ | 0.03 |

encode $\mathrm{h} = \sigma(\mathrm{W_e x} + \mathrm{b_e})$

x | 0 | 0 | 1 | 0 | 0 | 0 | ⋯ | 0 |

多値の特徴量をバイナリ列へ変換 (1-of-K変換)

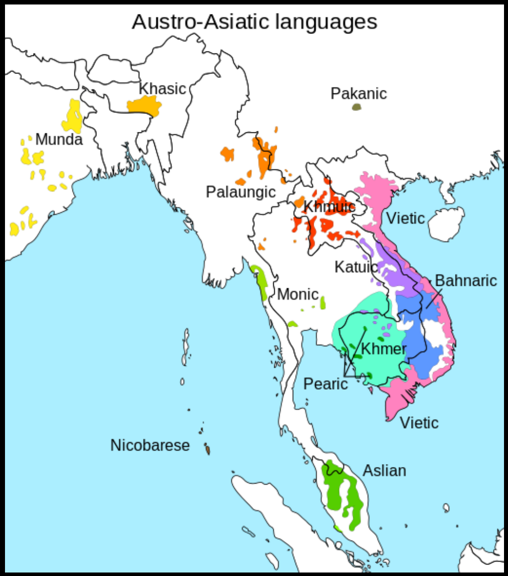
v | 2 | 2 | 0 | ⋯ | 3 | 3 |

自然さ判定

$$\mathrm{P(x)} = \frac{\exp(f(h))}{\sum_{\mathrm{x}'} \exp(f(\mathrm{h}'))}$$

(極端な) 例: オーストロアジア語族のムンダ諸語とモン・クメール諸語

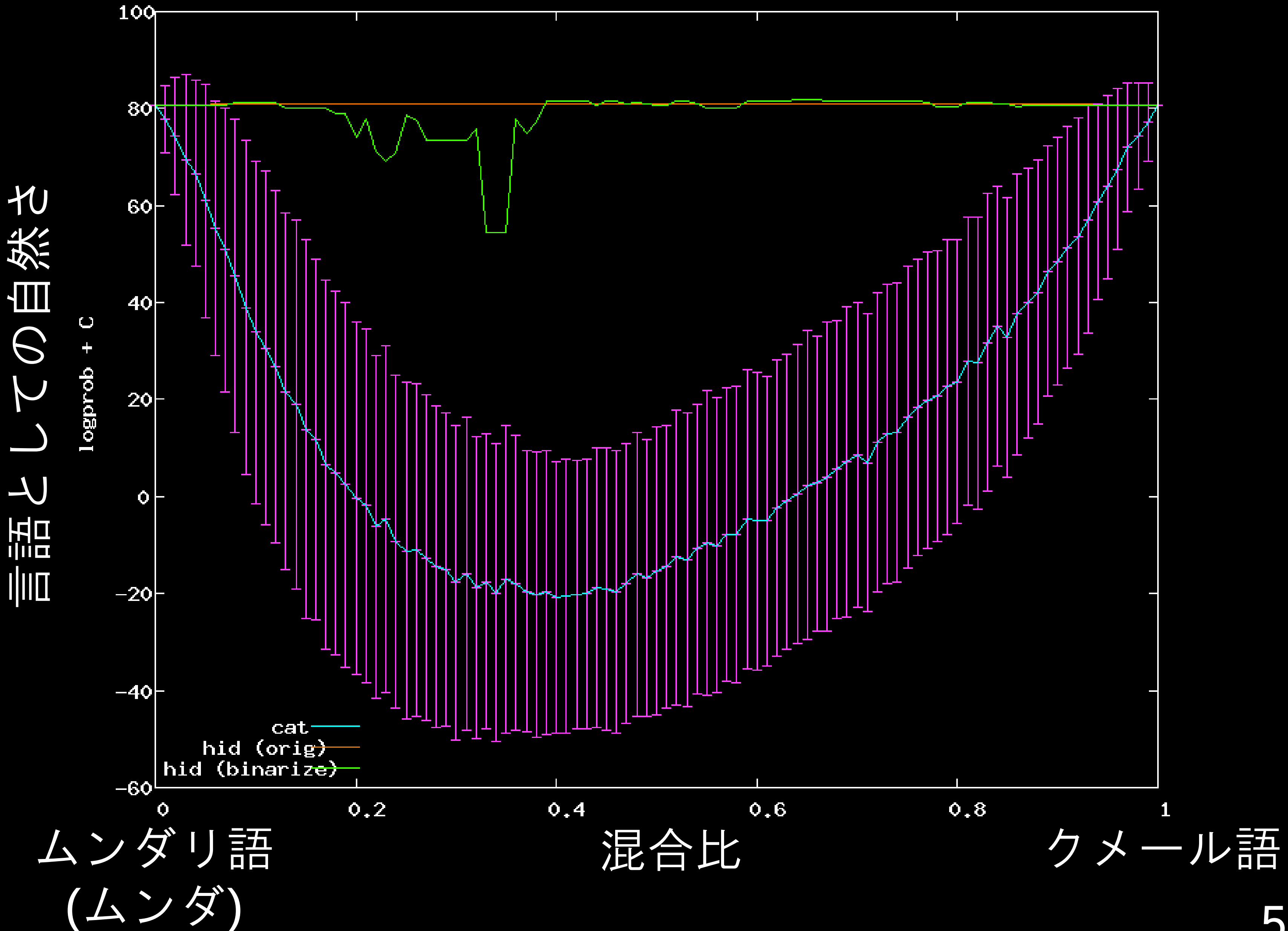
	ムンダ	モン・クメール
文法	統合的	分析的
語順	主辞後置	主辞前置
	OV	VO
	後置詞	前置詞
接辞	接頭辞/接中辞, 接尾辞	接頭辞/接中辞 孤立的



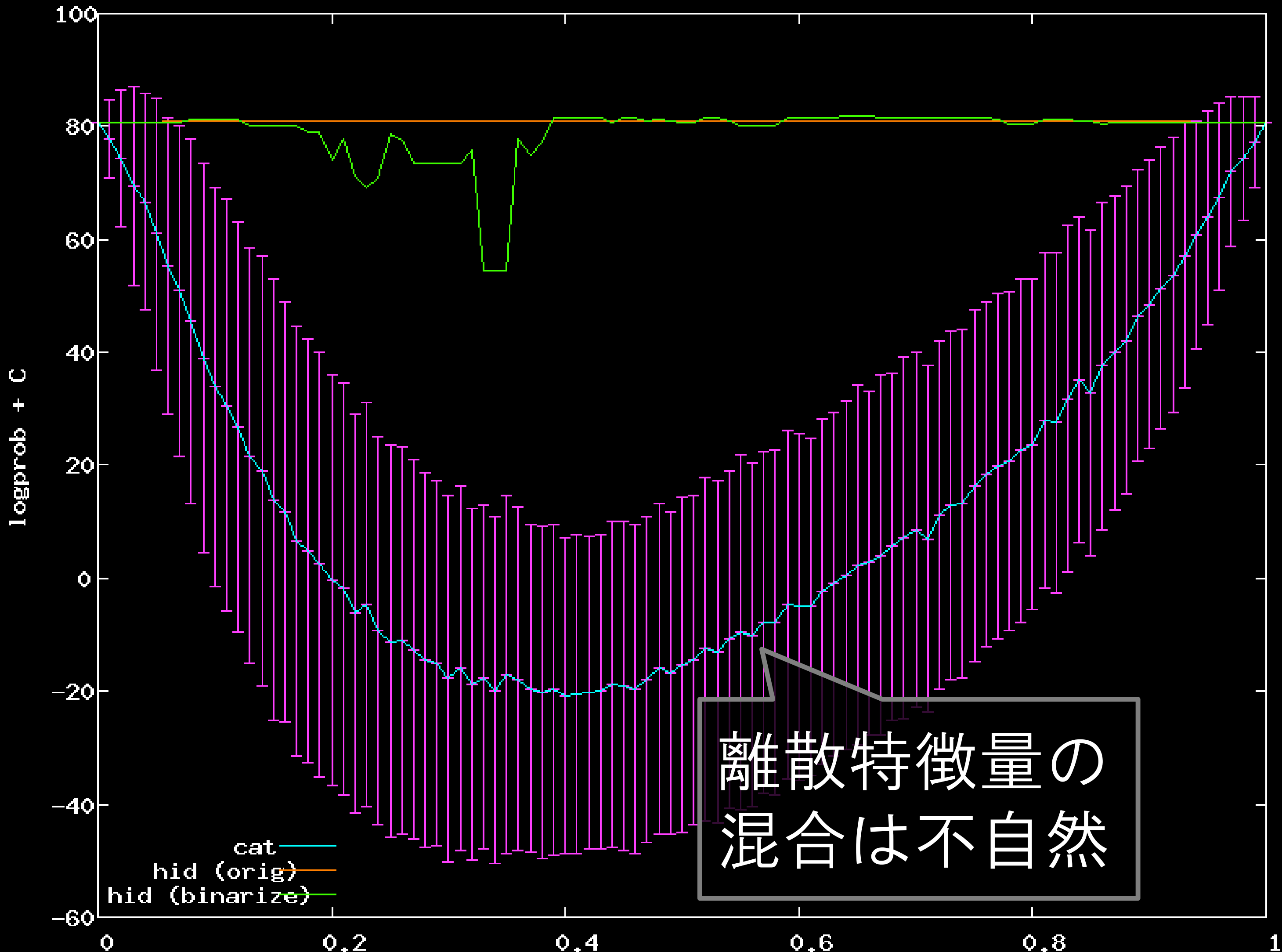
ソラ語 (ムンダ)	*ɑnin*	*dɔŋ-*	*ɲɛn*	*dɑrəj*	*-ən*	*ə-*	*tiy*	*-ben*	*idsɨm*	*-tɛ*	*ted*
	he/she	OBJ-	me	rice	-ART	INF-	give	–INF	want	-3PR	not

クメール語	*k*ət*	*ʔət*	*cɑŋ*	*ʔaoy*	*bay*	*kɲom*
	he/she	not	want	give	rice	me

[Donegan+, 2004]


言語としての自然さ
logprob + C
100
80
60
40
20
0
-20
-40
-60
0
0.2
0.4
0.6
0.8
1
cat
hid (orig)
hid (binarize)
ムンダリ語
(ムンダ)
混合比
クメール語

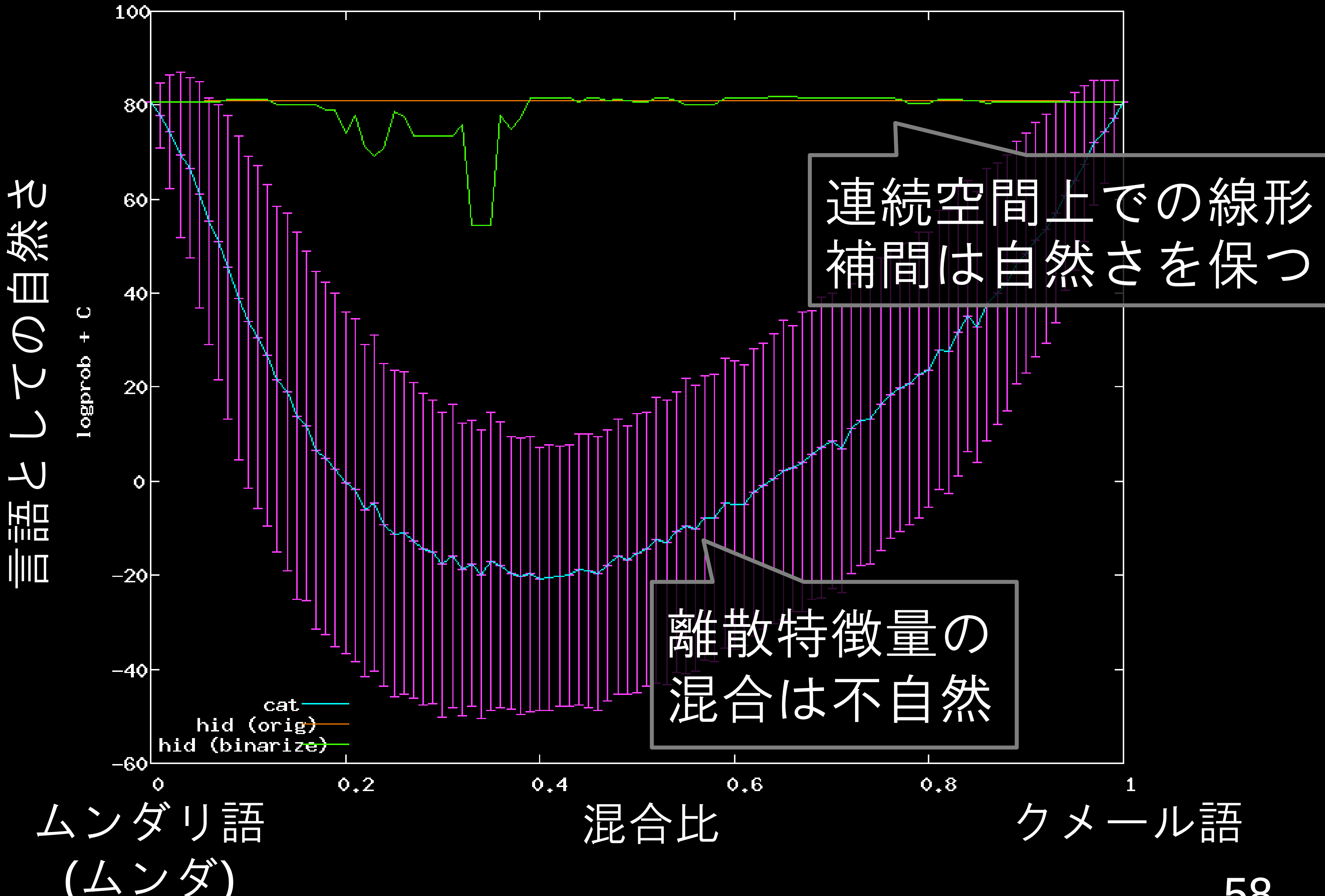
言語としての自然さ

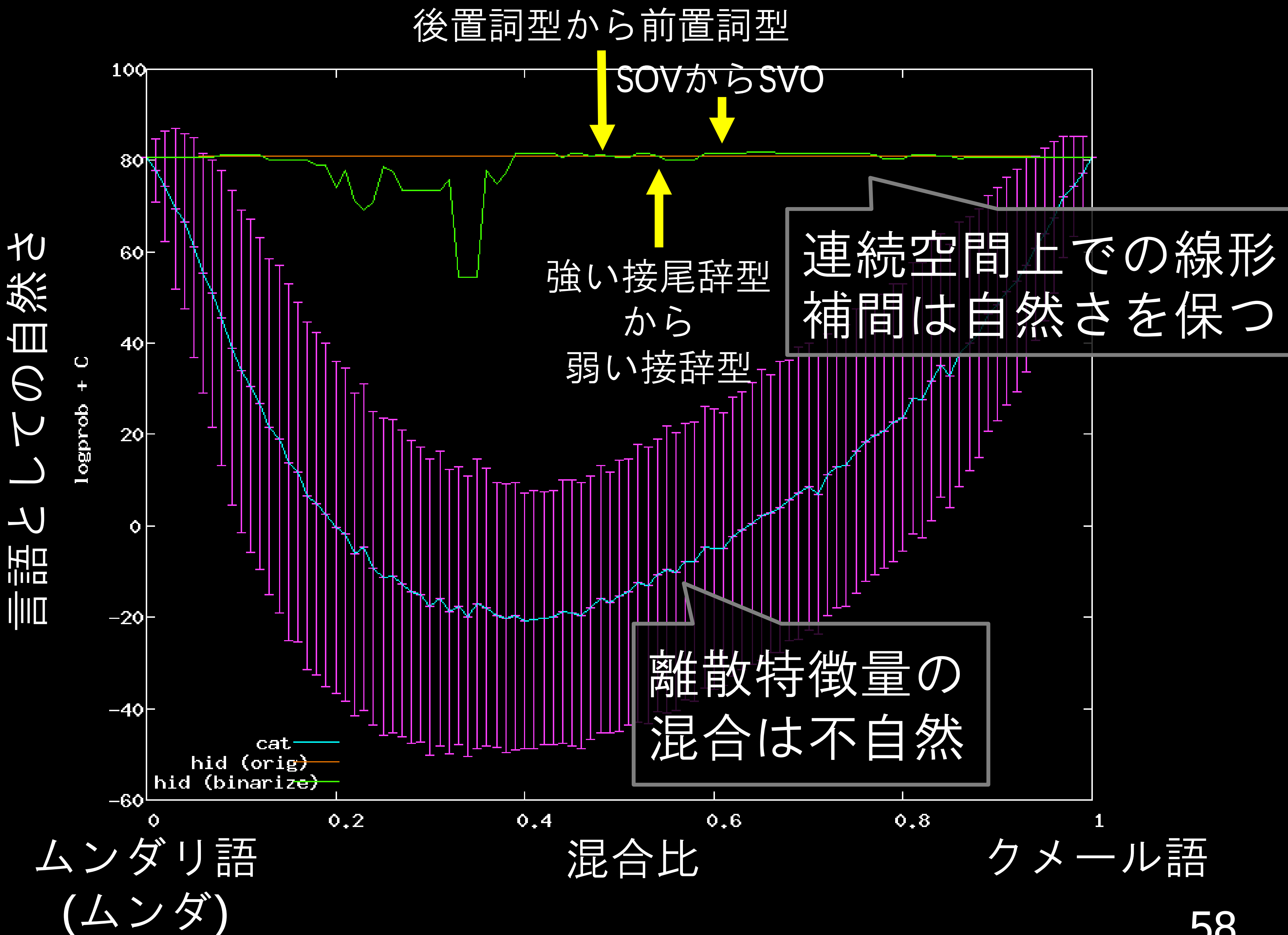


ムンダリ語
(ムンダ)

混合比

クメール語


言語としての自然さ
logprob + C
100
80
60
40
20
0
-20
-40
-60
0
0.2
0.4
0.6
0.8
1
cat
hid (orig)
hid (binarize)
連続空間上での線形
補間は自然さを保つ
離散特徴量の
混合は不自然
ムンダリ語
(ムンダ)
混合比
クメール語



後置詞型から前置詞型
SOVからSVO
強い接尾辞型
から
弱い接辞型
連続空間上での線形
補間は自然さを保つ
離散特徴量の
混合は不自然
言語としての自然さ
logprob + C
100
80
60
40
20
0
-20
-40
-60
cat
hid (orig)
hid (binarize)
0
0.2
0.4
0.6
0.8
1
ムンダリ語
(ムンダ)
混合比
クメール語

祖語の自然さと特徴量の安定性を考慮した世界の系統樹

連続空間上で安定性を考慮した日本語と他の現代語との距離

安定性を考慮した距離

1.	jpn		76
2.	ryu	Japonic	-33
3.	khk	Altaic->Mon.	-198
4.	lep	ST->Tib.-Bur.	-202
5.	chv	Altaic->Tur.	-209
6.	mvf	Altaic->Mon.	-213
7.	bxm	Altaic->Mon.	-217
8.	der	ST->Tib.-Bur.	-221
9.	uum	Altaic->Tur.	-228
10.	huu	Witotoan	-229

離散特徴量上の距離 (不一致率)

1.	jpn		0.0
2.	kxv	Dravidian	-0.394
3.	grt	ST->Tib.Bur.	-0.397
4.	ggo	Dravidian	-0.403
5.	lez	NC->EC->Legzi.	-0.409
6.	chv	Altaic->Tur.	-0.431
7.	huu	Witotoan	-0.436
8.	khk	Altaic->Mon.	-0.453
9.	ryu	Japinic	-0.460
10.	mal	Dravidian	-0.461

まとめ

- 諸言語の歴史的変化は計算でこそ解けるかも
  - 不確実性な手がかり、連続値 (変化率、年代、場所) が扱える
  - 組み合わせ爆発が (近似的に) 扱える
- 生物学由来のモデルをそのまま言語に適用することが多かった
- 言語の性質を踏まえてモデル化する必要