A Bayesian phylogenetic classification of the Siouan family using typological traits

Edwin Ko

University of California, Berkeley

25th International Conference on Historical Linguistics
University of Oxford, August 4th, 2022

Table of Contents

Table of Contents

Introduction

- Today, most linguistic phylogenetic studies use lexical cognate data (Greenhill et al., 2020; Macklin-Cordes et al., 2021).
- Although some studies use typological (or structural) data (e.g. Dunn et al., 2005, 2008; Sicoli and Holton, 2014), their use has been more controversial.
  1. Typological traits are by definition homoplastic (Nichols and Warnow, 2008); that is, they tend to develop independently.
  2. Genealogical signal: Dunn et al. 2005, 2007, 2008; Dunn 2009; Sicoli and Holton 2014; Bøegh et al. 2016
  3. Geographical signal: Donohue and Musgrave 2007; Donohue et al. 2008, 2011

Diffusion or inheritance: An old debate

- The debate of whether typological traits are reliable for classifying languages is over a century old.

(a) Franz Boas

(b) Edward Sapir

- "Sapir came to doubt that extensive morphological patterns could be borrowed [...] Boas came to emphasize the difficulty of distinguishing between the effects of borrowing and the effects of inheritance" (Campbell, 1997, 72)
- It is still unclear how reliable or useful typological features are in historical linguistics (see Wichmann and Saunders, 2007; Gray et al., 2010; Dunn, 2015; Greenhill et al., 2017).

Siouan language family: Traditional classification

- Siouan:
  - **Mandan** (Headley, 1971; Rankin, 2010)
  - **Missouri River** (Voegelin, 1941):
    - Crow, Hidatsa
  - **Mississippi Valley** (Koontz, 1988; Rankin et al., 2015):
    - **Dhegihan**:
      Quapaw, Osage, Omaha, Kansa (Rankin, 1988)
    - **Hocank-Chiwere**:
      Chiwere, Hocank (Miner and Dorsey, 1979)
    - **Dakotan**:
      Assiboine, Lakota, Dakota, Stoney (Parks and DeMallie, 1992)
  - **Ohio Valley** (Voegelin, 1938; Oliverio and Rankin, 2003):
    - Biloxi, Ofo, Tutelo
- Outgroup: **Catawba** (Siebert, 1945a,b; Rankin, 1998)

Rankin's (2010) classification of the Siouan family

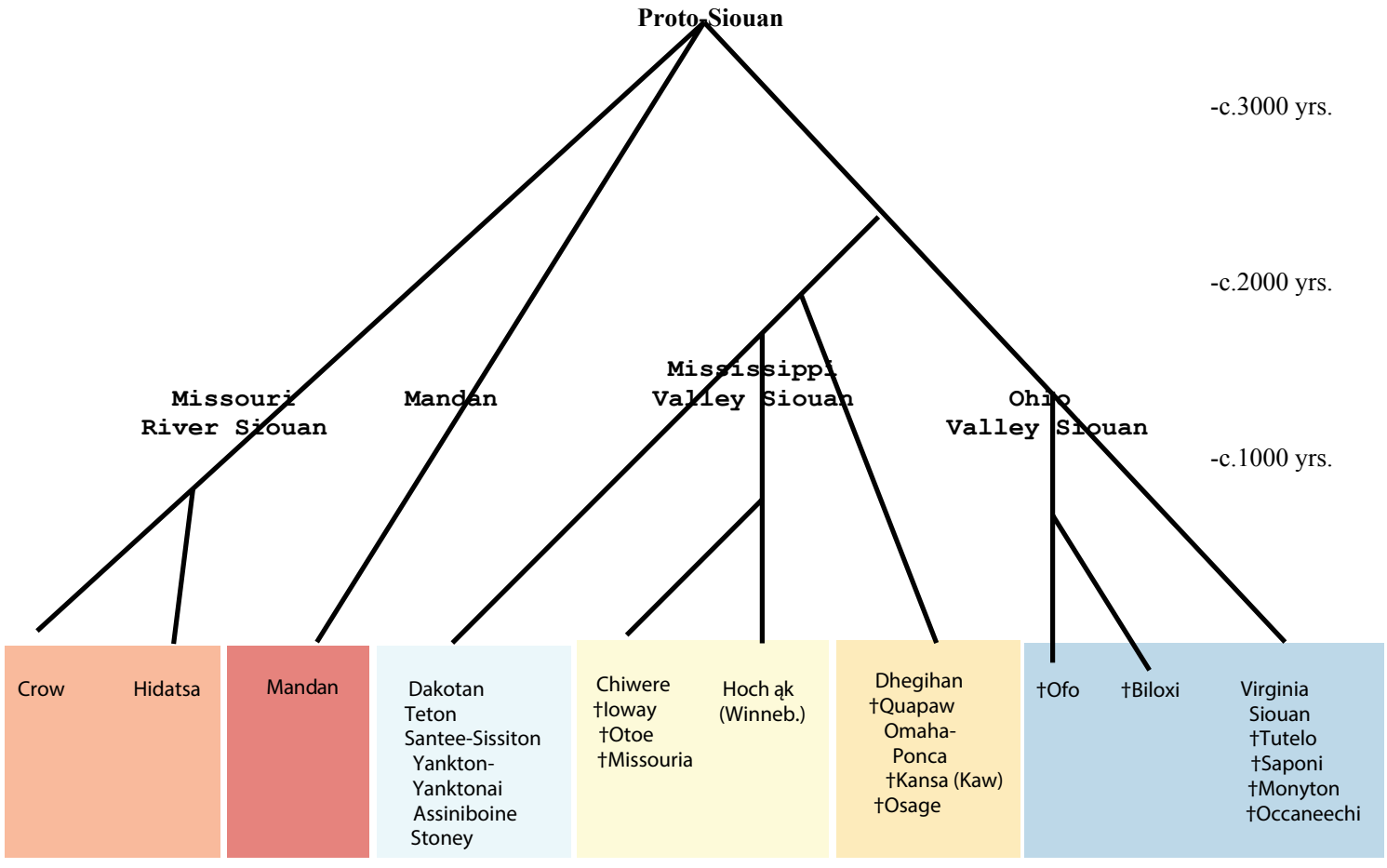
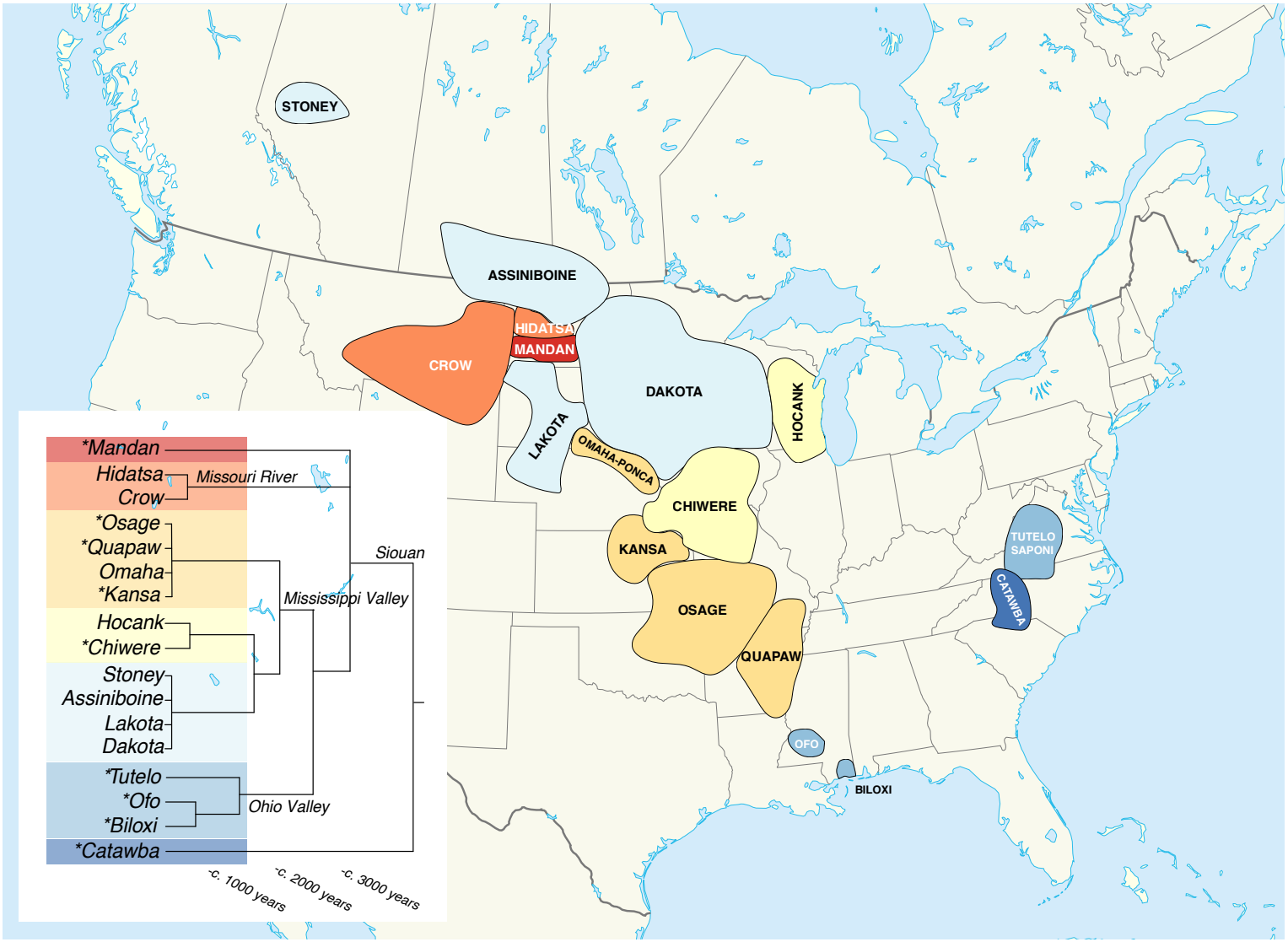


Figure: The proposed tree remains controversial. The placement of some subgroups are unexplained, such as Hocank-Chiwere and Ohio Valley.

Siouan language family: Geographical distribution



Main research questions

1. Can typological data detect a phylogenetic signal?
   - Typological data exhibits a phylogenetic signal.
   - Typological data exhibits parallel developments that are compatible with a contact scenario.
   - Typological data exhibits parallel developments that are incompatible with a contact scenario.

   (see Cathcart et al., 2018, 28–29)
2. How do different traits contribute towards tree inference? Which traits pick out which subgroups?

Key takeaways of this study

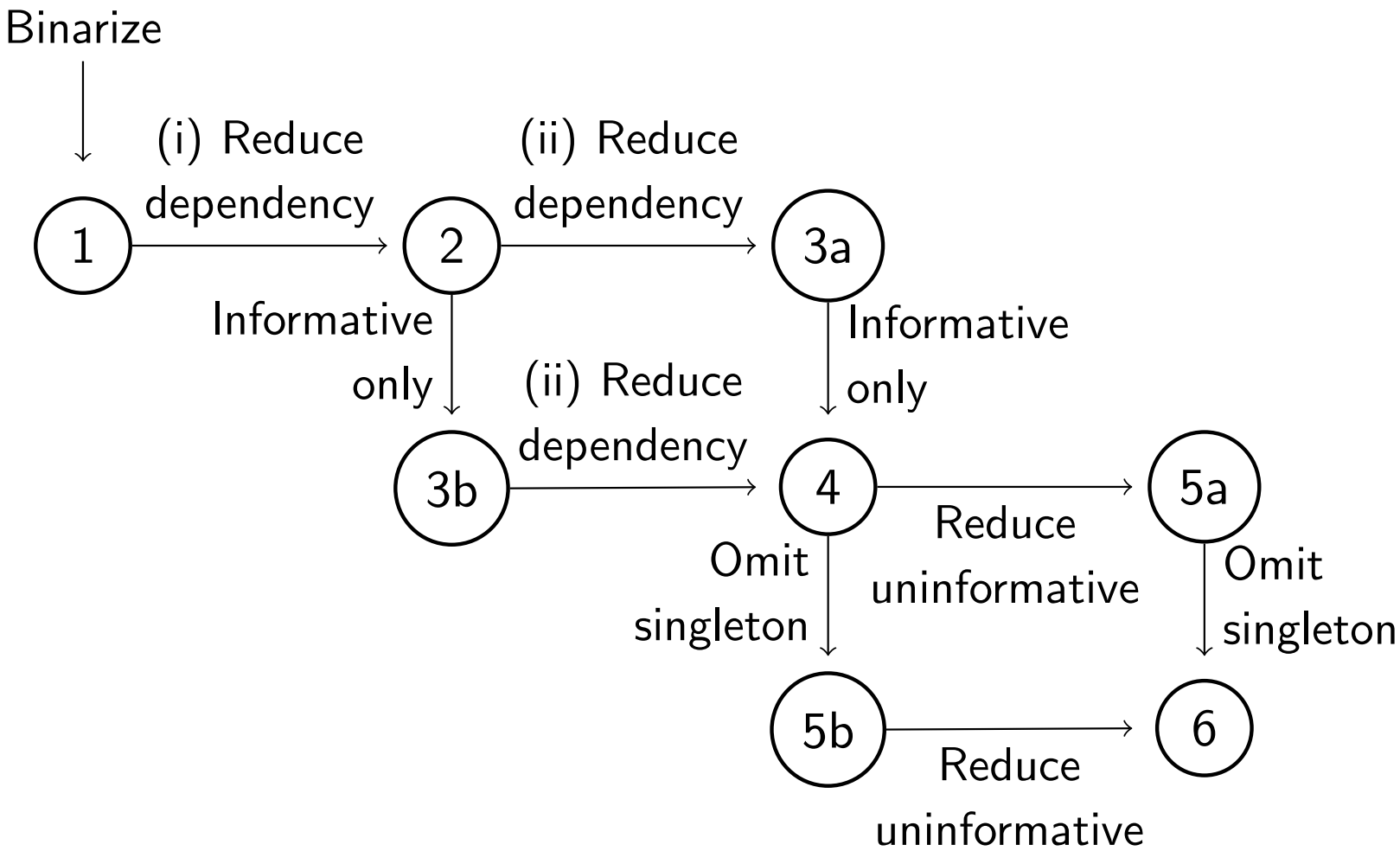
1. Typological data can infer a strong phylogenetic signal, but homoplasy and contact effects may obscure the signal.
2. Typological traits from various areas of grammar and varying degrees of granularity should be used in phylogenetic analyses involving classification.
3. Modifications to the original data set, especially the removal of dependencies, should not only be reported more clearly, but different versions of the data set should also be analyzed.

Table of Contents

Data used in this study

- All traits were coded *from scratch* using data from my own fieldwork on Crow, extant documentation, and personal correspondences with other Siouanists.
  - I will be conducting two months of archival research next week.
- After removing redundant traits across the three typological surveys WALS, Sherzer and Grambank:
  - 127 traits from WALS (Dryer and Haspelmath, 2013)
    - I collapsed the distinction between affix and clitic.
    - A few were adapted (e.g. `Number of Genders`) or omitted (e.g. `Inflectional Synthesis of the Verb`).
  - 93 traits from the modified version of Sherzer (1976) employed by Sicoli and Holton (2014)
    - See Yanovich (2020) for criticisms about their conclusions.
  - 84 traits from the list of morphosyntactic features and guidelines developed by the Grambank consortium (Skirgård et al., submitted)

Coding process: A schematic



Overview of data sets

	DATA SET	MISSING/ALL (% OF MISSING)	NO. OF SITES
(1)	Base	1044/7310 (14.3%)	430
(2)	Intra-trait dependencies	883/6222 (14.2%)	366
(3a)	Inter-trait dependencies	861/5831 (14.8%)	343
(3b)	Informative traits only	883/4947 (17.8%)	291
(4)	Both (3a) and (3b)	862/4624 (18.6%)	272
(5a)	Singleton values removed	842/3893 (21.6%)	229
(5b)	Informative traits only	509/3349 (15.2%)	197
(6)	Both (5a) and (5b)	509/2958 (17.2%)	174

- The amount of missing data varies between 14.3% to 21.6%.
- With small data sets, missing data can negatively impact tree inference using Bayesian analysis (Wiens and Moen, 2008).

Table of Contents

Detecting tree-likeness in the data

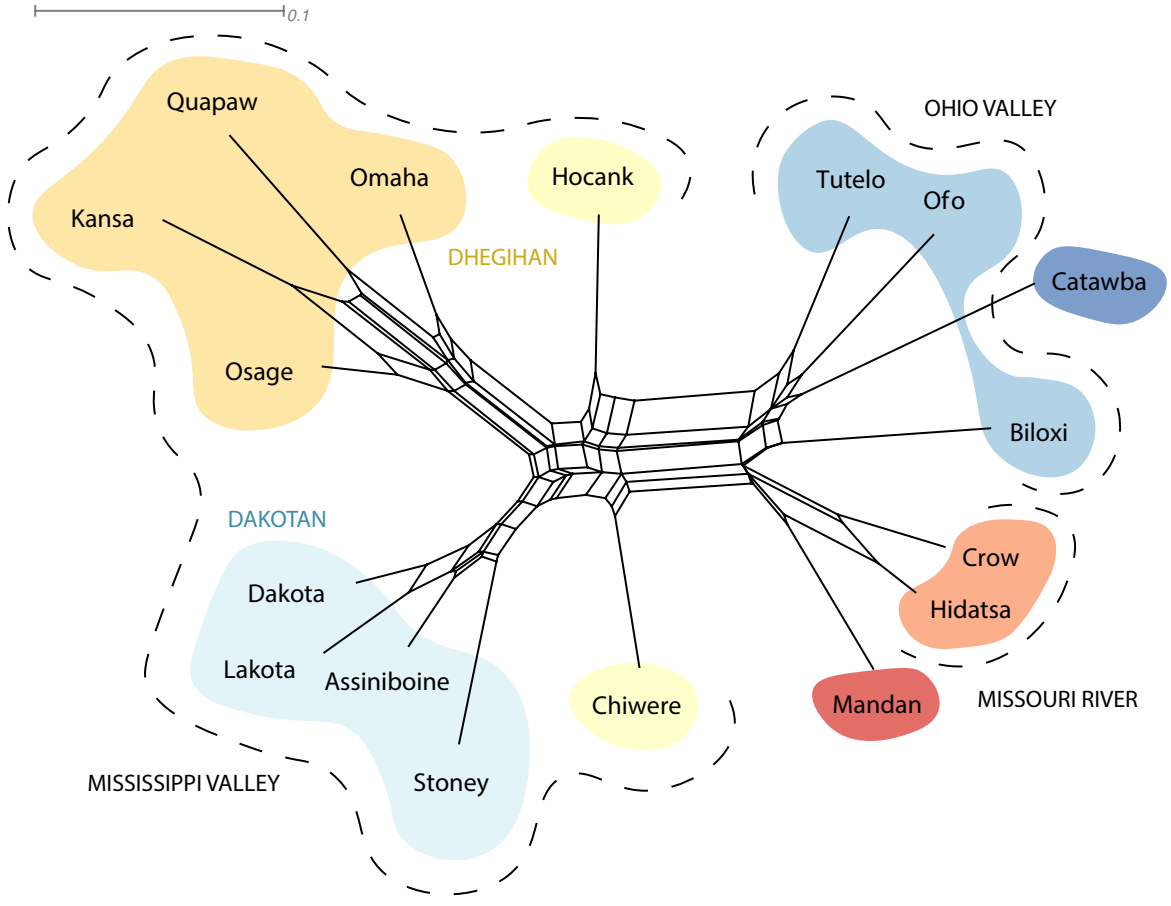


Figure: Splits graph using NeighborNet (Bryant and Moulton, 2004) in SplitsTree4 (Huson and Bryant, 2005). Data Set (4): δ-score $= 0.3077$, Q-residual 0.0294; Data Set (6): δ-score $= 0.3066$, Q-residual $= 0.032$.

Comparison of δ-scores and Q-residuals

Language group	δ-score	Q-residual	Data type	Source
Indo-European	0.20	0.001	Lexical	Kaiping and Klamer 2022
Indo-European	0.23	0.003	Lexical	Gray et al. 2010
Ryukyuan	0.23	0.004	Lexical	Lee and Hasegawa 2014
Ainu	0.25	0.010	Lexical	Lee and Hasegawa 2013
Timor-Alor-Pantar	0.26	0.005	Lexical	Kaiping and Klamer 2022
Chapacuran	0.26	0.016	Lexical	Birchall et al. 2016
Bornean	0.28	0.005	Lexical	Smith and Rama 2022
Tai	0.28	0.041	Lexical	Dockum 2018
Chinese	0.30	0.005	Lexical	Kaiping and Klamer 2022
Dravidian	0.30	0.007	Lexical	Kolipakam et al. 2018
Tai	0.30	0.026	Biphone transitions	Dockum 2018
Siouan	**0.31**	**0.029–0.032**	**Typological**	—
Turkic	0.34	0.001	Lexical	Savelyev and Robbeets 2020
Tai	0.31	0.039	Phonemes	Dockum 2018
Dene-Yeneseian	0.37	0.049	Typological	Sicoli and Holton 2014
Austronesian	0.38	0.006	Lexical	Greenhill et al. 2017
(Mainland) Japanese	0.39	0.002	Lexical	Lee and Hasegawa 2014
Tupí-Guaraní	0.40	0.032	Lexical	Gerardi and Reichert 2021
Polynesian	0.41	0.020	Lexical	Gray et al. 2010
Austronesian	0.44	0.035	Typological	Greenhill et al. 2017

- The splits graph, δ-score, and Q-residual suggest that the Siouan typological data is well within the range of what is considered **moderately tree-like**.

Model selection

- Marginal likelihood was estimated using nested sampling (Maturana Russel et al., 2019) with 100 particles for (4).

SUBSTITUTION MODEL	LOG MARGINAL LIKELIHOOD	SD	BF
Covarion, relaxed clock	**-1310.97**	**0.73**	—
GTR, relaxed clock	**-1311.74**	**0.71**	**1.54**
GTR+Γ, relaxed clock	-1315.58	0.71	9.22
Covarion, strict clock	-1316.39	0.68	10.84
GTR+I, relaxed clock	-1316.92	0.78	11.90
GTR, strict clock	-1317.72	0.69	13.50
GTR+Γ, strict clock	-1321.70	0.68	21.46
GTR+I, strict clock	-1323.69	0.72	25.44

- The covarion is a general form of the proportion invariant (+I) model (Huelsenbeck, 2002); it is worth also comparing with the +I model (p.c., Huelsenbeck, July 2022).

Bayesian inference

- I ran the analyses in `BEAST 2.6.7` (Bouckaert et al., 2019) which uses MCMC to sample the posterior distribution with 50 million generations with a 1,000 sampling frequency and 25% burn-in resulting in a total of 37,500 trees.
  - The number of generations was sufficient enough to yield a reasonable degree of convergence (i.e. $>$ 200 ESS, 'hairy caterpillar') for all analyses.
  - This process was repeated two additional times to ensure the results are similar across the three independent runs.
- The analyses I show here employ the covarion model with the (uncorrelated lognormal) relaxed clock under a constant-rate birth-death process and do not employ any clade constraints.
  - Other models were also used to check for robustness of topology inference (see Yanovich, 2020).
  - Ideally, each data set should have undergone the same process of model evaluation.

Table of Contents

Comparison between the Rankin tree and Analysis (1)

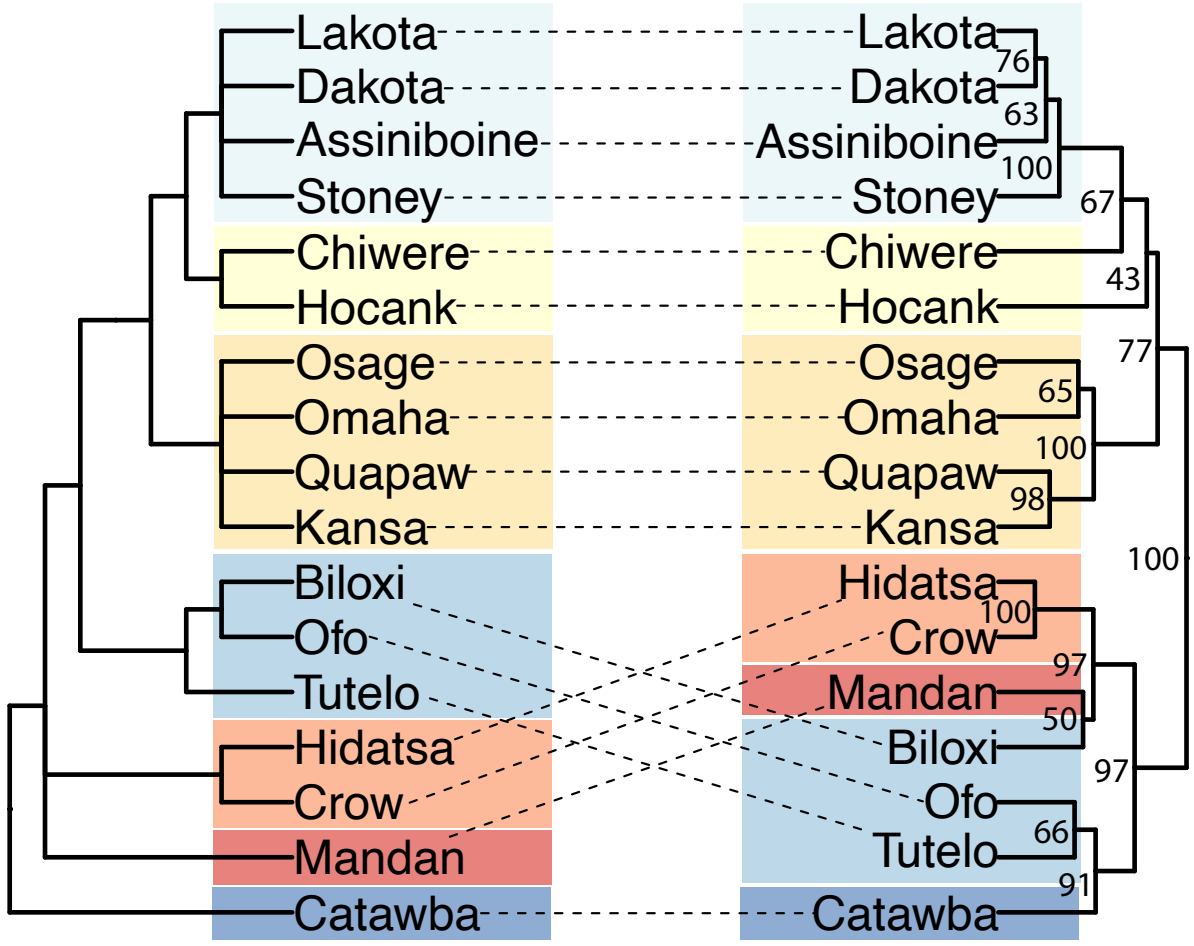


Figure: Rankin tree (left) and summary tree for Analysis (1) (right).

(4) Reduced dependencies and uninformative sites

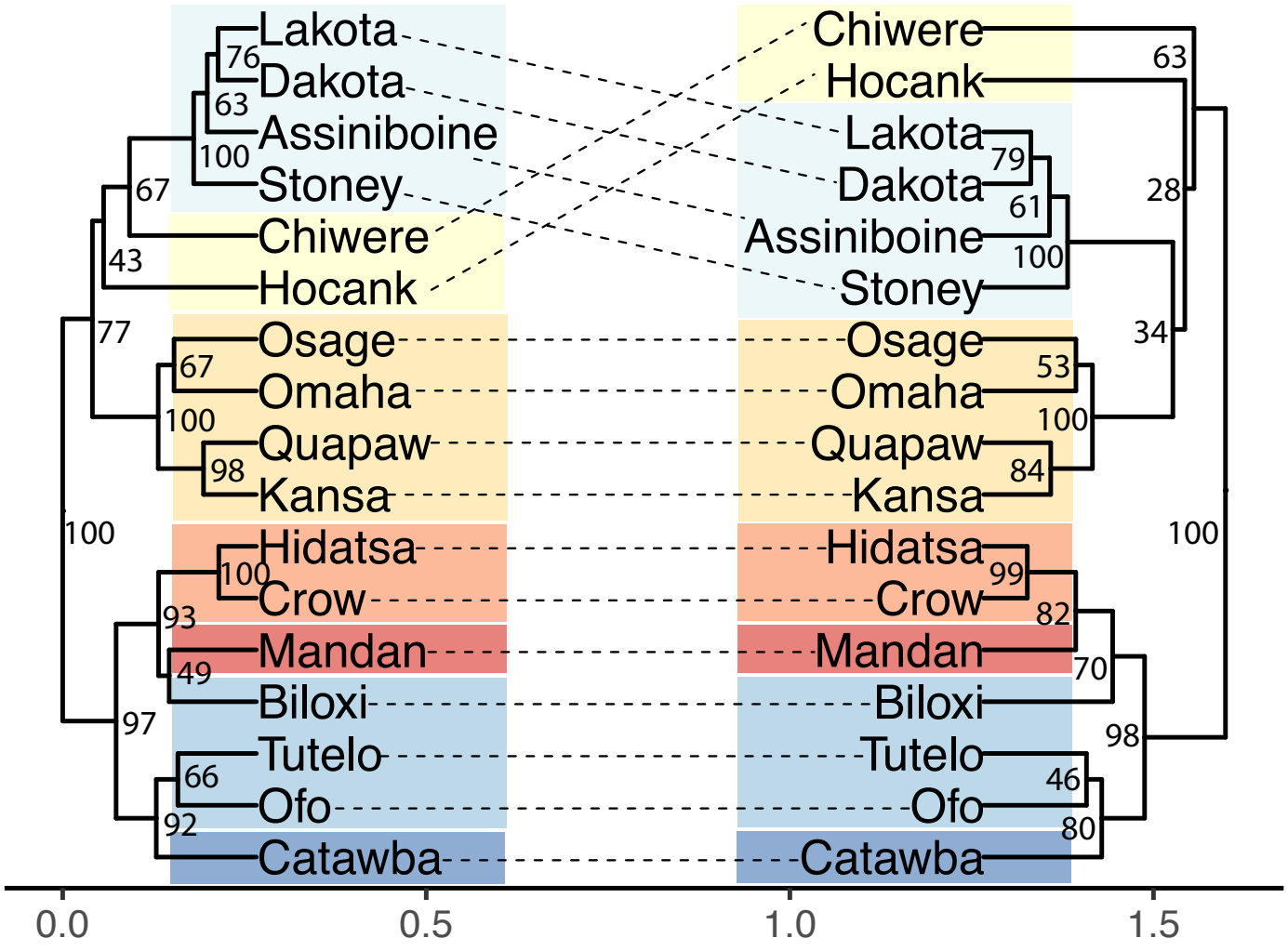


Figure: Summary trees for Analysis (1) (left) and Analysis (4) (right).

Investigating the contributions of specific traits

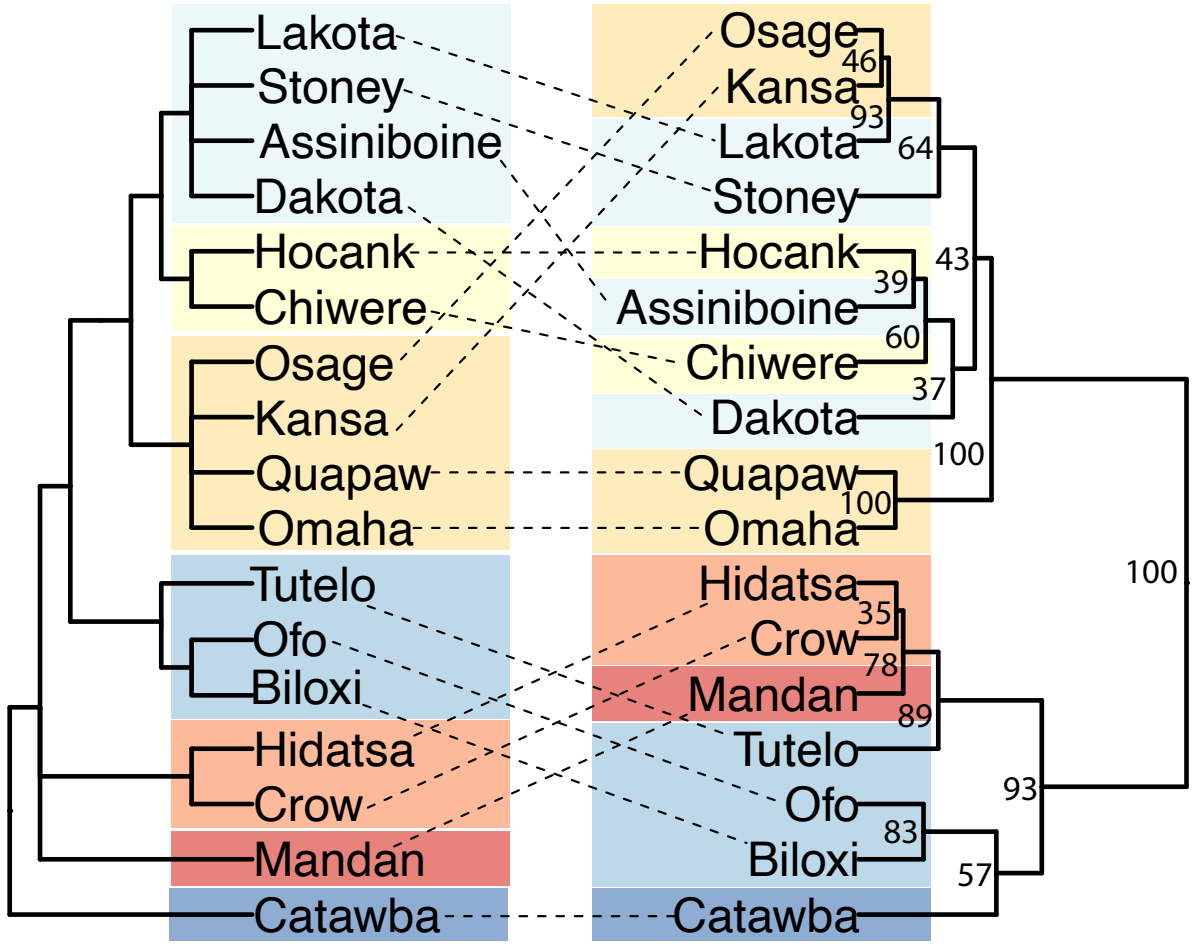


Figure: Phonology (36 sites) is responsible for the higher-order subgroups. Rankin tree (left) and summary tree (right).

Investigating the contributions of specific traits

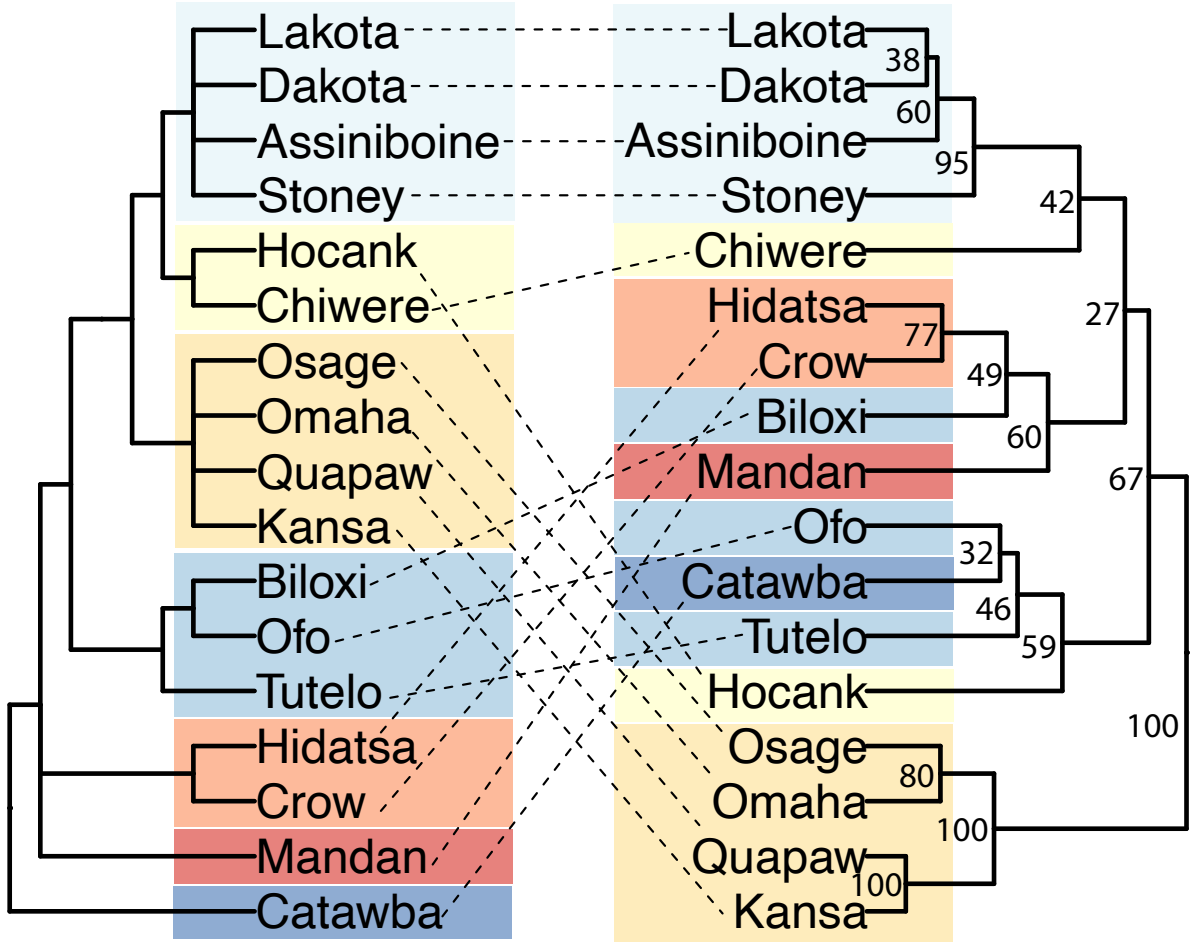


Figure: Morphology (110 sites) is responsible for the lower-order subgroups. Rankin tree (left) and summary tree (right).

Investigating the contributions of specific traits

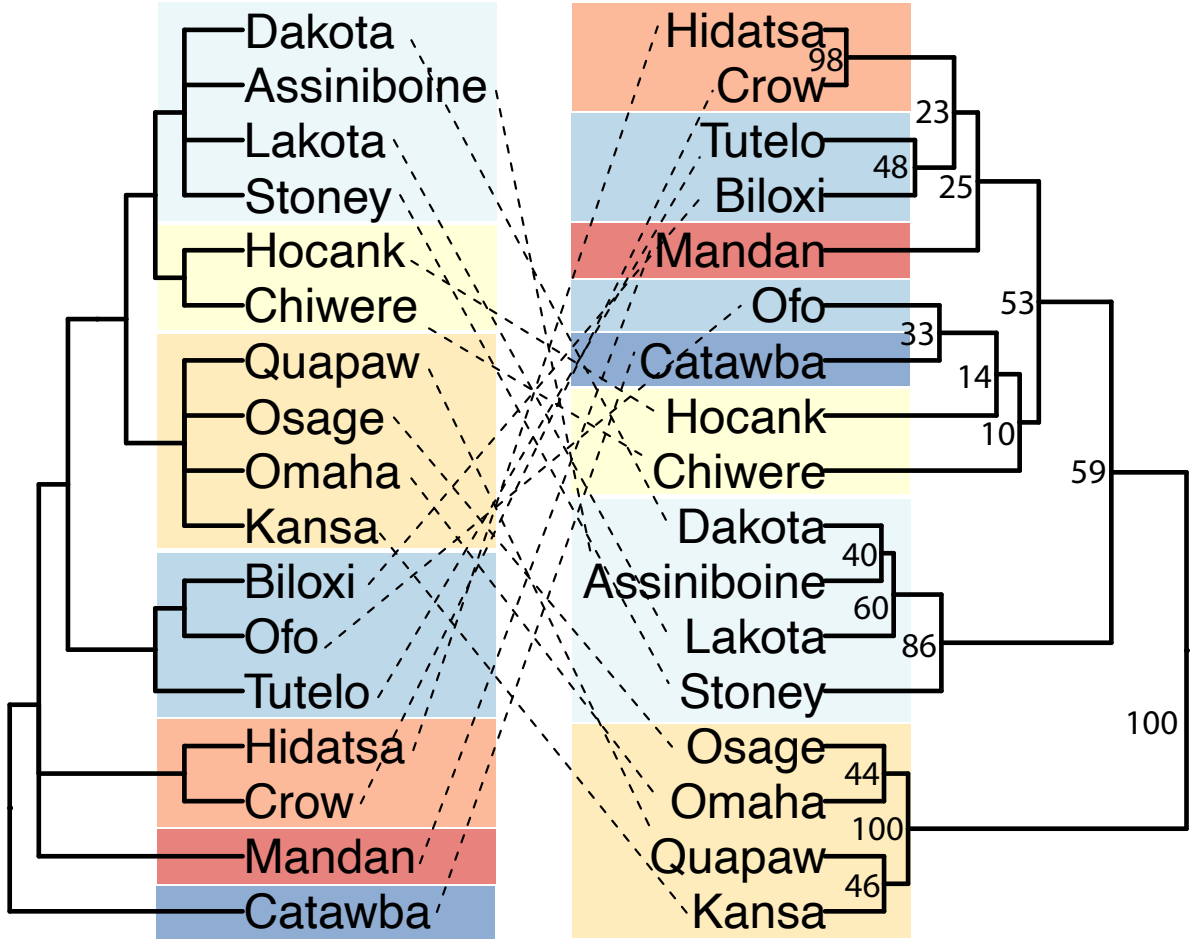


Figure: Nominal morphology (49 sites) is responsible for the lower-order subgroups. Rankin tree (left) and summary tree (right).

Investigating the contributions of specific traits

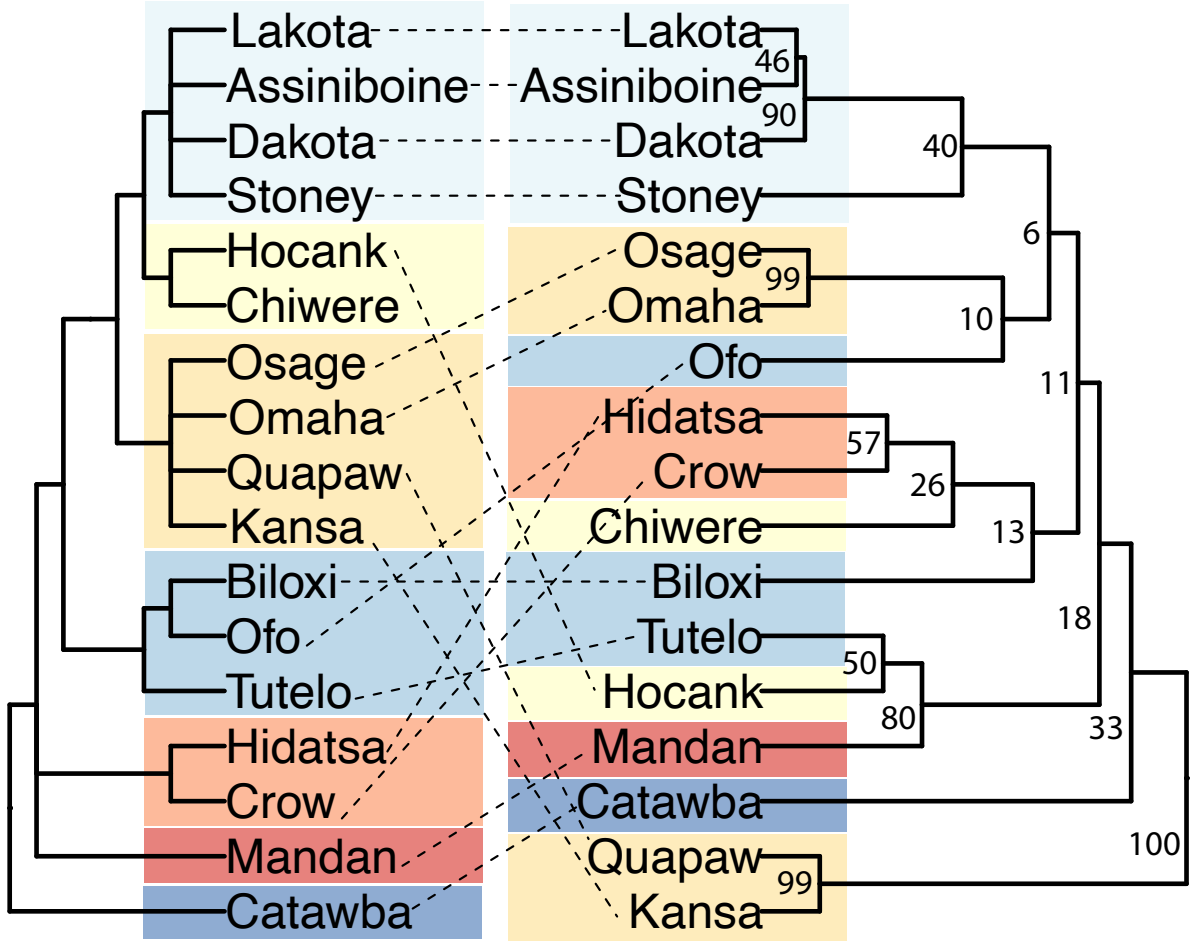


Figure: Verbal morphology (61 sites) is responsible for even lower-order subgroups. Rankin tree (left) and summary tree (right).

Table of Contents

Phylogeny and homoplasy

- Data sets (1)–(6) recover the Rankin tree fairly well, but there are some inconsistencies:
  1. The grouping of Catawba with Ofo and Tutelo is compatible with a contact scenario.
  2. The placement of Ohio Valley with Missouri River and Mandan is unexpected: Ohio Valley (Biloxi, Ofo, Tutelo), Missouri River (Crow, Hidatsa), and Mandan lost ejectives on stops and fricatives which were likely independent developments (cf. Rankin et al., 1997).
     - Domino effects: Loss of these phonemic contrasts therefore generally reduced the size of the consonant inventory and decreased the consonant-vowel ratio for these languages.
  3. The variable placement of Hocank and Chiwere and low posteriors reflect the sentiments expressed by Rankin (2010):
     - "There is, however, still controversy about the relative chronology of its internal splits. Do Dakotan and Chiwere pair up against Dhegiha or do Dhegiha and Chiwere pair up against Dakotan."

Speculations on phonological and morphological traits

1. Broader phonological traits tend to reflect changes that are shared across higher-order subgroups.
   - Individual sound changes may not be as reliable for subgrouping (Ringe et al., 2002; Babel et al., 2013).
   - Impressionistically, phonological traits that remain in the data set appear broader than morphological traits.
   - Coarse-grained phonological traits, such as size of consonant inventory or number of stop series, that capture phonological resembles may be due to not one but several shared changes.
2. Lower-order subgroups tend to exhibit more morphological similarities than higher-order ones.
   - Convergence may occur for languages that remain in contact after splitting (Garrett, 2006).
   - In Siouan, verbs are the most highly inflected (Rankin et al., 2003), and I conjecture that varieties that are in closer contact share more verbal morphology. Thus, verbal morphology recovers even lower-order subgroups than nominal morphology.

Dependencies in the data set

- Dependencies may introduce 'noise' or exaggerate certain subgroups regardless of whether they are consistent with traditional classifications (Reesink and Dunn, 2012).
- In this study, reducing almost 90 interdependent sites (i.e. 20% of the base data set) which are completely predictable or overlapping did not appear to seriously affect the results.
- Sites that are not totally predictable remained – these sites have information not expressed by other sites.
  - In fact, removing eight sites pertaining to phonological traits that are partially predictable and overlapping results in the Mississippi Valley no longer being recovered.
- The potential effects of dependencies (logical entailments or otherwise) on phylogenetic analyses are still largely unknown.

Table of Contents

Closing remarks

- Typological features were selected from surveys and databases that incorporated traits from different areas of grammar with varying degrees of granularity.
  - While a few traits were adapted for the Siouan languages, the vast majority of traits were selected to capture the world's linguistic diversity.
  - What level of granularity or areas of grammar provides the most information about the different subgroups? To what extent are the results lineage-specific or dependent on the selection of features?
- Such cases as independent developments and correlated evolution may attenuate the phylogenetic signal produced by the analysis; dependencies may further exacerbate this issue.

Closing remarks

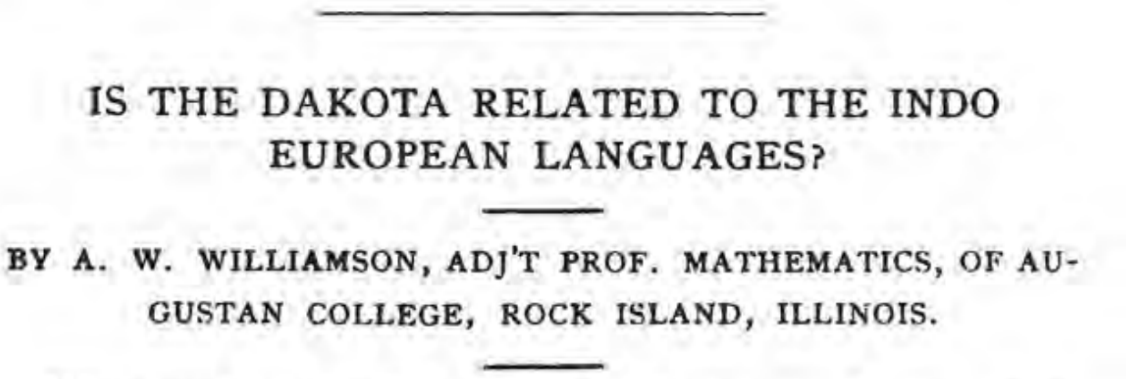
- There are assumptions that go into the design of phylogenetic analyses to determine what data gets included or excluded, as in the following remarks that can be found in the literature:
  - "we judge it innocuous to allow features with some degree of logical dependency between them to remain"
  - "Interdependent features were filtered as much as possible from the data set"
  - "characters with weaker tendencies to covariance were not excluded"
- It is important to clarify the coding process (see Wu and List, to appear) and report on results that use different data sets; doing so leads to greater transparency in research design and replicability.

Acknowledgements

Many thanks to Andrew Garrett, John Huelsenbeck, John Boyle, Gašper Beguš, Randolph Graczyk, Jill Greer, Iren Hartmann, David Kaufman, Armik Mirzayan, Corey Roberts, Catharine Rudin, Kathy Shea, and other participants at the 40th Siouan and Caddoan Conference for comments, discussion and suggestions at various stages of this work. Thanks also to my friends and collaborators on the Crow Indian Reservation, particularly Felice Big Day, Cyle Old Elk, Jack Real Bird, and Riley Singer for sharing their beautiful culture and language with me.

This material is based upon work supported by the National Science Foundation under Grant No. BCS 2215488 and the American Philosophical Society Daythal L. Kendall Fellowship.

Thank you for listening!



IS THE DAKOTA RELATED TO THE INDO EUROPEAN LANGUAGES?

BY A. W. WILLIAMSON, ADJ'T PROF. MATHEMATICS, OF AUGUSTAN COLLEGE, ROCK ISLAND, ILLINOIS.

"This paper is a preliminary result of my father's dying request to complete an article he was preparing showing that the Dakotas are of European origin" (Williamson, 1881, 139)

Bibliography I

Babel, M., Garrett, A., Houser, M. J., and Toosarvandani, M. (2013). Descent and diffusion in language diversification: A study of western numic dialectology. *International journal of American linguistics*, 79(4):445–489.

Beverley, R. (1705). History and Present State of Virginia.

Birchall, J., Dunn, M., and Greenhill, S. J. (2016). A combined comparative and phylogenetic analysis of the Chapacuran language family. *International Journal of American Linguistics*, 82(3):255–284.

Bøegh, K. F., Daval-Markussen, A., and Bakker, P. (2016). A phylogenetic analysis of stable structural features in west african languages. *Studies in African Linguistics*, 45(1 & 2):62–94.

Booker, K. M., Hudson, C. M., and Rankin, R. L. (1992). Place name identification and multilingualism in the sixteenth-century southeast. *Ethnohistory*, pages 399–451.

Bouckaert, R., Vaughan, T. G., Barido-Sottani, J., Duchêne, S., Fourment, M., Gavryushkina, A., Heled, J., Jones, G., Kühnert, D., De Maio, N., et al. (2019). BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. *PLoS computational biology*, 15(4):e1006650.

Bryant, D. and Moulton, V. (2004). Neighbor-net: an agglomerative method for the construction of phylogenetic networks. *Molecular biology and evolution*, 21(2):255–265.

Campbell, L. (1997). *American Indian languages: the historical linguistics of Native America*, volume 4. Oxford University Press.

Cathcart, C., Carling, G., Larsson, F., Johansson, N., and Round, E. (2018). Areal pressure in grammatical evolution: An Indo-European case study. *Diachronica*, 35(1):1–34.

Dockum, R. (2018). Phylogeny in phonology: How Tai sound systems encode their past. In *Proceedings of the Annual Meetings on Phonology*, volume 5.

Donohue, M. and Musgrave, S. (2007). Typology and the linguistic macrohistory of Island Melanesia. *Oceanic linguistics*, 46(2):348–387.

Donohue, M., Musgrave, S., Whiting, B., and Wichmann, S. (2011). Typological feature analysis models linguistic geography. *Language*, 87(2):369–383.

Bibliography II

Donohue, M., Wichmann, S., and Albu, M. (2008). Typology, areality, and diffusion. *Oceanic Linguistics*, 47(1):223–232.

Dryer, M. S. and Haspelmath, M., editors (2013). *WALS Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Dunn, M. (2009). Contact and phylogeny in Island Melanesia. *Lingua*, 119(11):1664–1678.

Dunn, M. (2015). Language phylogenies. *The Routledge handbook of historical linguistics*, pages 190–211.

Dunn, M., Foley, R., Levinson, S., Reesink, G., and Terrill, A. (2007). Statistical reasoning in the evaluation of typological diversity in Island Melanesia. *Oceanic Linguistics*, pages 388–403.

Dunn, M., Levinson, S. C., Lindström, E., Reesink, G., and Terrill, A. (2008). Structural phylogeny in historical linguistics: Methodological explorations applied in Island Melanesia. *Language*, 84(4):710–759.

Dunn, M., Terrill, A., Reesink, G., Foley, R. A., and Levinson, S. C. (2005). Structural phylogenetics and the reconstruction of ancient language history. *Science*, 309(5743):2072–2075.

Garrett, A. (2006). Convergence in the Formation of Indo-European Subgroups: Phylogeny and Chronology. In Forster, P. and Renfre, C., editors, *Phylogenetic methods and the prehistory of languages*, pages 139–151. Cambridge: McDonald Institute for Archaeological Research.

Gerardi, F. F. and Reichert, S. (2021). The Tupí-Guaraní language family: A phylogenetic classification. *Diachronica*, 38(2):151–188.

Gray, R. D., Bryant, D., and Greenhill, S. J. (2010). On the shape and fabric of human history. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 365(1559):3923–3933.

Greenhill, S. J., Heggarty, P., and Gray, R. D. (2020). Bayesian phylolinguistics. In Janda, R., Joseph, B., and Vance, B., editors, *The Handbook of Historical Linguistics, Volume II*, pages 226–253. Wiley Online Library.

Greenhill, S. J., Wu, C.-H., Hua, X., Dunn, M., Levinson, S. C., and Gray, R. D. (2017). Evolutionary dynamics of language systems. *Proceedings of the National Academy of Sciences*, 114(42):E8822–E8829.

Headley, R. K. (1971). The origin and distribution of the Siouan-speaking tribes. Master's thesis, Catholic University of America.

Bibliography III

Huelsenbeck, J. P. (2002). Testing a covariotide model of DNA substitution. *Molecular biology and evolution*, 19(5):698–707.

Huson, D. H. and Bryant, D. (2005). Estimating phylogenetic trees and networks using splitstree 4. *Manuscript in preparation, software available from www. splitstree. org*.

Kaiping, G. A. and Klamer, M. (2022). The dialect chain of the Timor-Alor-Pantar language family: A new analysis using systematic Bayesian phylogenetics. *Language Dynamics and Change*, 1(aop):1–53.

Kolipakam, V., Jordan, F. M., Dunn, M., Greenhill, S. J., Bouckaert, R., Gray, R. D., and Verkerk, A. (2018). A Bayesian phylogenetic study of the Dravidian language family. *Royal Society open science*, 5(3):171504.

Koontz, J. (1988). Isoglosses in Proto-Mississippi Valley Siouan. Paper presented at the Belcourt Lecture, University of Manitoba, Winnepeg, Canada.

Lee, S. and Hasegawa, T. (2013). Evolution of the Ainu language in space and time. *PLoS One*, 8(4):e62243.

Lee, S. and Hasegawa, T. (2014). Oceanic barriers promote language diversification in the Japanese Islands. *Journal of Evolutionary Biology*, 27(9):1905–1912.

Lewis, P. O. (2001). A likelihood approach to estimating phylogeny from discrete morphological character data. *Systematic biology*, 50(6):913–925.

List, J.-M. (2016). Beyond cognacy: Historical relations between words and their implication for phylogenetic reconstruction. *Journal of Language Evolution*, 1(2):119–136.

Macklin-Cordes, J. L., Bowern, C., and Round, E. R. (2021). Phylogenetic signal in phonotactics. *Diachronica*.

Mantel, N. (1967). The detection of disease clustering and a generalized regression approach. *Cancer research*, 27(2_Part_1):209–220.

Maturana Russel, P., Brewer, B. J., Klaere, S., and Bouckaert, R. R. (2019). Model selection and parameter inference in phylogenetics using nested sampling. *Systematic biology*, 68(2):219–233.

Miner, K. L. and Dorsey (1979). Dorsey's law in Winnebago-Chiwere and Winnebago accent. *International journal of American linguistics*, 45(1):25–33.

Bibliography IV

Nichols, J. and Warnow, T. (2008). Tutorial on computational linguistic phylogeny. *Language and Linguistics Compass*, 2(5):760–820.

Oliverio, G. R. and Rankin, R. L. (2003). On the Sub-Grouping of the Virginian Siouan Languages. In Rudes, B. and Costa, D., editors, *Essays in Algonquian, Catawban, and Siouan Linguistics in Memory of Frank T. Siebert, Jr. Memoir*, volume 16. Winnipeg, Manitoba: Algonquian and Iroquoian Linguistics.

Parks, D. R. and DeMallie, R. J. (1992). Sioux, Assiniboine, and Stoney dialects: A classification. *Anthropological Linguistics*, 4(1):233–255.

Rankin, R. L. (1988). Quapaw: Genetic and areal affiliations. In Shipley, W., editor, *In Honor of Mary Haas: From the Haas Festival Conference on Native American Linguistics*, pages 629–650. Berlin, Germany: De Gruyter Mouton.

Rankin, R. L. (1998). Siouan-Catawban-Yuchi genetic relationship: With a note on Caddoan. Paper presented at the *18th annual Siouan and Caddoan Languages Conference*, Bloomington, IN.

Rankin, R. L. (2010). The place of Mandan in the Siouan language family. Paper presented at the *30th Annual Siouan and Caddoan Languages Conference*, Chicago, IL.

Rankin, R. L., Boyle, J., Graczyk, R., and Koontz, J. E. (2003). Synchronic and diachronic perspective on 'word'in Siouan. In Dixon, R. and Aikenvald, A., editors, *Word: A cross-linguistic typology*, pages 180–204.

Rankin, R. L., Carter, R. T., and Jones, A. W. (1997). Proto-Siouan phonology and grammar. In Li, X., López, L., and Stroik, T., editors, *Papers from the 1997 Mid-America Linguistics Conference*, pages 366–375. Columbia, MO: University of Missouri Press.

Rankin, R. L., Carter, R. T., Jones, A. W., Koontz, J. E., Rood, D. S., and Hartmann, I., editors (2015). *Comparative Siouan Dictionary*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Reesink, G. and Dunn, M. (2012). Systematic typological comparison as a tool for investigating language history. *Language documentation and conservation*, (5):34–71.

Rexová, K., Bastin, Y., and Frynta, D. (2006). Cladistic analysis of bantu languages: a new tree based on combined lexical and grammatical data. *Naturwissenschaften*, 93(4):189–194.

Bibliography V

Ringe, D., Warnow, T., and Taylor, A. (2002). Indo-European and computational cladistics. *Transactions of the philological society*, 100(1):59–129.

Savelyev, A. and Robbeets, M. (2020). Bayesian phylolinguistics infers the internal structure and the time-depth of the Turkic language family. *Journal of Language Evolution*, 5(1):39–53.

Sherzer, J. (1976). *An Areal-Typological Study of American Indian Languages North of Mexico.* Amsterdam, Netherlands: North-Holland Publishing Company.

Sicoli, M. A. and Holton, G. (2014). Linguistic phylogenies support back-migration from Beringia to Asia. *PLoS One*, 9(3):e91722.

Siebert, F. T. (1945a). Linguistic Classification of Catawba: Part I. *International Journal of American Linguistics*, 11(2):100–104.

Siebert, F. T. (1945b). Linguistic Classification of Catawba: Part II. *International Journal of American Linguistics*, 11(4):211–218.

Skirgård, H., Haynie, H. J., Hammarström, H., Blasi, D. E., Collins, J., Latarche, J., Lesage, J., Weber, T., Witzlack Makarevich, A., Passmore, S., Maurits, L., Dunn, M., Reesink, G., Singer, R., Bowern, C., Epps, P., Hill, J., Vesakoski, O., Robbeets, M., Abbas, K., Auer, D., Bakker, N., Barbos, G., Borges, R., Danielsen, S., Dorenbusch, L., Dorn, E., Elliott, J., Falcone, G., Fischer, J., Ghanggo Ate, Y., Gibson, H., Göbel, H., G. J., Gruner, V., Harvey, A., Hayes, R., Heer, L., Herrera Miranda, R., Hübler, N., Huntington-Rainey, B., Ivani, J., Johns, M., Just, E., Kashima, E., Kipf, C., Klingenberg, J., König, N., Koti, K., Kowalik, R., Krasnoukhova, O., Lindvall, N., Lorenzen, M., Lutzenberger, H., Martins, T., Mata German, C., Meer, S., Montoya Samamé, J., Müller, M., Muradoglu, S., Neely, K., Nickel, J., Norvik, M., Oluoch, C. A., Peacock, J., Pearey, I., Peck, N., Petit, S., Pieper, S., Poblete, M., Prestipino, D., Raabe, L., Raja, A., Reimringer, J., Rey, S., Rizaew, J., Ruppert, E., Salmon, K., Sammet, J., Schembri, R., Schlabbach, L., Schmidt, F., Skilton, A., Smith, W. D., Sousa, H., Sverredal, K., Valle, D., Vera, J., Voß, J., Witte, T., Wu, H., Yam, S., Ye, J., Yong, M., Yuditha, T., Zariquiey, R., Forkel, R., E. N., Levinson, S. C., Haspelmath, M., Greenhill, S. J., Atkinson, Q. D., and Gray, R. D. (Submitted). Grambank data reveal global patterns in the structural diversity of the world's languages.

Smith, A. D. and Rama, T. (2022). Environmental factors affect the evolution of linguistic subgroups in Borneo. *Diachronica*, 39(2):193–225.

Bibliography VI

Smith, M. R. (2019). *Quartet: comparison of phylogenetic trees using quartet and split measures*. R package version 1.2.5.

Swanton, J. R. (1943). Siouan tribes and the Ohio Valley. *American Anthropologist*, 45(1):49–66.

Voegelin, C. F. (1938). Ofo-Biloxi sound correspondences. In *Proceedings of the Indiana Academy of Science*, volume 48, pages 23–26.

Voegelin, C. F. (1941). Internal relationships of Siouan languages. *American Anthropologist*, 43(2):246–249.

Wichmann, S. and Saunders, A. (2007). How to use typological databases in historical linguistic research. *Diachronica*, 24(2):373–404.

Wiens, J. J. and Moen, D. S. (2008). Missing data and the accuracy of bayesian phylogenetics. *Journal of Systematics and Evolution*, 46(3):307.

Williamson, A. W. (1881). Is the dakota related to the indo european languages? *Journal of the Minnesota Academy of Science*, 2(3):110–143.

Wu, M.-S. and List, J.-M. (to appear). Annotating cognates in phylogenetic studies of south-east asian languages. *Language Dynamics and Change*.

Yanovich, I. (2020). Phylogenetic linguistic evidence and the Dene-Yeniseian homeland. *Diachronica*, 37(3):410–446.

Multistate characters

- In principle, coding typological traits as multistate characters would preclude many of the issues that may be resulting from dependencies.
- The next step is to run the analysis using multistate, polymorphic characters with the `morph-models` (MM) package in BEAST 2.
- MM uses the Mk model and characters with different states can be partitioned into different alignments.
- It would be useful to examine what differences exist between binary and multistate (for discussion, see Rexová et al., 2006; Nichols and Warnow, 2008; List, 2016).

(1) Binarizing the WALS traits

- Binary coding for presence (1) or absence (0) of features with missing or unknown data coded as '?'.
- The WALS traits were coded as closely as possible based on the descriptions on the website.
  - Traits were coded with more than one value if applicable.
- Each distinct value was converted into a site/character:

LANGUAGE	DISTANCE CONTRASTS IN DEMONSTRATIVES		TWO	THREE
Crow	Three-way		1	0
Mandan	Three-way		0	1
Lakota	Three-way	→	0	1
Quapaw	Two-way		1	0
Biloxi	?		?	?

(2) Reducing intra-trait dependencies

- Values that indicate absence (e.g. 'None') were removed and represented with 0.
- Traits with two values were collapsed:

LANGUAGE	DISTANCE CONTRASTS IN DEMONSTRATIVES		TWO/THREE
Crow	Three-way		1
Mandan	Three-way		0
Lakota	Three-way	→	0
Quapaw	Two-way		1
Biloxi	?		?

- But what if Biloxi has neither two or three distance contrasts?

(3a) Removing parsimonious uninformative sites

- Sites that consist of all 0s or all 1s were removed.
  - E.g. `Presence of Common Consonants`, `Tone`, `Locus of Marking in the Clause`, `Associative Plural`
- These sites do not directly inform classification, but their removal gives rise to ascertainment bias: "Acquisition bias is problematic because mean rates of evolution embodied in the branch length parameters will be overestimated if only variable characters are present in the data set [...] such overestimation, if not corrected, would lead to *bias in tree topology inferences*" (Lewis, 2001, 917, emphasis mine).
- Even with ascertainment correction, the topology of the tree does not appear to change.

(3b) Reducing inter-trait dependencies

- Completely predictable or overlapping sites were also removed. For example:
  - The base of gender across the languages is completely predictable from number of genders:

LANGUAGE	NUMBER OF GENDERS	BASE OF GENDER
Crow	None	None
Mandan	Two	Sex-based
Hocank	None	None
Lakota	Two	Sex-based
Quapaw	None	None
Biloxi	Two	Sex-based

(3b) Reducing inter-trait dependencies

- Dependencies that were not completely predictable or overlapping were maintained.

LANGUAGE	ONE STOP SERIES	VOICING
Lakota	0	Fricatives/Plosives
Stoney	0	Fricatives
Biloxi	0	Plosives
Tutelo	1	None
Tutelo	0	Plosives

- Data Set (4) involved both removal of uninformative sites (3a) and reduction of inter-trait dependencies (3b).

(5a) Removing uninformative sites (incl. missing data)

- Missing data (?) are treated as ambiguous states (1s or 0s).
- Sites with missing data that otherwise have all 1s or all 0s were removed since it is likely, although not definitively, that these sites would end up being parsimony uninformative:

LANGUAGE	COMITATIVES AND INSTRUMENTALS
Crow	Differentiated
Stoney	?
Lakota	Differentiated
Osage	?
Biloxi	Differentiated

- It is an empirical question how this type of 'quasi-parsimony uninformative' data impacts topology estimation.

(5b) Omitting 'singleton' sites

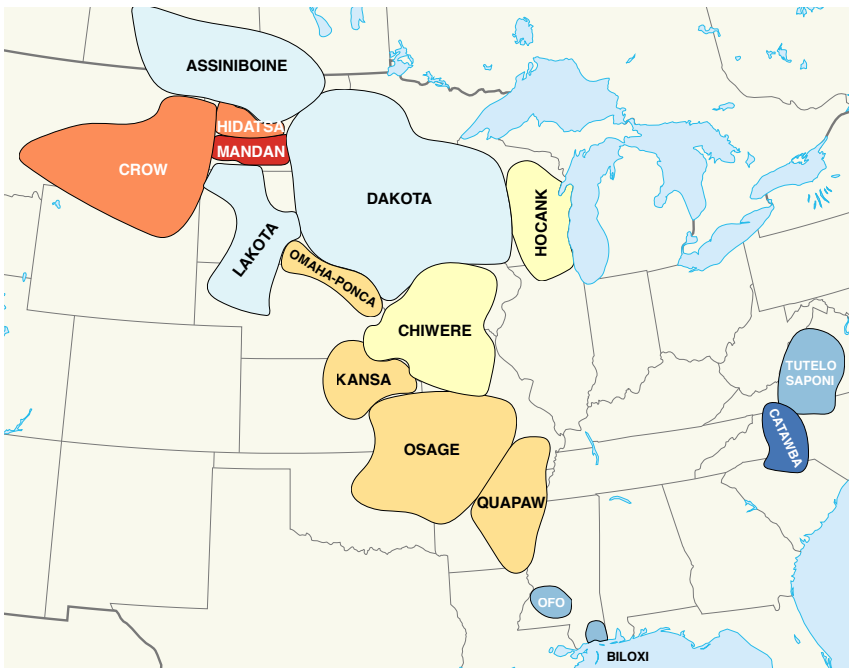
- Sites in which all but one language shares the same value were removed, such as the presence of the velar nasal:

LANGUAGE	VELAR NASAL
Crow	None
Lakota	None
Osage	None
Chiwere	Velar nasal
Biloxi	None

- The assumption is that these sites do not provide (direct) information about the internal structure of the family.
- Data Set (6) involved both removal of uninformative sites that include missing data (5a) and removal of 'singleton' sites (3b).

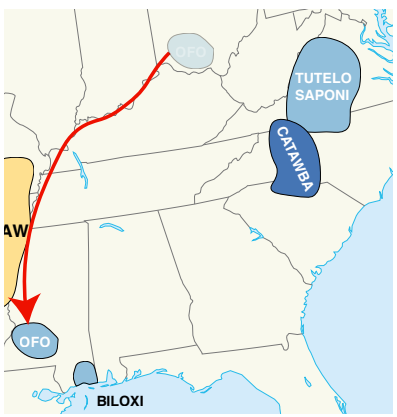
Phylogeny, geography and homoplasy

- The phylogeny reflects traditional classification to a reasonable degree but can they all be explained by geography?
  - Crow, Hidatsa and Mandan are geographically proximate to the Mississippi languages yet are placed separately.
  - Even unlikely subgroups consist of geographically distant languages, e.g. Ohio Valley with Crow, Hidatsa and Mandan.



Phylogeny, geography and homoplasy

- I employed the Mantel test (Mantel, 1967) to calculate correlations between geographical and typological distances.
  - Data Set (4) was not found to be significant ($p = 0.056$, $r = 0.190$), but Data Set (6), which excludes many sites with homogeneous values, was ($p = 0.035$, $r = 0.176$).
  - Using fixed geographical location are somewhat problematic.
- The subgroup consisting of Catawba, Ofo and Tutelo is potentially explained by contact (or at least homoplasy that is compatible with a contact scenario à la Cathcart et al. 2018).
  - According to Swanton (1943), the Ofo migrated in the 17th century from the Ohio Valley to Louisiana.



Possible contact effects in the Southeast

- The unattested language Occaneechi, which is mutually intelligible with Tutelo and was a lingua franca in parts of the Southeast, likely would have played some role in the spread of linguistic traits:
  - "Their [the Indians] Language differs very much [...] However, **they have a sort of general Language** [...] which is understood by the Chief men of many Nations, as Latin is in most parts of Europe, and Lingua Franca quite thro the Levant. **The general Language here us'd, is said to be that of the Occaneeches**, tho they have been but a small Nation, ever since those parts were known to the English" (Beverley, 1705, 23–24, emphasis mine)
- "Catawba grammar and vocabulary show evidence of language mixture, which is not surprising given the number of different groups that ultimately united with the Catawbas. It may, in fact, be the descendant of a creolized language" (Booker et al., 1992, 410)

(2) Reducing intra-trait redundancies

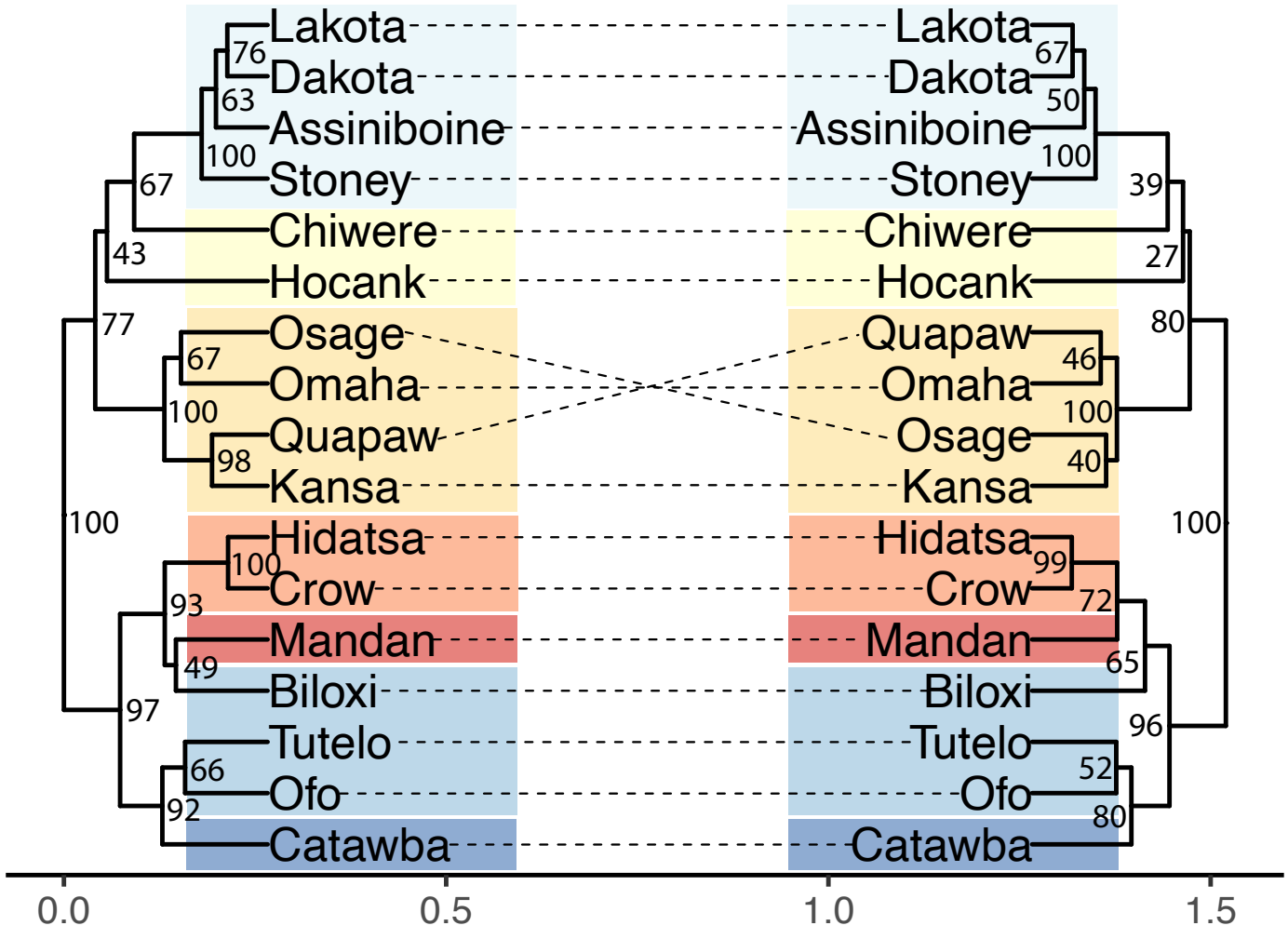


Figure: Summary trees for Analysis (1) (left) and Analysis (2) (right).

(3a) Removing parsimony uninformative sites

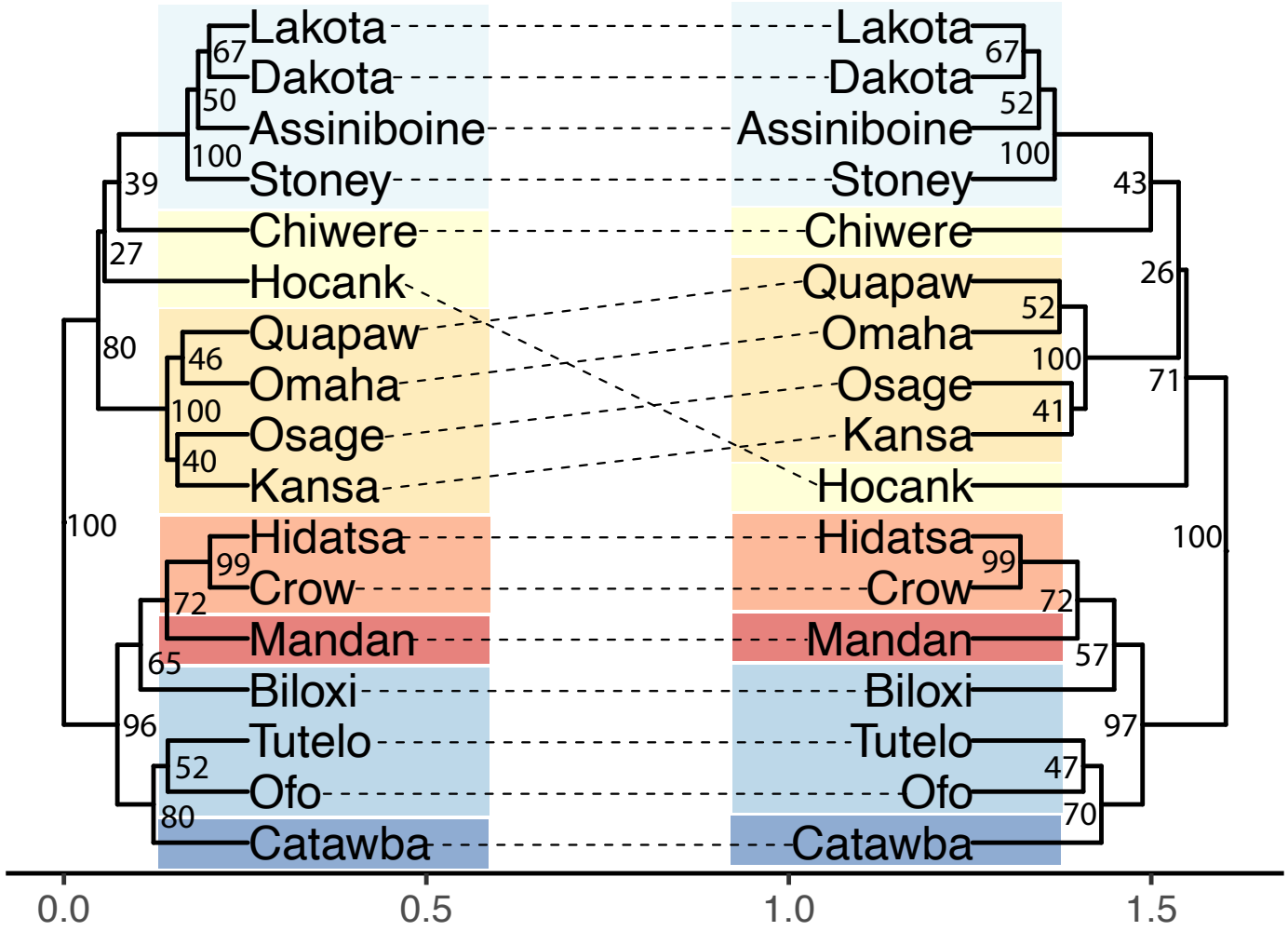


Figure: Summary trees for Analysis (2) (left) and Analysis (3a) (right).

(3b) Reducing inter-trait dependencies

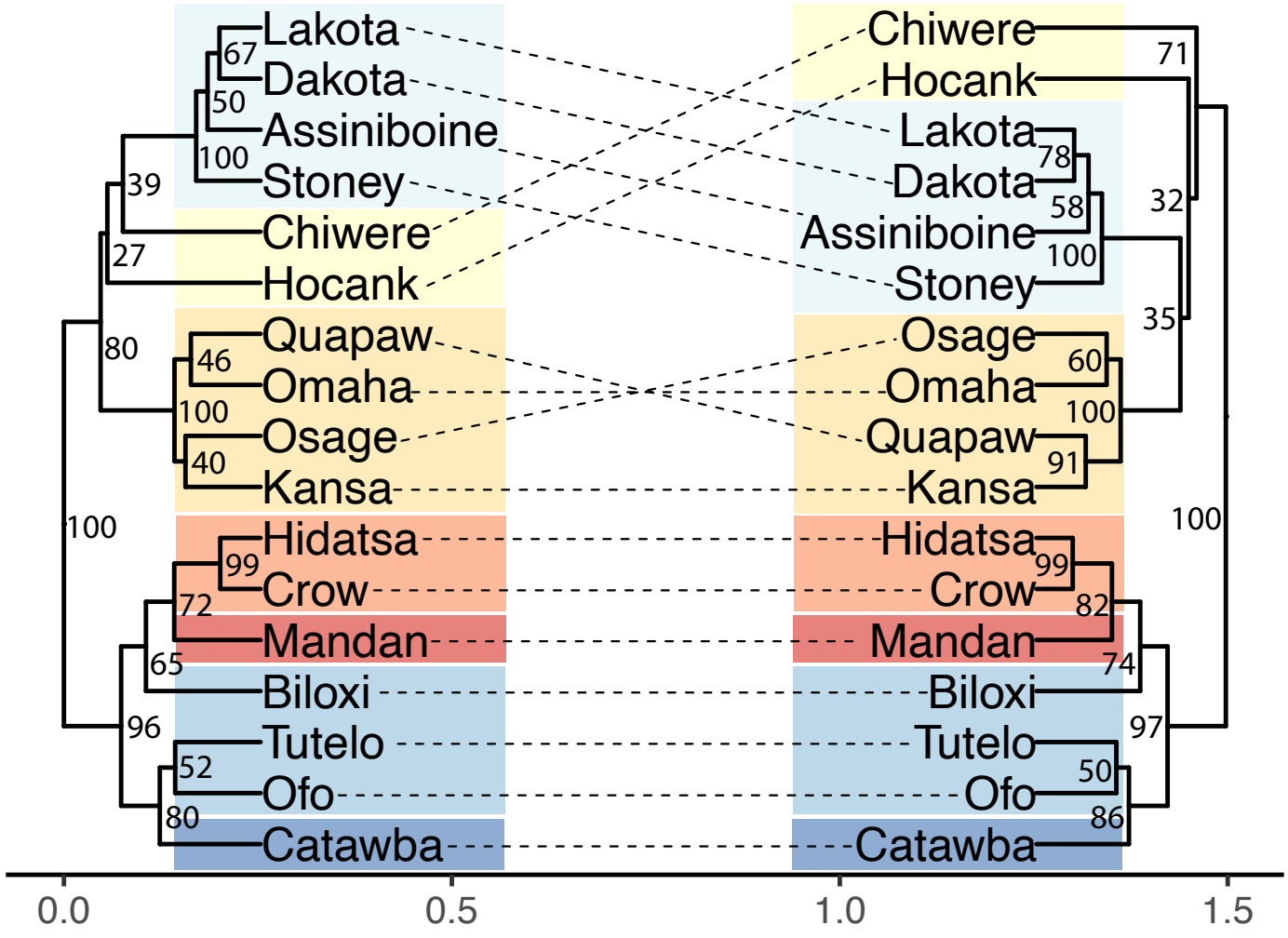


Figure: Summary trees for Analysis (2) (left) and Analysis (3b) (right).

(5a) Removing uninformative sites (incl. missing data)

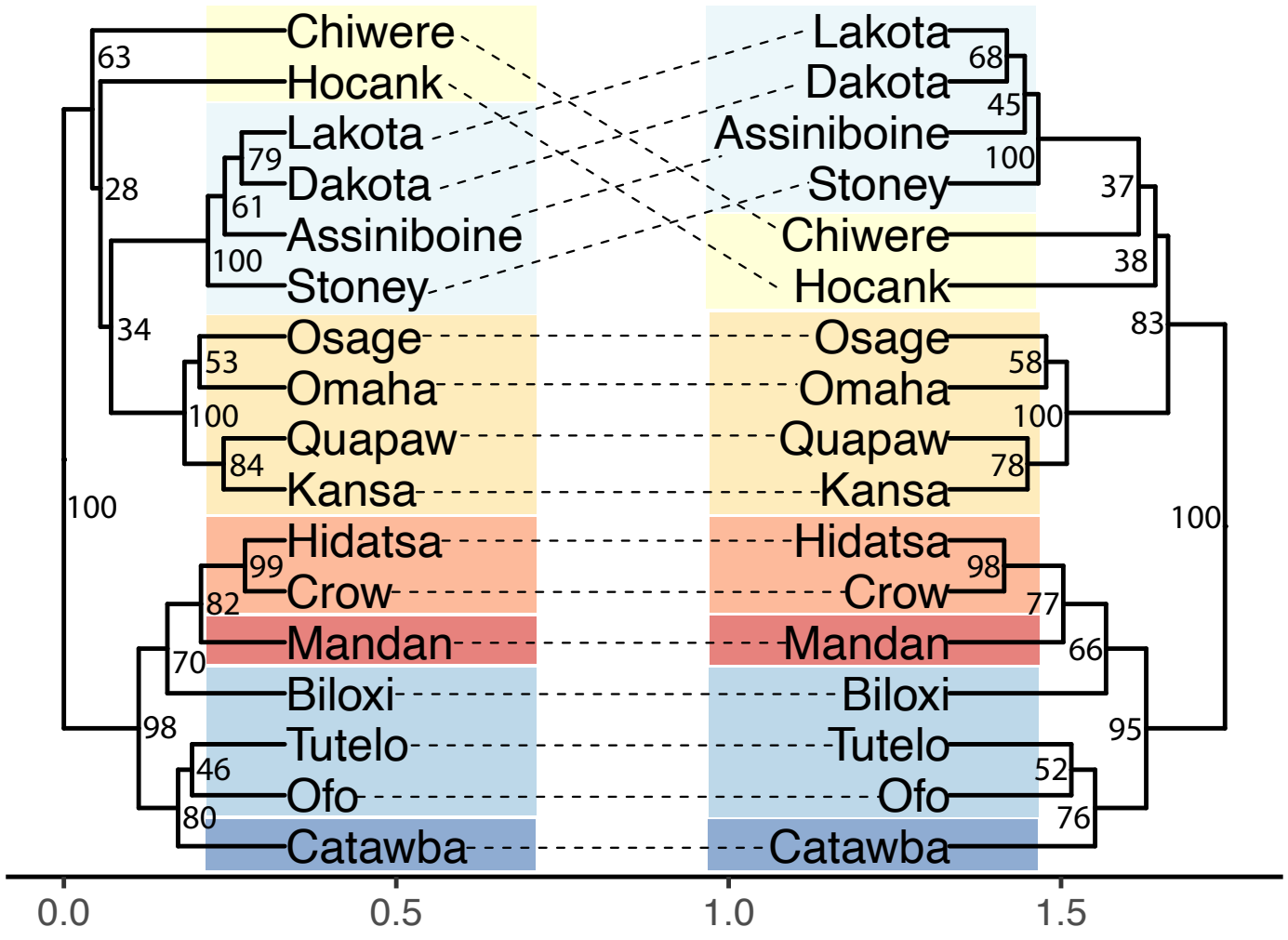


Figure: Summary trees for Analysis (4) (left) and Analysis (5a) (right).

(5b) Omitting 'singleton' sites

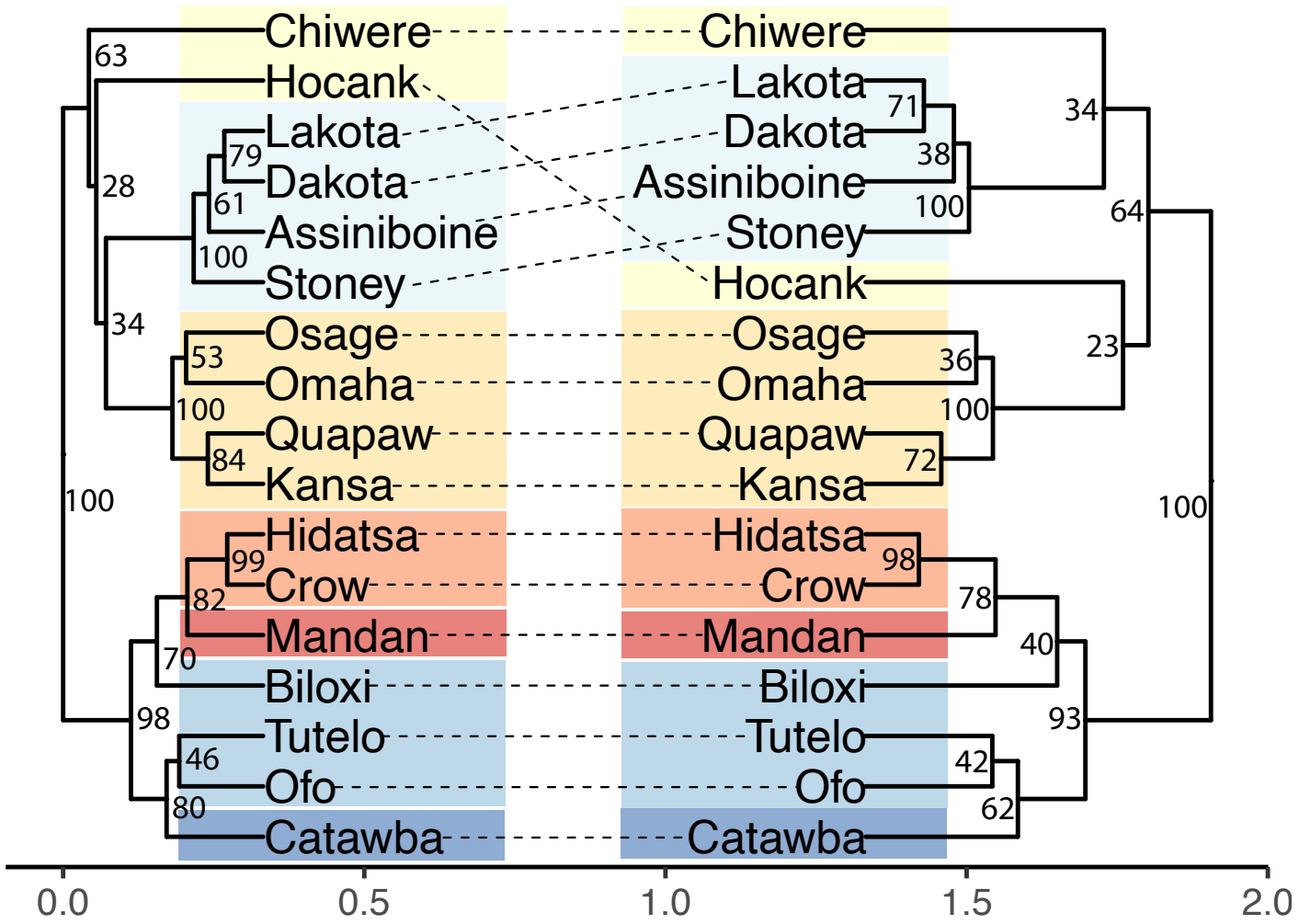


Figure: Summary trees for Analysis (4) (left) and Analysis (5b) (right).

(6) Informative (incl. missing data), non-singleton sites

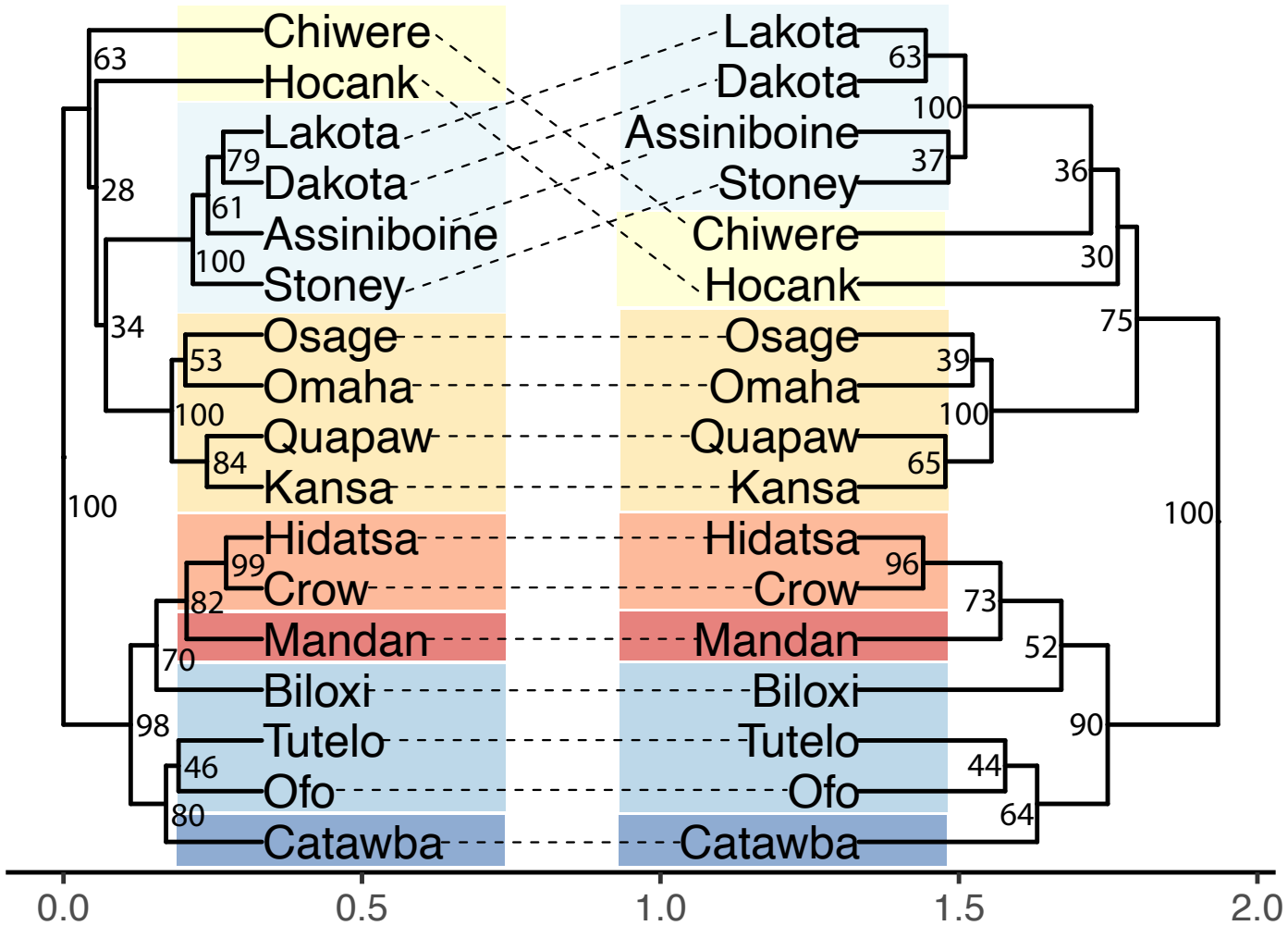


Figure: Summary trees for Analysis (4) (left) and Analysis (6) (right).

Investigating the contributions of specific traits

- I extracted from Data Set (6) all the phonological and morphological (i.e. nominal and verbal) traits and analyzed them using `BEAST 2.6.7`.
  - Even with a small number of sites, we can gain some insights into this question.
  - 50 million generations, relaxed clock, 200+ ESS, the main results were robust to choice of priors and models.
  - Results were comparable even when traits were extracted from Data Set (4).

DATA SET	δ-SCORE	Q-RESIDUAL	SITES	MISSING/ALL (%)
Phonology	0.28	0.0340	36	18/612 (2.9%)
Nominal morphology	0.35	0.0791	49	158/833 (19.0%)
Verbal morphology	0.42	0.0841	61	237/1037 (22.9%)
Morphology	0.37	0.0561	110	395/1870 (21.1%)

Quartet distances from the Rankin tree

ANALYSIS	QUARTET DISTANCE
Rankin 2010	—
Analysis (1)	0.1261
Analysis (6)	0.1261
Phonological only	0.2941
Morphological only	0.3466
Phonology + Morphology	0.1076

Table: Distance from the Rankin tree using the `QuartetDivergence` function from in R package `Quartet` (Smith, 2019).

Analysis using only traits from WALS and Sherzer

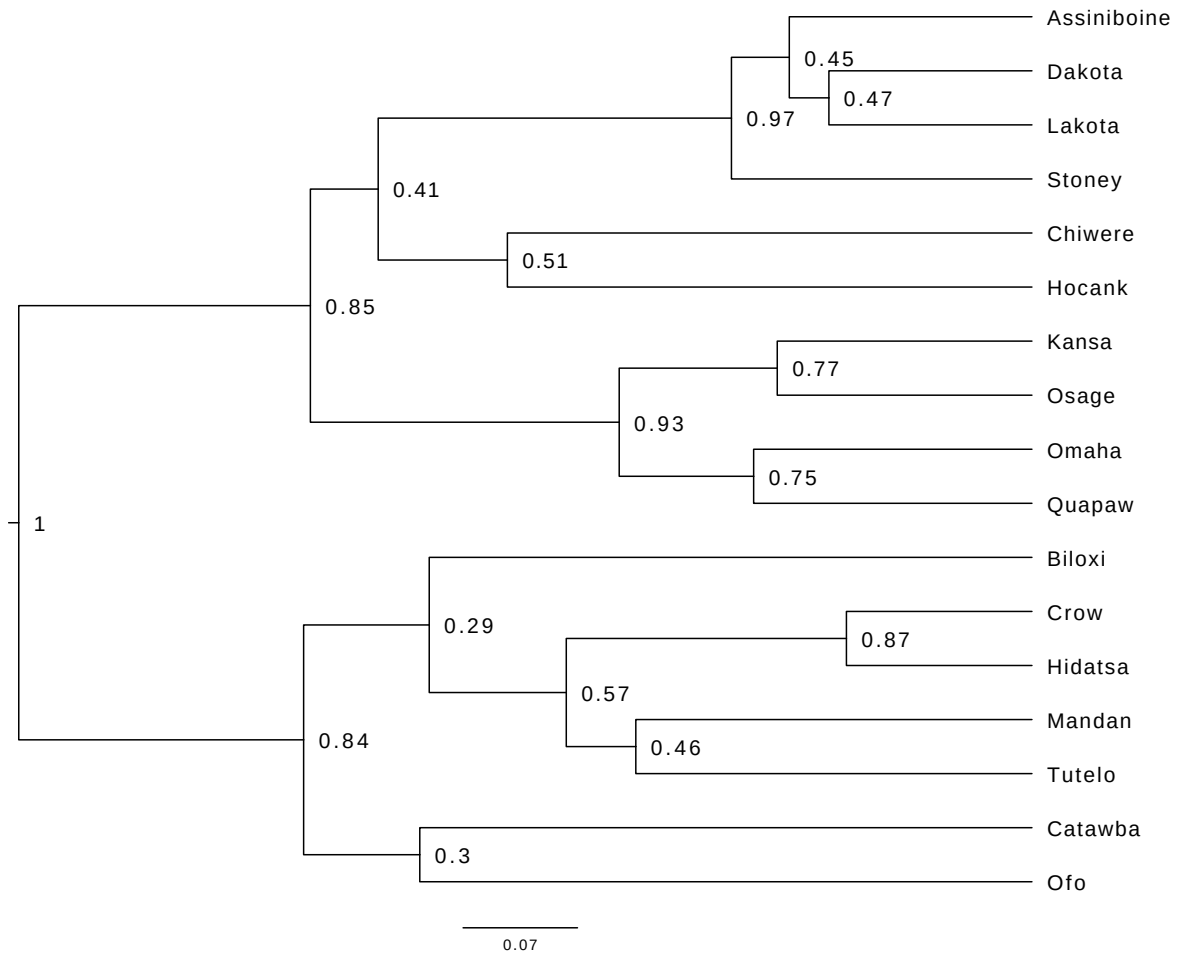


Figure: Summary tree of the data set consisting of only traits from WALS and Sherzer (138 sites).

Investigating the contributions of specific traits

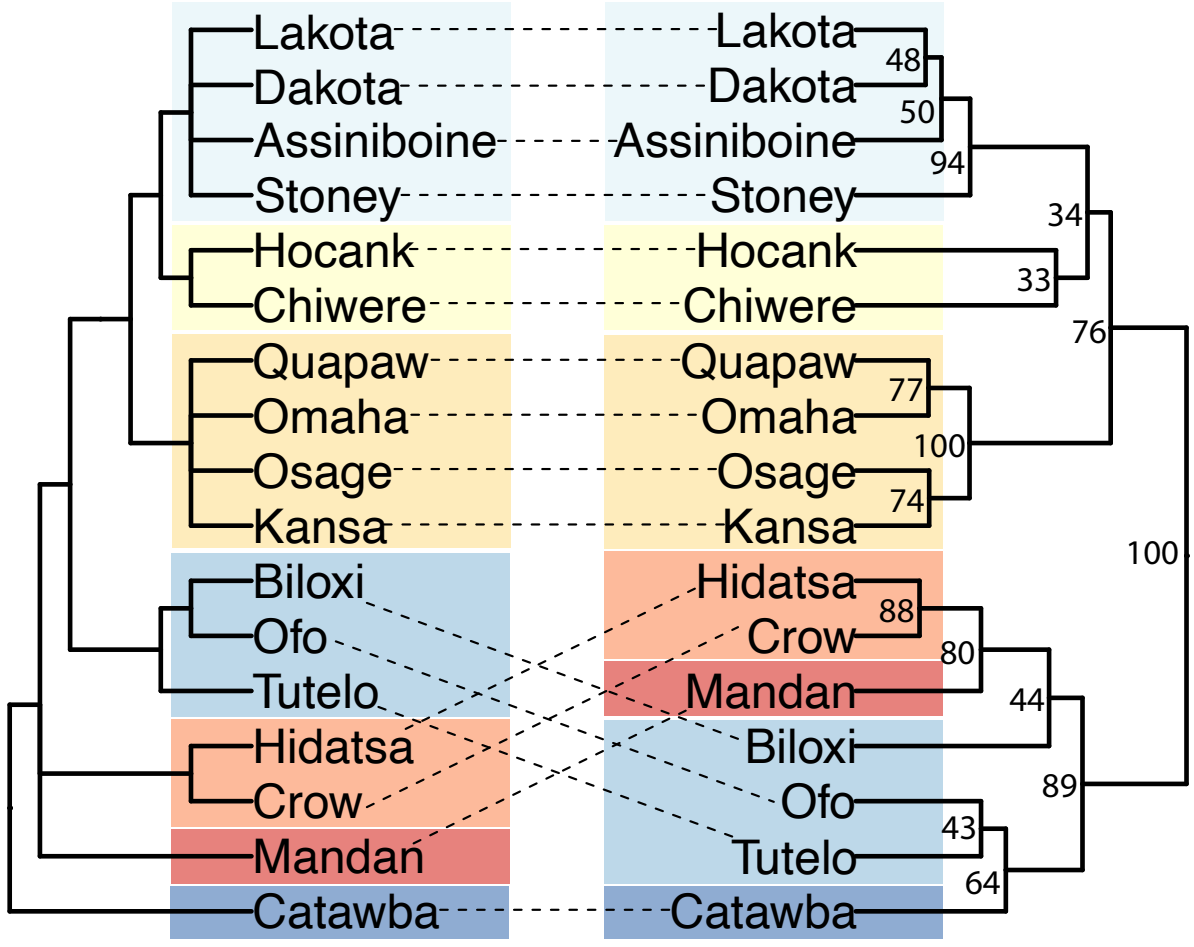


Figure: Phonology and morphology (146 sites) recover most of the inferred tree. Rankin tree (left) and summary tree (right).

List of phonological traits

- Consonant Inventories
- Vowel Quality Inventories
- Consonant-Vowel Ratio
- Voicing in Plosives and Fricatives
- Voicing and Gaps in Plosive Systems
- Glottalized Consonants
- Lateral Consonants
- Vowel Nasalization
- Syllable Structure
- 221
- not aeiou
- vowel length contrast
- one stop series: voiceless
- two stop series: voiceless/voiced
- three stop series: voiceless/voiced/glottalized
- four stop series
- k/č
- one fricative series: voiceless
- two fricative series: voiceless/voiced
- three fricative series: voiceless/voiced/glottalized
- labial fricative
- s/ʃ
- z
- x
- ɣ
- r

List of nominal morphological traits

- Coding of Nominal Plurality
- Occurrence of Nominal Plurality
- Plurality in Independent Personal Pronouns
- Definite Articles
- Indefinite Articles
- Inclusive/Exclusive Distinction in Independent Pronouns
- Inclusive/Exclusive Distinction in Verbal Inflection
- Distance Contrasts in Demonstratives
- Indefinite Pronouns
- Numeral Classifiers
- Position of Pronominal Possessive Affixes
- Possessive Classification
- Adjectives without Nouns
- Order of Demonstrative and Noun
- possessive pronouns independent morpheme
- reduplication = distributive or plual
- animate/inanimate gender
- plural in pronouns
- dual in pronouns
- demonstratives for visible/invisible objects
- numerals classified by form or shape of object

List of nominal morphological traits (cont.)

- GB052 Is there a noun class/gender system where shape is a factor in class assignment?
- GB059 Is the adnominal possessive construction different for alienable and inalienable nouns?
- GB170 Can an adnominal property word agree with the noun in noun class/gender?
- GB171 Can an adnominal demonstrative agree with the noun in noun class/gender?
- GB172 Can an article agree with the noun in noun class/gender?
- GB184 Can an adnominal property word agree with the noun in number?
- GB185 Can an adnominal demonstrative agree with the noun in number?
- GB186 Can an article agree with the noun in number?
- GB187 Is there any productive diminutive marking on the noun (exclude marking by system of nominal classification only)?
- GB188 Is there any productive augmentative marking on the noun (exclude marking by system of nominal classification only)?
- GB204 Do collective ('all') and distributive ('every') universal quantifiers differ in their forms or their syntactic positions?
- GB431 Can adnominal possession be marked by a prefix on the possessed noun?
- GB433 Can adnominal possession be marked by a suffix on the possessed noun?
- GB325 Is there a count/mass distinction in interrogative quantifiers?

List of verbal morphological traits

- Perfective/Imperfective Aspect
- The Future Tense
- The Perfect
- Position of Tense-Aspect Affixes
- The Prohibitive
- Imperative-Hortative Systems
- The Optative
- Situational Possibility
- Epistemic Possibility
- Coding of Evidentiality
- Polar Questions
- Predicative Possession
- Nominal and Locational Predication
- Comparative constructions
- Third Person Zero of Verbal Person Marking
- Order of Person Markers on the Verb
- Reciprocal Constructions
- Passive Constructions
- Antipassive constructions
- Nonperiphrastic Causative Constructions
- Negative Morphemes
- Want' Complement Subjects
- Order of Negative Morpheme and Verb
- locative suffifxes
- locative-directional markers prefix
- locative-directional markers suffix
- GB312 Is there overt morphological marking on the verb dedicated to mood?
- GB519 Can mood be marked by a non-inflecting word ('auxiliary particle')?
- GB520 Can aspect be marked by a non-inflecting word ('auxiliary particle')?

List of verbal morphological traits (cont.)

- GB422 Is there a postposed complementizer for complements of verbs of thinking and/or knowing?
- GB402 Does the verb for 'see' have suppletive verb forms?
- GB403 Does the verb for 'come' have suppletive verb forms?
- GB300 Does the verb for 'give' have suppletive verb forms?
- GB099 Is there verb suppletion for participant person?
- GB081 Is there productive infixation in verbs?
- GB114 Is there a phonologically bound reflexive marker on the verb?
- GB120 Can aspect be marked by an inflecting word ('auxiliary verb')?
- GB127 Are different posture verbs used obligatorily depending on an inanimate locatum's shape or position (e.g. 'to lie' vs. 'to stand')?
- GB151 Is there an overt verb marker dedicated to signaling coreference or noncoreference between the subject of one clause and an argument of an adjacent clause ('switch reference')?
- GB177 Can the verb carry a marker of animacy of argument, unrelated to any noun class/gender of the argument visible in the NP domain?
- GB298 Can standard negation be marked by an inflecting word ('auxiliary verb')?
- GB324 Is there an interrogative verb for content interrogatives (who?, what?, etc.)?
- GB152 Is there a morphologically marked distinction between simultaneous and sequential clauses?
- GB146 Is there a morpho-syntactic distinction between predicates expressing controlled versus uncontrolled events or states?
- GB117 Is there a copula for predicate nominals?