Российский государственный гуманитарный университет
Russian State University for the Humanities

Russian State University for the Humanities
Institute of Linguistics of the Russian Academy of Sciences

Journal of Language Relationship

International Scientific Periodical

№ 18/4

Moscow 2020

Российский государственный гуманитарный университет
Институт языкознания Российской Академии наук

Вопросы языкового родства

Международный научный журнал

№ 18/4

Москва 2020

Advisory Board:
H. EICHNER (Vienna) / Chairman
W. BAXTER (Ann Arbor, Michigan)
V. BLAŽEK (Brno)
L. HYMAN (Berkeley)
F. KORTLANDT (Leiden)
A. LUBOTSKY (Leiden)
J. P. MALLORY (Belfast)
A. YU. MILITAREV (Moscow)
V. F. VYDRIN (Paris)

Editorial Staff:
V. A. DYBO (Editor-in-Chief)
G. S. STAROSTIN (Managing Editor)
T. A. MIKHAILOVA (Editorial Secretary)
A. V. DYBO
† S. V. KULLANDA
M. A. MOLINA
M. N. SAENKO
I. S. YAKUBOVICH

Founded by Kirill BABAEV

© Russian State University for the Humanities, 2020

Редакционный совет:
Х. Айхнер (Вена) / председатель
В. Блажек (Брно)
У. Бэкстер (Анн Арбор)
В. Ф. Выдрин (Париж)
Ф. Кортландт (Лейден)
А. Лубоцкий (Лейден)
Дж. Мэллори (Белфаст)
А. Ю. Милитарев (Москва)
Л. Хайман (Беркли)

Редакционная коллегия:
В. А. Дыбо (главный редактор)
Г. С. Старостин (заместитель главного редактора)
Т. А. Михайлова (ответственный секретарь)
А. В. Дыбо
† С. В. Кулланда
М. А. Молина
М. Н. Саенко
И. С. Якубович

Журнал основан К. В. Бабаевым

© Российский государственный гуманитарный университет, 2020

Вопросы языкового родства: Международный научный журнал / Рос. гос. гуманитар. ун-т; Рос. акад. наук. Ин-т языкознания; под ред. В. А. Дыбо. — М., 2020. — № 18/4. — xiii + 87 с.

Journal of Language Relationship: International Scientific Periodical / Russian State University for the Humanities; Russian Academy of Sciences. Institute of Linguistics; Ed. by V. A. Dybo. — Moscow, 2020. — No. 18/4. — xiii + 87 p.

ISSN 2219-3820

http://www.jolr.ru/
gstarst1@gmail.com

Дополнительные знаки: С. Г. Болотов
Add-on symbols by S. G. Bolotov

Подписано в печать 10.12.2020. Формат 60×90/8.
Усл. печ. л. 12,5. Уч.-изд. л. 10,0. Заказ № 1126.
Тираж 1050 экз.

Издательский центр
Российского государственного гуманитарного университета
125993, Москва, Миусская пл., 6
www.rggu.ru
www.knigirggu.ru

Table of Contents / Содержание

Articles / Статьи

Contributors

Rafael ANDRADE — graduate student, University of São Paulo (Brazil), rafael.andrade.almeida@usp.br

Anton KOGAN — doctor of sciences (Philology), head of Department of Asian and African languages, Institute of Oriental Studies of the Russian Academy of Sciences (Moscow), kogan_anton@yahoo.com

Andrey NIKULIN — Laboratório de línguas e literaturas indígenas, University of Brasilia, andre.n.guzman@gmail.com

Mikhail N. SAENKO — candidate of sciences (Philology), researcher, Institute of Slavic studies of the Russian Academy of Sciences, Moscow, veraetatis@yandex.ru

Andrei SHATSKOV — candidate of sciences (Philology), researcher, Institute of Linguistic studies of the Russian Academy of Sciences (Saint-Petersburg), andrey.shatskov@gmail.com

Mikhail E. VASILYEV — junior researcher, Institute of Linguistics of the Russian Academy of Sciences, Moscow, mvhumanity@gmail.com

Сведения об авторах

АНДРАДЕ, Рафаэль — студент Университета Сан-Паулу (Бразилия), rafael.andrade.almeida@usp.br

ВАСИЛЬЕВ, Михаил Евгеньевич — мл. науч. сотрудник Института языкознания РАН (Москва), mvhumanity@gmail.com

КОГАН, Антон Ильич — доктор филол. наук, зав. Отдела языков Азии и Африки Института востоковедения РАН (Москва), kogan_anton@yahoo.com

НИКУЛИН, Андрей Владимирович — сотрудник лаборатории по изучению индейских языков и литератур, Бразильский университет, andre.n.guzman@gmail.com

САЕНКО, Михаил Николаевич — канд. филол. наук, науч. сотрудник Института славяноведения РАН (Москва), veraetatis@yandex.ru

ШАЦКОВ, Андрей Владимирович — канд. филол. наук, науч. сотрудник Института лингвистических исследований РАН (Санкт-Петербург), andrey.shatskov@gmail.com

Note for Contributors

Journal of Language Relationship welcomes submissions from everyone specializing in comparative-historical linguistics and related disciplines, in the form of original articles as well as reviews of recent publications. All such submissions should be sent to the managing editor:

G. Starostin
Institute for Oriental and Classical Studies
Russian State University for the Humanities
125267 Moscow, Russia
Miusskaya Square, 6
E-mail: gstarst1@gmail.com

Articles are published preferably in English or Russian, although publication of texts in other major European languages (French, German, etc.) is possible. Each article should be accompanied with an abstract (not exceeding 300 words) and keywords.

For more detailed guidelines on article submission and editorial policies, please see our website at: http://www.jolr.ru or address the editorial staff directly at gstarst1@gmail.com.

Будущим авторам

Журнал *Вопросы языкового родства* принимает заявки на публикацию оригинальных научных статей, а также рецензий от всех, кто специализируется в области сравнительно-исторического языкознания и смежных дисциплин. Рукописи можно высылать непосредственно заместителю главного редактора по адресу:

125267 Москва
Миусская площадь, д. 6
Российский государственный гуманитарный университет
Институт восточных культур и античности
Г. Старостину
E-mail: gstarst1@gmail.com

Предпочтительные языки публикации — английский или русский, хотя возможна также публикация статей на других европейских языках (французский, немецкий и т. п.). К каждой статье обязательно прикладывается резюме (не более 300 слов) и список ключевых слов.

Подробнее о требованиях к оформлению рукописи, редакционной политике журнала и т. п. вы можете узнать на нашем сайте по адресу: http://www.jolr.ru или же непосредственно, обратившись к редакции по электронной почте (gstarst1@gmail.com).

In memory of Sergey Kullanda (1954–2020)

It is with the deepest regret that we announce the loss of our friend and colleague Sergey Vsevolodovich Kullanda, one of the long-term members of the Editorial Board of the Journal of Language Relationship; he passed away on November 30, 2020, after a long and difficult illness.

Sergey Kullanda was first and foremost a historian, specializing in such areas as the history of Indonesia and the ancient Scythian culture. But he had always shared a deep interest in all things related to the history of languages, meaning that most of his research was tied to linguistic issues one way or another.

His studies of ancient social history, driven by a desire to "get to the foundation of each social phenomenon" (p.c.) in the pre-written period, quite logically caused Sergey to fully embrace comparative-historical linguistics. Linguistic data were already quite prominent in his early works on the history of Java and Austronesian studies in general; later on, his focus largely shifted to Indo-European studies and comparative Iranistics, where he used the methodology of linguistic reconstruction to combine anthropological data with the results of studies in diachronic semantics. This was most prominent in his works on the Indo-European system of kinship terms, the main point of those being that the primary semantics of the studied terms must have originally referred not so much to actual kinship as to their social functions, distributed across various age and sex groups. Future research will no doubt show just how close this point of view was to the actual truth; but in any case, Kullanda's interdisciplinary hypotheses were quite immaculately constructed both from the point of view of social anthropology and comparative Indo-European linguistics, and will be highly relevant for subsequent generations of scholars in those fields.

Another of Sergey's favorite topics concerned the role of comparative-historical linguistics in elucidating the ancient history of Iranian and Indo-Iranian peoples. Here must be mentioned his paper on Proto-Indo-Iranian and Proto-Iranian homelands, with a clear and meticulous focus on any potential linguistic contacts between speakers of these protolanguages and other linguistic units; and, most importantly, his 2016 monograph "The Scythians: language and ethnogenesis" (in Russian), in which he has summarized most of the preceding research on Scythian culture, provided a critical evaluation of all pieces of evidence on the Scythian language, established a firm position for Scythian on the genealogical tree of Iranian languages, and scrutinized all possible linguistic contacts between Scythian and the other ancient languages of West Asia and Eastern Europe.

All of these works testify to Sergey Kullanda's huge potential — much of it, alas, unrealised — as a prominent researcher in comparative-historical linguistics, making his passing a truly significant loss for the scientific community at large. This is not even to mention, of course, the significance of this loss for all of Sergey's friends, colleagues, and disciples — many of the latter still remember with fondness his lectures on Sanskrit at the Russian State University for the Humanities. Sergey was one of those true friends one could always count on, and, among other instances, he proved this while serving on the Editorial Board of JLR, never downplaying the importance of even the most routine technical work on submitted papers and sacrificing his personal needs so that the next number of the Journal could be published on time. He will be sorely missed, but never forgotten.

Anna Dybo
on behalf of the Editorial Board of The Journal of Language Relationship

For all those interested in works by Sergey Kullanda, a complete list (in Russian) may be found on the website of the Institute of Oriental Studies of the Russian Academy of Sciences: https://www.ivran.ru/persons/SergeyKullanda.

Памяти С. В. Кулланды (1954–2020)

30 ноября 2020 г. после тяжелой продолжительной болезни умер наш друг и коллега, член редакционной коллегии журнала «Вопросы языкового родства», Сергей Всеволодович Кулланда.

Сергей Всеволодович был в первую очередь историком, специалистом по древней Индонезии и по скифологии. Но интерес и внимание к истории языков, к древним языковым свидетельствам придавали его работам значительную лингвистическую составляющую.

Его занятия древней социальной историей, желание «доискаться до того, как возникает то или иное явление» (цитата из его устного выступления) в дописьменные периоды существования человечества совершенно логично привели Сергея Всеволодовича в объятия сравнительно-исторического языкознания. Языковые данные активно использовались и в его ранних работах по яванской истории и в целом австронезистике. Впоследствии он серьезно углубился в индоевропеистику, в сравнительно-историческую иранистику и использовал в своих исследованиях методы лингвистической реконструкции, в первую очередь стараясь сопрягать семантическую реконструкцию с антропологическими данными. Так сделаны его работы по индоевропейской терминологии родства; основной их посыл — семантика исследуемых терминов должна была в древности относиться не к родственным, а к социальным (в первую очередь половозрастным) отношениям. Будущие исследования покажут, насколько такие построения приближаются к истине; но, во всяком случае, междисциплинарные конструкции С. В. Кулланды были выстроены весьма добросовестно и с точки зрения социальной антропологии, и с точки зрения использования данных индоевропейской этимологии, и разобранный в них материал может быть использован в дальнейшей работе.

Другая область его исследований, связанная со сравнительно-историческим языкознанием — древняя история иранцев и индоиранцев. Здесь прежде всего следует отметить статью об индо-иранской и иранской прародинах, в которой с особой тщательностью разбираются возможности праязыковых контактов для праиндоиранцев и праиранцев; и, главное, вышедшую в 2016 году монографию «Скифы: язык и этногенез», в которой подводятся итоги предыдущих скифологических исследований, критически оцениваются все имеющиеся свидетельства, касающиеся скифского языка, устанавливаются особенности скифского языка в отличие от других восточноиранских и его место на генеалогическом древе иранских языков, а также внимательно рассматриваются все возможные контакты скифского языка с другими древними языками Западной Азии и Восточной Европы. Книга снабжена отработанным с филологической и этимологической точек зрения словариком скифских слов, что делает ее особенно значимой для последующих исследований в этой области.

Все эти работы позволяют оценивать научную деятельность С. В. Кулланды как значительный вклад в сравнительно-историческое и историческое языкознание, а его болезнь и смерть — как большую потерю для научного сообщества.

Это огромная потеря и для всех друзей, сотрудников и учеников Сергея Всеволодовича. Он был очень хорошим, отзывчивым, дружелюбным человеком. Долгое время он преподавал санскрит и ведийский язык в РГГУ, и его студенты вспоминают его с нежностью. Так же помнят его и все «взрослые» друзья. На его помощь можно было рассчитывать в любой сложной ситуации. В частности, его работа в редколлегии «Вопросов языкового родства» содержала элементы самопожертвования — он никогда не отказывался даже от вполне технических занятий по вычитке статей, откладывая собственные дела ради своевременного выхода номера журнала.

Мы с ним дружили. Он катал моего старшего сына на лошадке, когда тому было 4 года. Мне нравилась его старомодная вежливость по отношению к женщинам (подавать пальто, уступать сиденье, наливать вино...). Наши попытки писать совместные работы в основном потерпели фиаско (кроме критических), но мы с удовольствием подолгу обсуждали работы друг друга и взаимно старались прислушиваться к критике. Известие о его болезни, а потом и смерти проделало очередную черную дыру в моей модели мира.

А. В. Дыбо

Список основных трудов С. В. Кулланды можно найти на сайте Института востоковедения РАН: https://www.ivran.ru/persons/SergeyKullanda.

Anton Kogan
Institute of Oriental Studies, Russian Academy of Sciences, Moscow; kogan_anton@yahoo.com

Notes on the historical phonology of Indo-Iranian loanwords in Northwestern Tibetan dialects

Recent research has shown that the vocabulary of certain Northwestern Tibetan dialects contains a significant number of Indo-Iranian loanwords. It is, however, still unclear if these loanwords have been borrowed from a single or from several sources, and whether their presence is the result of substratum or adstratum interference. Likewise, the exact genetic position of the donor-language(s) within the Indo-Iranian group so far remains undetermined. The study of all these issues should, no doubt, be based on facts of historical phonology. In this article I attempt to identify the most conspicuous historical-phonological features of the Indo-Iranian elements found in Northwestern Tibetan varieties. Furthermore, I make some preliminary conclusions concerning the linguistic geography of the region in the pre-Tibetan period, as well as the direction and relative chronology of Tibetan migrations.

Keywords: historical phonology; language contact; linguistic substratum; Indo-Iranian languages; Dardic languages; Tibetan dialects; Ladakhi language; Balti language; Purik language.

A recent etymological study of the vocabulary of certain Northwestern Tibetan varieties, namely Ladakhi, Balti and Purik, has revealed the existence of a significant Indo-Iranian lexical stratum in these dialects (Kogan 2019). This discovery is to some extent in line with the hypothesis put forth more than a century ago by the German scholar August Hermann Francke, and still popular among Tibetologists, according to which the pre-Tibetan population of Ladakh and adjoining areas was somehow related to the present-day speakers of Dardic languages (Francke 1907). It should, however, be kept in mind that the overall picture of ethnic and linguistic history of the region is still far from clear. The presence of Indo-Iranian loanwords in the Tibetan dialects of Ladakh and Baltistan raises more questions than it answers. We do not know if these loanwords have been borrowed from a single source or several sources, whether their existence is the result of substratum or adstratum interference,[1] and finally, which branch or branches of Indo-Iranian they represent.

It goes without saying that the answers to these questions must be based on facts of historical phonology. The only possible way to establish the exact genetic position of the donor language is to compare its historical-phonological peculiarities with those of all the four known branches of the Aryan subfamily, namely Indo-Aryan, Iranian, Dardic and Nuristani. The strongest indication of borrowing from several related lects is, no doubt, the presence of more than one type of phonological development in the same position of the word, inexplicable by secondary processes like analogical changes. The choice of substratum or adstratum alternative can hardly be made either without taking into account historical-phonological data. In our case, such a choice is actually a complex task. The present-day Indo-Iranian neighbors of Northwestern Tibetan dialects are two Dardic languages: Kashmiri and Shina.[2] There are

[1] Based on our present knowledge of language situation in the Tibetan Empire and later northwestern Tibetan kingdoms, the influence of an unknown Indo-Iranian superstrate on Tibetan dialects should be considered improbable.

[2] The Ladakhi and Purik dialects are also in contact with Brokskat, a Dardic language genetically close to Shina. However, as shown in the above-cited article (Kogan 2019), the speakers of this language must be comparatively recent migrants to their present habitat.

strong reasons to believe that both of them spread to their present-day speaking areas in the Middle Ages due to ethnic migrations. The population of Kashmir until the 13th century spoke an Indo-Aryan language, most likely closely related to West Pahari dialects (Kogan 2016), whereas in the now Shina-speaking valleys of Karakoram the language spoken one thousand-plus years ago was probably Burushaski (Jettmar 1975). It should be born in mind that Indo-Iranian influence on Tibetan varieties under study, whatever its nature, cannot be exactly dated, even though the Tibetan conquest of Ladakh and adjoining areas in the 8th century A.D. provides a *terminus post quem* for the process. In light of this fact, it cannot be ruled out that at least a part of loanwords were borrowed in the 2nd millennium A.D. directly from Shina and/or Kashmiri. Another possible source of lexical loans may have been East Iranian Saka dialects spoken before the 11th century A.D. in the western part of the present-day Xinjiang, immediately to the north-east of Ladakh and Baltistan.[3] To sum up, borrowing from an ad-strate language should be considered probable if a particular Indo-Iranian word shows Indo-Aryan, East Dardic (Shina or Kashmiri)[4] or Iranian historical-phonological traits.

In the following sections an attempt will be made to identify the most conspicuous historical-phonological features of Indo-Iranian elements found in Northwestern Tibetan dialects.[5] First, I shall address the development of vowels, syllabic sonorants and some sequences containing vowels and sonorants, then the development of single consonants and consonant clusters. Finally, some preliminary conclusions will be drawn.

Vowels and syllabic sonorants

Vowels (monophthongs and diphthongs)

There is no phonological vowel length in Northwestern Tibetan dialects. That is why PII **a* and **ā* are usually reflected in the same way:

1. Ladakhi ***asur*** 'mustard seed as a spice used in Ladakhi pickle' (Norman 2010: 1073).[6] Cf. OIA *āsurī* 'Sinapis ramosa', Sindhi *ahuri* 'mustard seed', Lahnda *ôhur*, *ahũr*, *āhur*, Punjabi *āhur*, Hindi-Urdu *āsurī* 'mustard', Kashmiri *āsoru* 'plant, Sinapis ramosa'.

2. Balti ***basanda*** 'dandelion' (Sprigg 2002: 27). Cf. OIA *vāsanta-* 'vernal, pertaining to spring', OIA *vasanta-*, Shina *bazōn*, Phalura *basānd*, Bashkarik *basan,* Torwali *basān,* Kalasha *bāsun,* Khowar *bosun*, Pashai *wahán*(*d*), Gawar-Bati *wasand* 'spring'.

3. Balti ***bat*** 'boiled mixture of germinated grain flour and ordinary flour (made during Ramzan)' (Sprigg 2002: 27). Cf. OIA *bhakta-* 'food; boiled rice', Khowar *bot* 'evening meal', Kalasha *batay* 'flour taken to be eaten in high pastures' (Trail 1999), Shina *bat*, Bashkarik *batt,* Torwali *bāt,* Kashmiri *batɨ*, Lahnda, Punjabi *bhatt*, Hindi-Urdu, Nepali, Gujarati, Marathi *bhāt* 'boiled rice', Wakhi *bat* 'wheat flour gruel', Ishkashimi *bat*, Shughni *bāt*, Munji *bātək* 'ritual food made of wheat flour and butter, cooked in water or milk'.[7]

[3] It is, however, worth noting that a tentative attempt to find Irainan phonological features in Ladakhi, Balti and Purik words of Indo-Iranian origin made in Kogan 2019 was not successful.

[4] The existence of the East Dardic subbranch consisting of Kashmiri, Shina and Kohistani languages was first hypothesized by George Abraham Grierson (1906; 1919). For the latest views on this group and its historical-phonological features see Kogan 2015; Kogan 2016.

[5] The material studied here as well as lexical comparisons are drawn chiefly from Kogan 2019, where the respective etymologies are also discussed. If a word not mentioned in this article is analyzed, its etymology will be discussed in the present text.

[6] For ease of reference, etymologies are numbered consecutively.

[7] The above-cited East Iranian words are most probably borrowed from some Dardic source.

4. Balti, Purik ***chal*** 'overflow, spill over', Ladakhi (Leh and Shamskat dialects) ***chal-ces*** 'to splash, to spill over' (Sprigg 2002: 41; Norman 2010: 297). Cf. OIA *kṣarati, kṣalati* 'flows, trickles', *kṣālayati* 'washes', Proto-Iranian **xšar-* 'to flow' (> Persian *šārīdan* 'to trickle', *(āb)šār* 'waterfall', Ossetic *äxsärdzän* 'waterfall' (< **xšar-čana-*), Middle Persian *Xšart* 'the river Jaxartes' (Livshits 2003)), Kashmiri *čhalun* 'to wash', *čhar* 'a sprinkle of water etc. from the fingers'), Gawar-Bati *ṭchār* 'rapids in a stream', Phalura, Indus Kohistani *ṭchār*, Bashkarik, Kalasha *uṭchār*, Shina *ṭchar* 'waterfall'.

5. Ladakhi ***darak***, Purik ***deraq*** 'stiff, hard' (Norman 2010: 438) < PII **dhāraka-*. Cf. Kashmiri *dor* 'firm, hard, strong, compact, durable, solid', *darun* 'to become steady (of something in motion), to become firm, to stand steady' (Grierson 1915–1932: 238, 246), Pashai *d(h)ar-*, Wotapuri *dar-*, Phalura *dháara-* 'to remain, stay', Torwali *dērī* 'they remained', Khowar *dorik* 'hold back, wait, keep', Hindi, Punjabi *dharnā* 'to keep', Gujarati *dharvũ* 'to hold, catch', OIA *dhārayati, dharati* 'holds, keeps', Av. *dāraiiehi* '(you) hold'.

6. Balti ***gzar*** 'to flow' (Sprigg 2002: 72), Purik ***zar*** 'id.', Ladakhi ***zar-ces, dzar-cas*** 'to drip, to run down, to trickle out' (Norman 2010: 790). Cf. Av. *γžar-* 'to flow', Ossetic *ğzælyn* 'to pour down, drip', OIA *jhara-* 'waterfall', *jharī* 'river', Prakrit *jharaï* 'drips', Hindi *jharnā* 'to ooze, trickle away' < PII **gjhar-* < PIE **$d^hg^{wh}er$-* (Cheung 2007: 124) or **$g^ugh'er$-* (LIV: 213–214).

Etymological **u* and **ū* also merge:

7. Balti, Purik, Ladakhi ***mulṭuk***, ***multuk*** 'fist' (Norman 2010: 705; Sprigg 2002: 118) < **mulṭak* < **muṣṭaka-*.[8] Cf. OIA *muṣṭi-*, Av. *mušti-*, Khotanese *muṣṭu*, Sindhi *muṭhi*, Lahnda, Punjabi *muṭṭh*, Hindi-Urdu, Gujarati, Marathi *mūṭh*, Nepali *muṭhi*, Persian *mošt*, Shughni *mut*, Wakhi *məst*, Shina *muṭ(h)*, Kashmiri *mŏṭh*, Phalura, Kalasha, Khowar *muṣṭi*, Gawar-Bati *muṣṭāk*, Burushaski (loanword) *muč̣* 'fist'.

8. Balti ***mulak***, ***mulu***, Purik ***mulaq*** 'turnip' (Sprigg 2002: 118; Zemp 2018: 944). Cf. Burushaski *múlo*, Shina *muúlo* 'id.', OIA *mūla-* 'root', *mūlaka-* 'radish', Shina *mŭlī*, Khowar *mūḷ* 'root', Pashai *mūluk*, Hindi-Urdu *mūlī* 'radish' < PII *mūla-* 'root'.[9]

No reflexes of PII initial or medial **i* and **ī* have been attested in our material. There is, however, one instance of the drop of ** ī* in the word-final position (see **1.**).

As we can see, Balti shows double reflexes of PII word-final **a*. This vowel is sometimes preserved (cf. ***basanda*** 'dandelion') and sometimes dropped (cf., e.g. ***bat*** 'boiled mixture of germinated grain flour and ordinary flour (made during Ramzan)'). It cannot be ruled out that the difference in reflexes is due to different borrowing sources.

In one example, we find an unusual vowel sequence *aa* in Balti, which may correspond to short *a* in Old Indo-Aryan as well as in Iranian and Dardic:

9. Balti ***baan*** 'man or men who sing religious songs and foretell the future' (Sprigg 2002: 24). Cf. Kashmiri *wan-*, Indus Kohistani *ban-* 'to say', Kalasha *bandek* '1. to teach; 2. To announce so as to inaugurate, to order or command' (Trail 1999: 27), OIA *vandate* 'praises, worships', Av. *vaṇd-*, Khotanese *van-*, Parthian *wynd-* 'to praise, honor, worship'.

It is, however, quite possible that the source of the Balti word reflects some lengthened-grade derivative of the above-cited root. Such derivatives are attested in several Dardic languages. Cf., e.g. Woṭapuri *-bān* '*sprechend*, *Sprecher*' in *alik-bān* '*Lügner*' (*alik* '*Lüge*') with the medial *ā* being the regular continuant of the etymological **ā* (Buddruss 1960: 20, 87, 92).

Before a final nasal the change **a > o* takes place:

10. Balti (Skardu dialect) ***dom*** 'sadness, trouble, difficulty, adversity' (Norman 2010: 493). Cf. Shina *dămizhār* 'adversity, trouble' (Bailey 1924), Burushaski *dʌm(i)jar* 'trouble, inconven-

[8] For more details on the change **ṣṭ > lṭ* in Northwestern Tibetan dialects see Kogan 2019.

[9] Probably, related to Middle High German *mūl*, German *Maul* 'muzzle' (Mayrhofer 1996: 369).

ience, worry' (Lorimer 1938),[10] OIA *damayati* 'tames, subdues', Ossetic *domyn* 'to tame; exhaust; demand'.

11. Balti ***chon*** 'vain' (Sprigg 2002: 43), Purik ***chon*** ("tʃʰon") 'useless, in vain, unfounded' (Zemp 2018: 109), Ladakhi ***chon*** 'spontaneously, for no reason, gratuitously; in vain, for nothing, to no avail, useless; free, for free, at no cost' (Norman 2010: 310). Cf. Bashkarik *ṭhan*, Phalura *čheṇiko* (*e* < *a* in i-umlaut position), Shina *ṭhon*, Kashmiri *čhon* (< **čhanu* with *o* < *a* in u-umlaut position), Gawar-Bati *ṭhēnika* (*e* < *a* in i-umlaut position), Pashai (Kurangali dialect) *čhāni* (*ā* < *a*) 'empty, void', Burushaski *ṭhan* (Hunza), *ṭan* (Yasin) '*leer (Schüssel)*, *unterbeschäftigt, ohne Arbeit*, *frei*' (Berger 1998: 106).[11]

See also **48**.

This change does not occur before clusters and the historical vowel **i*. The etymological *a* is preserved in these positions:

12. Balti ***zan-zos***, ***dzan-zos*** (Turtuk dialect), Purik ***zan-zos*** 'wife; family' (Norman 2010: 842; Sprigg 2002: 180).[12] Cf. Phalura *jeeni* 'female person' (Liljegren, Haider 2011: 76), Kashmiri *zən*', Sindhi, Lahnda, Punjabi *jaṇī* 'woman', Bashkarik *jin kar-* 'to marry', Indus Kohistani *zhāl* 'marriage' (Zoller 2005: 202), OIA *jani-*, Av. *ǰaini-* 'woman, wife'.

See also **2.**

The Proto-Indo-Iranian diphthong **au* is monophthongized (**au* > *u*):

13. Ladakhi, Purik ***kulik*** 'lock; key; the joint of the jaw' (Norman 2010: 12) < PII **kaula-*, **kaulikā-* 'curved' (= Proto-Iranian **kaura-* > Khotanese *kūra-* 'crooked' (Bailey 1979: 62)). Cf. Tirahi *kṓolə*, Pashai *kōlā*, Shumashti *kolāṇṭa*, Khowar *koli*, Bashkarik *kōl*, Torwali *kōl*, Phalura *kūulo*, Shina *kōlu̯* 'curved, crooked'[13].

I.2. Syllabic sonorants

The Proto-Indo-Iranian syllabic sonorant **r̥* is vocalized in interconsonantal position:

14. Ladakhi ***kit-ces*** 'to catch, seize, capture' (Norman 2010: 82). Cf. Bashkarik *gītə* 'took, bought', Woṭapuri *gat* 'took', Gawar-Bati *gūtím* 'I took' < **gr̥pta-* < **gr̥b-ta-* (Vedic *gr̥bdha-*, Av. *gərəpta-* 'taken').

15. Ladakhi ***shen-ces*** 'to squash' (Norman 2010: 986). Cf. OIA *śr̥ṇāti* 'crushes, breaks', Av. *a-sarə-ta-* 'not discouraged (lit. 'not broken')', Khowar *šeník* 'to crush' < PIE **k'erh₂-* '*brechen*, *zerbrechen* (intr.)' (LIV: 327–328).

[10] The Burushaski word is likely to have been borrowed from Shina.

[11] In theory, there is a possibility that the Balti and Purik words have been borrowed from Shina, whereas the Ladakhi word is a loan from Kashmiri. Should this be the case, the Ladakhi example cannot illustrate the aforesaid historical phonological process, because in Kashmiri the change **a* > *o* was caused by *u*-umlaut rather than the following nasal. Semantic data, however, show that separate borrowing of the lexeme under analysis is rather unlikely. In all the three Tibetan varieties, this lexeme has undergone the same semantic development ('empty' > '(in) vain'). Technically, it cannot be ruled out that such a development took place independently in all the dialects, but it seems more probable that it occurred either in the donor language before borrowing or after borrowing but before the divergence of Ladakhi, Balti and Purik. In either case, the source language could hardly have been Kashmiri. In the latter language the development of *u*-umlaut dates back not earlier than to the 17th century (Kogan 2016), i.e. to a period when linguistic Tibetanization of not only Ladakh but also Baltistan was, no doubt, an accomplished fact. This said, direct borrowing from Kashmiri into Balti is scarcely possible, because these two languages, being geographically separated by the Shina-speaking area, are not in close contact with each other.

[12] This word is most probably a compound. Its second component may be connected with Tibetan *tshos* 'color' (Kogan 2019).

[13] For the semantic change 'crooked' > 'key' cf. Greek κληῖς, Latin *clāvis*, Old Church Slavic *kl'učь* (ключь) 'key' < PIE **klēu-* '*Haken*, *krummes Holz*' (Pokorny 1959: 604–605).

The vowel *e* in the latter example may have been a free allophone of *i* in the donor language. Cf. the free variation of *i* and *ĕ* after *š* in Kashmiri.

I.3. Vowel sequences and sequences with intervocalic sonorants

Secondary vowel sequences which appear due to deletion of intervocalic consonants contract to monophthongs.

**a+*u > o*:

16. Ladakhi ***po-ze*** 'ram, full-grown male sheep' (Norman 2010: 540) < **pau-* < PII **paśu-* 'cattle' (> OIA *paśu-*, Av. *pasu-* 'id.', Pashto *psə*, Ossetic *fɨs* 'sheep'). [14]

**a+*i > e*:

17. Ladakhi ***tshele*** 'thatched roof, shade made of branches and leaves, thatched-roofed house or hut' (Norman 2010: 773), Balti ***tshele*** 'hut' (Sprigg 2002: 170) < **sčaila-* < PII **śćad-ila-*. Cf. OIA *chādayati* 'covers', *chadis-* 'cover, roof', Lahnda, Punjabi *chatt*, Hindi *chat* 'roof', Kashmiri *chĕy* 'a kind of thatching grass' (Grierson 1915–1932: 1066), Proto-Iranian **sād-* 'to cover' (> Pashto *psoləl* 'to adorn' from **upa-* or **pati-sād-*) < PIE **sk'ed-*.

Sequences containing intervocalic sonorants, both historical and secondary, yield monophthong vowels.

**aya > e*:

18. Ladakhi ***shen*** (Leh dialect), ***ṣen*** (Shamskat dialect) '(wooden) floor' (Norman 2010: 986) < PII **śrayana-* 'leaning, foothold'. Cf. OIA *śray-*, Av. *sray-* 'to lean'.

**āwa > o*:

19. Ladakhi ***phok*** 'incense or burning juniper leaves or other fragrant burning materials' (Norman 2010: 580) < PII **pāwaka-* 'purifier'. Cf. OIA *pāvayati* 'purifies', Persian *pāk* 'pure' (< **pāwaka-*).

**āiya > e:*

20. Balti ***be-kar*** 'court singer and dancer who improvises poems and songs' (Sprigg 2002: 28), Ladakhi ***be-da*** 'member of the caste that used to be itinerant musicians', ***be-mo*** f. (Norman 2010: 640–641) < **bāiya-* < PII **wād(i)ya-* 'music, musical instrument', cf. OIA *vādya-* 'musical instrument, music', *vādyakara-* 'musician', *vādayati* 'plays a musical instrument', Punjabi *vajjā* 'musical instrument', Hindi-Urdu *bājā* 'music', Kashmiri *waz-* 'to sound (of bell, clock etc.)', *wāy-* 'to play a musical instrument', Shina (Guresi) *baž-* 'to strike (of a gong etc.)', Ossetic *wadynʒ* 'flute, panpipe' < **vādəničī*.

I.4. Vowel assimilation

Vowel assimilation, both progressive and regressive, was noted:

21. Balti ***men-ze*** 'lump of dough' (Sprigg 2002: 113), Ladakhi ***men-ze*** 'ball of dough ready to be shaped and cooked, flattened dough ball, or shaped bread ready to be cooked' (Norman 2010: 710). Cf. Burushaski *máano* '*grösserer Teigklumpen*' (Berger 1998: 272), Shina *míno* 'id.', Brokskat *manīli*, Dameli *man* 'bread', Bashkarik *man*, Phalura *māṇḍili* 'very soft bread' (Morgenstierne 1940; 1941), Phalura *māṇḍ-* 'to knead', Romany *ma(n)ro* 'bread', Sindhi *mānī* 'bread, loaf, food', Punjabi *maṇḍā* 'a thick cake', West Pahari (Bhalesi) *mánni* 'a large cake', Assamese *mar-* 'to knead (dough)', Maithili *māṛ-* 'to knead', OIA *mr̥dnāti* 'crushes, kneads, rubs', Av. *mōrəndən* '(they) destroy'.

[14] As can be seen, the development of this sequence differs from that of the Proto-Indo-Iranian diphthong **au*. This fact implies that the two sound changes could hardly be simultaneous. The element *-ze* in the Ladakhi word is historically a suffix found also in Classical Tibetan (Kogan 2019).

22. Balti ***monṭhok*** ‘clod of earth’ (Sprigg 2002: 118) < **maṭ-ṭhok*? Cf. OIA *mṛttikā-* ‘earth, clay’, Hindi-Urdu, Punjabi *miṭṭī*, Nepali *māṭo*, Dameli *maṭhi* ‘id.’, Phalura *mēṭhi*, Shina *māṭi* ‘clay’; Burushaski *ṭhóko* ‘*Klumpen*, *Ballen* (*von Zucker*, *Salz*, *Butter*)’ (Berger 1998: 452), Indus Kohistani *ṭhokh* ‘a clod (earth, salt)’ (Zoller 2005: 220).

See also **7**.

Consonants

Single consonants

The devoicing of the historical initial voiced consonants, characteristic of Central and Upper Ladakhi dialects including Leh, affected several Indo-Iranian loanwords:

23. Ladakhi ***pat, paddi, patsi*** ‘totally, completely’ (Norman 2010: 533) < **bad.* Cf. OIA *baddha-* ‘bound’, Gujarati *bādhũ* ‘whole, entire’, West Pahari (Kotgarhi) *baddhɔ* ‘all, entire, (pl.) all together’, Kashmiri *bod* ‘handful’.

24. Ladakhi ***tul*** ‘powdered dung’ (Norman 2010: 448) < **dul*. Cf. OIA *dhūli-* ‘dust, powder’, *dhūlikā-* ‘pollen, fog, mist’; Prakrit *dhūlī*, Hindi *dhūl*, Punjabi, Lahnda, Gujarati *dhūṛ*, Sindhi *dhūṛi*, Kumauni *dhuli*, Bengali *dhul*, Marathi *dhūḷ* ‘dust’; Nepali *dhulo* ‘dust, powder’; Tirahi "*dùda*", Kalasha *udhrũ*, Shina *ŭdū*, Phalura *dúṛi* ‘dust’; Pashai (Wegali dialect) *duṛī* ‘dust-storm’; Torwali *dur* ‘mist’.

See also **14.**

The above examples give reason to believe that the process of borrowing may have begun before and finished after the devoicing of the initial mediae (Kogan 2019).

When not devocalized, the etymological initial voiced aspirated obstruents lose aspiration and merge with their plain voiced counterparts. Thus, **bh* > *b*, **dh* > *d.* See **3, 5.**

The process of deaspiration, followed by dentalization and assibilation, seems to have affected the Proto-Indo-Iranian affricate **jh (*jh* >**j* > *(d)z)*: ***gzar*** ‘to flow’ < PII **gjhar-* (see **6**).

25. Another probable instance of this sound change is Ladakhi ***zanggi*** ‘a species of tiny flying biting insect’ (Norman 2010: 825). This word is likely to reflect some derivative of PII **jhan-* ‘to strike, kill, injure’ (> OIA *han-*, Av. *jan-* ‘id.’). Cf. the semantic development of this root in some Iranian languages, e.g. Talysh *žan-* ‘to bite, sting (of insects and snakes)’ (Rastorguyeva, Edelman 2007: 136). If we assume the same development for the Indo-Iranian lect from which the Ladakhi noun has been borrowed, we may consider the word in question to be the reflex of PII **jhan-aka-* with the meaning ‘sting, stinger’ to which the adjectivizing suffix *-ī* (< PII **-in/-ī*) has been added. This suffix was widely used, e.g. in Old Indo-Aryan, to form adjectives with a possessive sense, often prone to nominalization: *keśin-*, Nom Sg *keśī* ‘long-haired’ (cf. *keśa-* ‘hair’), *pakṣin-* Nom Sg *pakṣī* ‘winged; bird’ (cf. *pakṣa-* ‘wing’), *hastin-* Nom Sg *hastī* ‘possessing a hand; elephant’ (cf. *hasta-* ‘hand’). The original meaning of the source form of the Ladakhi word should thus have been “possessing a sting”. The elision of the interconsonantal short *a* (**jhanaka-* > **zang-*) is not uncommon in some Dardic languages (cf., e.g. Pashai *šūŋg* ‘dog’ < **śunaka-*, *sāŋg* ‘earth’ < **samaka-*). This sound change must have resulted in the formation of the secondary cluster *ng*, which in turn must have precluded the change *a* > *o* in the initial syllable.

Other etymological palatal affricates are also dentalized word-initially (**č* > *c* (=*ts*),[15] **j* > *(d)z*):

[15] Balti, Purik and Ladakhi words are transcribed using the standard Tibetological transcription system, where *c* and *ch* stand for the voiceless palatal and voiceless palatal aspirated affricate respectively, while their den-

26. Ladakhi ***tsapik*** 'a little, a bit, a little while' (Norman 2010: 731) < PII **čap-* 'to catch, snatch, pick, pinch', cf. Shina *čap-*, Indus Kohistani *cap-*, Gawar-Bati *cep-* 'to bite', Pashai *čip-* 'to bite off', Kashmiri *cop* 'a bite', Balochi *čāmp-* 'to snatch', Ossetic *cæfsyn* 'to stick, glue', Munji *cəb-* 'to pluck'.

See also **12**, **17**.

The initial palatal affricates in turn usually correspond to retroflex affricates or their regular reflexes in the Dardic languages (see **4**, **11**).

The two above types of correspondences may have resulted from a kind of affricate shift involving the dentalization of the historical palatal affricates and the concomitant palatalization of the retroflex ones. A similar historical-phonological process has affected the affricate system of Kashmiri (Kogan 2016).

It is worth noting that the initial palatal affricates in Burushaski loanwords are sometimes preserved as such:

27. Ladakhi ***cancil***, ***chancil*** 'the green outer shell or fruit of walnut' (Norman 2010: 267). Cf. Burushaski *ćhanjíl* (Nagir, Hunza), *ćanjíl* (Yasin)[16] 'die grüne äussere Schale der Walnuss, Häutchen zwischen den Teilen des Walnusskerns', Shina *ćhaćíil* 'id.' (Berger 1998: 96).

28. Balti ***cangti*** 'drop', Purik, Ladakhi (Shamskat and Leh dialects) ***cangti*** 'leak in a roof' (Norman 2010: 267). Cf. Burushaski *ćhaṅ man-*, Shina *ćhaćhaṅ-* '(*Wasser*) *sickern*, *herabfliessen*' (Berger 1998: 97).

29. Balti ***cha*** 'millet' (Sprigg 2002: 41), Ladakhi (Shamskat dialect) ***cha*** 'a variety of millet, a cereal grain which was grown in Ladakh in the past, but very little now' (Norman 2010: 290). Cf. Burushaski *ćha* (Nagir, Hunza), *ća* (Yasin) '*Hirse*, *Kolbenhirse*, *Setaria italica*' (Berger 1998: 95).

Etymological intervocalic stops, both voiceless and voiced, are dropped in a number of cases:

30. Balti (Skardu dialect) ***spa*** 'taste, tasty (neutral), (sexual) enjoyment', Purik, Ladakhi (Nubra dialect) ***spa*** 'taste, flavour' (Norman 2010: 542; Sprigg 2002: 159) < PII * *swāda-*. Cf. OIA *svāda-* 'taste', Iranian **χʷāda-* (> Persian *χvā* 'good taste', Balochi *wād* 'salt'), Shina *ispāvu* 'tasty'.

31. Ladakhi ***perak*** 'Ladakhi women's head-dress, covered with turquoise and coral' (Norman 2010: 539) < PII **paridhāka-*. Cf. OIA *paridhā-* 'to put on (clothes)', Sindhi *paharaṇu*, Nepali *pairanu*, Hindi *pahirnā* 'to put on, wear', Kashmiri *pārun* 'to put on; to adorn, ornament', Khowar *purduik* 'to cover oneself, put on a cloak'.

32. Ladakhi ***shanṭi*** 'a leafy vegetable' (Norman 2010: 975) < PII **śāka-* 'green vegetable'. Cf. OIA *śāka-*, Shina, Indus Kohistani *šā*, Phalura *šō*, Bashkarik *ša*, Kashmiri *hākh*, Kalasha *šak*, Khowar *šax*, Waigali, Dameli *cā* 'id.', Indus Kohistani *šʌṛ̃ī* 'a green vegetable with round leaves' (Zoller 2005: 378) < PIE **k'ēko-* (Mayrhofer 1996: 628).

See also 17, 20, 48.

Intervocalic voiceless stops, however, are sometimes preserved or sonorized. Such instances require a separate analysis. The retention of intervocalic *k* is observed exclusively in probable reflexes of the reconstructed formations with the suffix **-k-*, i.e. masculine nouns in **-aka-* or feminine nouns in **-ikā-*:

33. Ladakhi ***kuruk*** 'donkey foal; foal' (Norman 2010: 12), Balti ***bong-kúru*** 'donkey colt' (Sprigg 2002: 30) < **kurtaka-*. Cf. Kalasha *kúḷak (<* **kuṛak)* 'child (male or female); the offspring of a human or animal' (Trail 1999), Shumashti *kuṛ*, Dameli *kuŕa* 'child', Ashkun *kūṛə* 'child, foetus', Kati *kŕu, kuŕuk* 'young of animals', Prasun *kyüru* 'young of animals, child' < **kurtaka-*, Kurdish *kurr* 'son', Middle Persian *kurrak*, Persian *kurra* 'foal' < **kurna(ka)-*.

---

tal counterparts are conveyed by *ts* and *tsh*. In Dardic and Iranian examples, as well as in Proto-Indo-Iranian reconstructions, *č, čh* mark the palatal affricates, while *c, ch* are used for the dental ones.

[16] In Berger's notation *ć* stands for the voiceless palatal affricate.

See also **5**, **7**, **8**, **13**, **19**, **31**, **48**.

All such cases can be easily explained, if we assume that the suffix *-*k*- retained its productivity in the donor language for a long time and was perceived as a separate morpheme when intervocalic consonants were dropped. In this connection, an interesting fact is that this formant has probably been suffixed to some roots of Burushaski origin:

34. Balti ***kulak*** 'meal (quickly made mixture of buttermilk and flour)' (Sprigg 2002: 94), Purik ***kholak*** 'a certain dish', Ladakhi (Leh, Shamskat and Nubra dialects) ***kholak, qholak*** 'ready-to-eat dough of roasted flour' (Norman 2010: 117). Cf. Burushaski *(d)-q(h)ul-an-* (Hunza, Nagir), *d-χul-an-* (Yasin) '(*Teig*) *kneten*' (Berger 1998: 357).

The sonorization of the historical intervocalic *t* can be observed after a non-etymological nasal, or perhaps nasalization:

35. Balti ***rindi*** 'lead, bullet' (Sprigg 2002: 139), Ladakhi ***rindi*** 'bullet; lead (metal)' (Norman 2010: 911) < PII **rīti-* 'flowing, melting, (metal) casting' Cf. Bashkarik *rīd*, Torwali *žit* (*ž* < *r*) 'brass', Shina *rīl* 'brass, bronze, copper', Gawar-Bati *rīt* 'copper', OIA *rīti-* 'stream; yellow brass, bell-metal'.

The development of secondary nasals and nasalization is a well-attested phenomenon in a number of Dardic, Indo-Aryan and East Iranian languages.

In the language under study rhotacism, i.e. the change of PIE **l* to *r*, seems not to have been a regular process in intervocalic position. Like in Nuristani, Dardic and most Indo-Aryan dialects, the distribution of reflexes of the two sonorants is not always clear. Possible, albeit not always provable cases of retention of PIE **l* as well as lambdacism (the change of PIE **r* to *l*) have been attested:

36. Ladakhi ***puli***, ***polo*** 'Ladakhi biscuits of a particular type' (Norman 2010: 541). Cf. OIA *pūra-* 'cake', *pauli-* 'a cake of scorched grain and ghee', Sindhi, Punjabi, Hindi, Kumauni *pūrī*, Gujarati, Marathi *purī* 'fried cake', Kashmiri *pūr* 'a kind of cake fried in ghee' < PIE **pūr-* 'corn, wheat' (Gamkrelidze, Ivanov 1984: 657).

See also **4**,[17] **8**.

PII **s*, both prevocalic (word-initial) and intervocalic, is preserved:

37. Ladakhi ***sale*** 'knitting needle' (Norman 2010: 1005). Cf. Phalura *silēni* 'needle', OIA *sīvyati* 'sews', Khotanese *hīya* 'sewn stuffs', Ossetic *x*^{w}*ɨyɨn* 'to sew'.[18]

See also **2**.

PII initial **ś* (< Proto-Aryan[19] **ć* < PIE **k'*) is retained:

38. Balti ***shang*** 'wisdom, sense' (Sprigg 2002: 151)[20], Purik ***šaŋ*** 'consciousness' (Zemp 2018: 931), Ladakhi ***shang*** 'alertness, awareness, caution, prudence' (Norman 2010: 553). Cf. Burushaski (loanword) *šʌŋ* 'awake, aware; care, heed, attention' (Lorimer 1938: 322), Shina *šo̯ṅ* 'care, anxiety; awake, alert' (Bailey 1924), Khowar *šaṅg* 'fear, suspicion' (Morgenstierne 1973), OIA *śaṅkā-* 'fear, distrust', *śaṅkate* 'is afraid, distrusts' < PIE **k'enk-* '*in der Schwebe sein, hängen* (intr.)' (LIV: 325).

See also **15, 32.**

[17] Possibly, PII **kšar-* (>Balti, Purik, Ladakhi ***chal*** 'overflow, spill over', see **4**) is related to PII **gjhar-*, reflecting another variant of the same Proto-Indo-European root (Rastorgueva, Edelman 2007).

[18] Burushaski *sel* 'Nadel, Stecknadel' (Berger 1998b: 377) may have been borrowed from the same Indo-Iranian source as the Ladakhi word.

[19] The terms "Aryan" and "Indo-Iranian" are not used as synonymous in the present work. Following D.I. Edelman (1992), we believe that the split of the Aryan unity began with the separation of the Nuristani branch. The remaining three branches, i.e. Indo-Aryan, Iranian and Dardic, form the Indo-Iranian unity, whose disintegration took place at a later date. Thus, Indo-Iranian is considered a lower-level taxon within the Aryan subfamily.

[20] In traditional Tibetological transcription *sh* stands for voiceless palatal sibilant.

As for intervocalic PII *ś, there are possible cases of its retention as well as elision:

39. Balti ***lashi*** 'a resinous wood used as a candle or torch because it burns slowly' (Sprigg 2002: 98). Cf. Kashmiri *làsh*^{i} 'a torch' (Grierson 1915–1932: 533), Shina (Gilgit dialect) *lăi* 'torch (unlit)' (*i* regularly < *ś in intervocalic position).

See, by contrast **16.**

Since the Indo-Iranian etymology of the above word for 'torch' is unclear, and no apparent cognates outside Dardic have been found for it so far, this example should be considered doubtful.

PII intervocalic *š[21] (=OIA *ṣ*) > *y*:

40. Balti ***múyu***, ***myo*** 'mouse' (Sprigg 2002: 119). Cf. OIA *mūṣ-*, *mūṣā-*, *mūṣika-* 'mouse, rat', Persian *mūš*, Pashto *məžak*, Ossetic *myst*, Pashai *mūč*, Shumashti *múṣo*, Gawar-Bati *muṣa*, Torwali *mūṣ*, Phalura *mũṣo* 'mouse', Bashkarik *mūṣ* 'mouse, rat', Shina (Guresi dialect) *mūẓṵ* 'rat', *mŭẓai* 'mouse', West Pahari (Jaunsari) *mūśā*, Romany *mušó* 'mouse', Kumauni, Nepali *muso* 'mouse, rat'.

After a non-etymological nasal (or secondary nasalization) this sibilant is sonorized (*š > ž):[22]

41. Balti ***munzhur*** 'small mole-like mouse' (Sprigg 2002: 119).

PII initial *w > *b*. See **2, 9, 20.**

One probable instance of consonant dissimilation across a morphemic boundary has been attested. See **22.**

Consonant clusters

PII initial *dw > *b*:

42. Balti ***bar-ban*** 'window (in a wall), glass-pane window' (Sprigg 2002: 26), Purik ***barban*** 'window' (Zemp 2018: 945). Cf. Brokskat ***barban*** 'id.', Pashai *darī*, Gawar-Bati *derī*, Kalasha *durík*, Phalura *darúṛi*, Indus Kohistani *dʌrī*, Shina *darií*, Kashmiri *dār* 'window', OIA *dvār-* 'door, gate', Av. *duuar-* 'gate'.

PII *kš (> OIA *kṣ*, Proto-Iranian *xš*, Dardic **čh*) > *čh*. See **4.**

This process may have gone through the intermediate stage of *čh*, the latter having changed to palatal *čh* in the wake of affricate shift.

Like in most Dardic and New Indo-Aryan languages, in etymological groups of two voiceless stops the first component is lost (**kt*, **pt* > *t*).[23] See **3, 14.**

The etymological voiceless stops are sonorized after nasals (**nk* > *ng* > *ŋ*, **nt* > *nd*).[24] See **2, 38**.

Proto-Indo-Iranian clusters of the type "nasal + voiced stop" develop in different ways. Word-finally PII **ndh* > *n*:

43. Balti ***ban*** 'fence' (Sprigg 2002: 25) < PII **bandha-*, cf. OIA *bandha-* 'border, framework, damming'.

In the same position PII **ngh* has apparently lost its nasal component and was simplified to *g* with subsequent devoicing (**ngh* > *g* > *k, q)*:

44. Ladakhi ***tak-tak*** (Shamskat dialect), ***ṭak-ṭak***, (Leh dialect) 'taut, stretched tight, tight (*e.g.* of curtain, clothing, greenhouse plastic)', Purik ***taqtaq*** 'tight' (Norman 2010: 370), Balti

[21] Developed from PIE **s* by the RUKI-rule.

[22] In Tibetological transcription the voiced palatal sibilant is conveyed by *zh.*

[23] It is, however, unclear, whether or not the development of these clusters involved regressive assimilation and the formation of geminates, as was the case, e.g. in the history of Indo-Aryan.

[24] Cf. the aforementioned sonorization of voiceless consonants after a non-etymological nasal or secondary nasalization.

ṭak-ṭak ‘hard’ (Sprigg 2002: 184) < **tag* < PII **tangh-* (> Proto-Iranian **ϑanj-* ‘to pull, draw’(> Av. *ϑaŋj-* ‘id.’, Persian *sanjīdan* ‘to measure; reflect; compare, put in balance’, Ossetic *tɨnʒɨn* ‘to spread, stretch out; to crucify’), Kashmiri *ṭanz* ‘extreme and urgent desire’, Khowar *tonjeik* ‘to destroy, pull down’).

The Proto-Indo-Iranian cluster **rt* seems to have undergone two different developments. In one Balti example the first component of this cluster moved from word-medial to word-initial position:

45. Balti ***rkat*** ‘to cut down with a sword’ (Sprigg 2002: 140). Cf. PII **kart-* ‘to cut’ > OIA *kṛ̊ntati, kartati* ‘cuts’, Av. *kərəṇtaiti* ‘cuts’, *karəta-* ‘knife’, Hindi-Urdu *kattā* ‘curved knife’, *kattī* ‘sword, knife, dagger’, Tirahi *katāri* ‘knife’, Pashai *kāṭare* ‘spear’, Gawar-Bati *kaṭāro* ‘large knife’, Kalasha *katār*, Khowar *kuter* ‘knife, dagger’, Bashkarik *kāṭer* ‘knife’, Savi *kaṭārēi*, Phalura *kaṭōro* ‘dagger’, Shina *khăṭa̯rʊ* ‘knife’.

On the other hand, there is an apparent instance of the development **rt* > *r* (perhaps, through the retroflex *ṛ* as an intermediate stage): Ladakhi ***kuruk*** ‘donkey foal; foal’ (Norman 2010: 12), Balti ***bong-kúru*** ‘donkey colt’ (Sprigg 2002: 30) < PII **kurtaka-* (see **33**).

It should, however, be noted that, technically, PII **kurtaka-* cannot so far be proven to be the only possible prototype for the immediate Indo-Iranian source of the Ladakhi and Balti words. Since the reflex of PII **rn* in the language under study is not clear, the prototype **kurnaka-*, similar to that which is reconstructed for the Iranian forms, cannot be excluded.

The PII cluster **rth* seems to be reflected as the retroflex or dental unaspirated stop (**rth* > Balti, Purik *ṭ*, Ladakhi *t*):

46. Balti, Purik ***gaṭ*** ‘knot, joint of body’ (Sprigg 2002: 58; Zemp 2018: 64), Ladakhi (Leh dialect) ***changgat*** ‘knee-joint’ (Norman 2010: 313), ***gat*** ‘obstacle’ (Norman 2010: 241) < **garthi-*, cf. Indus Kohistani *gāṛ*, Brokskat *gaṭhi* ‘knot’, Pashai *gaṭanā* ‘joint’, *gaṭh-* ‘to tie’, Hindi, Bengali, Marathi, Gujarati *gāṭh*, Punjabi, Lahnda *gaṇḍh*, Sindhi *gʼaṇḍhi* ‘knot’, Pashto *γaṛəl* ‘to twist, spin, plait’ (< **garϑ-*), Burushaski (loanword) *gaṭ* ‘*Knoten* (*auch im Stengel von Pflanzen*), *Knöchel* (*des Fingers*)’ (Berger 1998: 150). The reconstructed prototype **garthi-* is, no doubt, a derivative of PII **grath-*/**granth-* ‘to tie, bind’ > OIA *granthi-* ‘knot’, *granthayati* ‘ties’, Middle Persian *grih*, Persian *girih*, Khotanese *grratha,* Sogdian *γrʾnš*, Ishkashimi *γureẍ*, Ossetic *ælxync* ‘knot’.

PII **sć* (< PIE **skʼ*) > *ch*. See **17.**

This change has probably passed through the intermediate stage of **čh*.

PII **sw* > *sp* (see **30**).

This phonetic change is regular in a number of Dardic languages (e.g., Tirahi, Kalasha, Khowar). In Shina, however, it coexists with the more widespread change **sw* > *s*: *sa* ‘sister’ < **swasar-*, *sãc̣u* ‘dream’ < **swāpra-* (Turner 1966: 805), *so-* ‘to sleep’ < **swapa-*. It seems likely that words with *sp* < **sw* were borrowed from some other, perhaps now extinct, Dardic lect.

PII initial **śr* > *ṣ, š*. See **18.**

The cerebral and palatal reflexes are found in the Shamskat and Leh dialects of Ladakhi respectively. Since no examples of secondary cerebralization of the palatal sibilant were noted in Shamskat, it is reasonable to assume that the retroflex *ṣ* (< PII **śr*) of the Indo-Iranian donor language was preserved intact in this dialect and palatalized to *š* (*sh*) in the Leh variety due to the influence of the following front vowel.

PII **št* > *ṣṭ* > *lṭ, lt*:

Balti, Purik, Ladakhi ***mulṭuk***, ***multuk*** ‘fist’ (Norman 2010: 705; Sprigg 2002: 118) < **mulṭak* < **muṣṭaka-*. See **7.**

The change of a sibilant to *l* before a voiceless retroflex stop is apparently recent. As was shown in Kogan 2019, this process has even affected English loanwords. The presence of the retroflex *ṭ* in the above Tibetan words clearly indicates that the Proto-Indo-Iranian cluster **št*

has been cerebralized in the donor language, i.e. the development *št* > *ṣṭ* has taken place. The existence of the variant with dental (*multuk*) may be attributed to the fact that the sequence *lṭ* is rather rare in Northwestern Tibetan dialects.

The OIA cluster *ṣṭh*, unlike its counterpart with unaspirated stop (*ṣṭ*), seems to correspond to a single retroflex consonant in the language under study. This consonant is reflected as *ḍ* in Balti and *ṭ* in Ladakhi:

47. Balti ***kaḍik*** 'small branches' (Sprigg 2002: 82), Ladakhi ***kaṭik*** 'branches and leaves of trees as fodder for animals in the spring' (Norman 2010: 4). Cf. OIA *kāṣṭha-* 'piece of wood', *kāṣṭhikā-* 'small piece of wood', Lahnda, Punjabi, Hindi-Urdu, Kumauni, Nepali, Gujarati *kāṭh*, Sindhi *kāṭhī* 'wood', Kati *kåṭ* 'branch', Kalasha *kaṭ* 'board' (Trail 1999), Shina *kāṭ* 'wood', Kashmiri *kāṭh* 'wood', *kāṭh* 'small stick'.

The Ladakhi form looks more archaic than the Balti one. In the latter the intervocalic sonorization of an earlier voiceless retroflex stop must have occurred. The origin of the lexeme under discussion is still somewhat unclear. The most plausible and reliable etymology seems to be the one advanced by Otakar Klima and supported by Thomas Burrow (Klima 1970; Burrow 1975). According to these scholars, the Indo-European protoform of OIA *kāṣṭha-* should be reconstructed as **kolstho-* and construed as a derivative of PIE **kelə-* 'to hew' with an *s*-extension.[25] The element **-tho-* in this protoform is most probably etymologically identical to OIA *-thă-*, Av. *-ϑă-* < PII **-thă-* (cf. OIA *artha-* 'object, aim', Av. *arəϑa-* 'object, matter' < *ar-* 'to move, reach'; OIA *gāthă-*, Av. *gāϑā-* 'singing, chant' < *gā-* 'to sing'). For more details on this suffix see Wackernagel, Debrunner 1954: 717–722.

Based on the last two groups of examples, the following development of cluster-initial sibilants may be hypothesized: a sibilant disappears before a historical (Indo-Iranian) voiceless aspirate, and is retained, sometimes with subsequent changes, before an unaspirated voiceless stop. The same phonological development is characteristic of many Dardic languages and may be tentatively reconstructed for Proto-Dardic.

PIE **tk'* (> OIA *kṣ*, Proto-Iranian **š*) > *ch*:[26]

48. Balti ***tshon*** 'injury' (Sprigg 2002: 171), ***tshak*** 'rheumatism' (Sprigg 2002: 169), Purik ***tshaq*** 'pain with difficulty of breathing', ***tshak yong*** 'to ache', Ladakhi ***tshak*** 'sprain, pulled muscle, sudden cramp, sudden sharp pain' (Norman 2010: 759–760). Cf. OIA *kṣaṇoti* 'injures, hurts', *kṣata-* 'wounded', *kṣataka-* 'wound', *kṣaṇana-*, *kṣati-* 'injury, damage', Pali *khaṇati* 'destroys', Khowar *čay* 'illness' (= OIA *kṣati-*?), Khotanese *vaṣanaurau* 'destructive' (Bailey 1979: 379) < **vi-šana-bara-*, Manichaean Sogdian *p'šyyn* 3 Sg 'to trap' < **apa-* or **upa-šan-*, Christian Sogdian *ptšng*, Buddhist Sogdian *ptš'nkh* 'cross, torture-instrument' < **pati-šana-ka-* (Gershevitch 1954: 25, 96), Greek κτείνω 'I kill' < PIE **tk'en-*(LIV: 645).

The dental affricate *ch* must have regularly evolved from the earlier palatal *čh*. The latter phoneme is the likely Proto-Dardic reflex of PIE **tk'*. As was shown in Kogan 2019, this historical-phonological feature of the aforesaid Indo-Iranian loanwords is a strong argument in favor of their Dardic origin.

In addition to the historical-phonological innovations discussed above, one interesting archaism seems to be noteworthy. An Indo-Iranian loanword in Balti shows the retention of the initial stop in the Proto-Indo-Iranian cluster **gjh* (> Indo-Aryan *jh*, Proto-Iranian **gž*): ***gzar*** 'to flow' < PII **gjhar-* (see **6**).[27]

[25] In *Lexikon der indogermanischen Verben* this root is reconstructed as **kelh₂-* '*schlagen*' (LIV: 350). For PIE **kolstho-* Klima adduces probable reflexes in Slavic.

[26] In Indo-Iranian languages this Proto-Indo-European cluster has reflected in the same way as PIE **k's*. This suggests that the two groups may have merged in Proto-Indo-Iranian.

[27] On the Iranian reflexes of PII **gjhar-* see also (Cheung 2007; Rastorgueva, Edelman 2007).

Preliminary conclusions

The above analysis has revealed a number of facts that are directly relevant to the questions formulated at the beginning of the present work. First of all, four apparent cases of non-uniform phonological development in the same position have been detected. These are:

1) PII medial **a* or **ā* > *aa* in Balti ***baan*** 'man or men who sing religious songs and foretell the future' but > *a* elsewhere;
2) PII final **a* preserved in Balti ***basanda*** 'dandelion' but dropped elsewhere;
3) the initial *r* of PII **-rt-* has moved to the word-initial position in Balti ***rkat*** 'to cut down with a sword', whereas in Ladakhi ***kuruk*** 'donkey foal; foal', Balti ***bong-kúru*** 'donkey colt' the same intervocalic cluster evolves into *r* (perhaps, through retroflex *ṛ*);
4) PII intervocalic **ś* is probably preserved in Balti ***lashi*** 'a resinous wood used as a candle or torch because it burns slowly' but dropped in Ladakhi ***po-ze*** 'ram, full-grown male sheep'.

It is easy to see that all the above cases involve words peculiar to Balti. As shown in Kogan 2019, the majority of Indo-Iranian and Burushaski loanwords in Northwestern Tibetan fall into two groups: those characteristic of all the dialects, and those found only in Muslim varieties, i.e. in Balti and/or Purik. All the attested instances of seemingly unmotivated irregularities in historical phonology belong to the latter category. This fact seems to be explicable, if we bear in mind that the route of Tibetan migration to the present-day Muslim regions must have inevitably passed through Ladakh. The Indo-Iranian speakers of this or adjoining areas must have been the first people whom the Tibetans had encountered in the course of their northwestward movement. It was the contact with these people that had most probably resulted in the acquisition of loan vocabulary, now shared by Ladakhi, Purik and Balti. Later on, a certain part of Tibetan speakers migrated from Ladakh further northwest to Baltistan, where a number of lexical items from local or neighboring Indo-Iranian dialects could have been borrowed. In other words, Muslim dialects seem to possess at least two chronologically different Indo-Iranian lexical strata. A remarkable fact is that the earliest stratum, common with Ladakhi, does not show any historical-phonological irregularities. It means that there is so far no indication that loanwords belonging to this stratum have been borrowed from more than one source. This is why we consider it reasonable to accept the single-source hypothesis as a working one.

Although the lexical data analyzed above are rather scanty, they nevertheless do give us some hints as to the genetic position of the donor language. In my previous paper it was noted that no unquestionable examples of typical Indo-Aryan phonological developments had been so far attested in the material (Kogan 2019). Now it can be added that certain historical-phonological features, discovered during our study, make it virtually impossible to classify the source-language of at least some of the Indo-Iranian loans as Indo-Aryan. Here undoubtedly belong the deaspiration of PII **jh* (< PIE **gh* and **gʷh* in the palatalizing position) with the subsequent change *j* > *z*,[28] and the distinction, at least in word-initial position, between the reflexes of PII **kš* and PIE **tk'*.[29] On the other hand, the retention of the prevocalic (word-initial) and intervocalic PII **s* suggests that the language in question was not Iranian, since, as is widely known, in Iranian the lenition **s* > *h* occurred in both positions. The Nuristani origin of the above loanwords is hardly probable either, since no instance of a regular Nuristani development of the PIE palatal **k'* into the dental affricate *c* has been noted. Instead, as we have shown, the change **k'* > *š* has taken place. Apart from this, in Nuristani etymological voiceless

[28] See the note on Ladakhi *zanggi* 'a species of tiny flying biting insect'.

[29] As noted above, in Indo-Iranian this cluster seems to have merged with PIE **k's*. Its Proto-Indo-Iranian reflex may perhaps be reconstructed as **ćs*.

aspirates have been deaspirated with the result that the opposition in aspiration ceased to exist for consonants, whereas our data contain quite a number of words with voiceless aspirated stops and affricates.

Unlike Indo-Aryan, Iranian or Nuristani, certain apparently Dardic historical-phonological isoglosses have been detected in our material. They are as follows:

1) the deaspiration of Proto-Indo-Iranian voiced aspirates, including the aspirated affricate **jh*;
2) loss of historical sibilants in Indo-Iranian clusters of the type "sibilant + voiceless aspirate";[30]
3) PIE **tk'* > **čh* with subsequent dentalization to *ch*.

Two of the three above isoglosses are found outside Dardic (the isogloss (1) in Iranian and Nuristani, and the isogloss (2) in Middle and New Indo-Aryan) but the bundle as a whole seems to be unique to the Dardic group. The isogloss (3) indicates that the donor language, being Dardic, could not, however, have belonged to the East Dardic subgroup, where word-initially the Proto-Dardic **čh* was cerebralized into **ċ̣h* and thus merged with the reflex of PII **kš* (Kogan 2016; 2019).

Importantly, none of the three above-mentioned sound changes is found in lexical items peculiar to Muslim dialects alone. It means that these historical-phonological features most probably characterize the earliest stratum of Indo-Iranian loanwords, common to all the Northwestern Tibetan varieties. This fact substantially contributes to clarifying the nature of the Tibetan-Indo-Iranian language contact. As already noted, there are reasons to believe that the earliest ("common Northwest Tibetan") loanwords have been borrowed from a single source. Since this source-language could have been neither East Dardic nor Indo-Aryan or Iranian, and it is to these groups that all the known Indo-Iranian neighbors of Ladakhi, Balti and Purik belong or belonged, the substratum influence in our case seems to be much more likely than the adstratum one.

The tentative picture of linguistic and ethnic contact obtained from the present research may thus be represented in the following way. As a result of the northwestward expansion of the Tibetans in the early Middle Ages, a number of Tibetan dialects fell under the influence of a certain Indo-Iranian substratum lect. Geographically, the zone of this substrate contact was most probably located in present day Ladakh. The substrate language must have belonged to the Dardic group but, no doubt, was not an early form of Shina, Brokskat or Kashmiri, nowadays spoken in zones adjacent to Ladakhi and Balti. After a certain time period, during which considerable loan vocabulary was acquired, a new migration to the northwest took place. Its consequence was the spread of the Tibetan language in what is now Baltistan and Kargil. The majority of the local pre-Tibetan population seems to have spoken a certain variety of Burushaski (Kogan 2019)[31] but the presence of some Indo-Iranian forms of speech is also probable. The latter lects must have influenced the early form of Balti, as is clear from the existence of several Indo-Iranian loanwords specific to the Balti dialects. The exact source of such loanwords, as well as the type of contact that has resulted in their adoption, remain unknown because of the extreme scarcity of the material. There is, however, a possibility that future field linguistic research will yield sufficient new lexical data, instrumental in the clarification of these both issues.

[30] See the note on the intervocalic retroflex in Balti *kaḍik* 'small branches' and Ladakhi *kaṭik* 'branches and leaves of trees as fodder for animals in the spring'.

[31] Burushaski seems to have exerted some influence on the Indo-Iranian language of pre-Tibetan Ladakh. This is indicated by the existence of Burushaski loanwords common for Ladakhi, Balti and Purik. It is remarkable that some of these loanwords are attested with the Indo-Iranian formant *-k* (see **34**).

Abbreviations for language names

Av. — Avestan; OIA — Old Indo-Aryan; PIE — Proto-Indo-European; PII — Proto-Indo-Iranian

References

Bailey, T. Grahame. 1924. *Grammar of the Shina Language ("Royal Asiatic Society, Prize Publication Fund", vol. VIII).* London: Royal Asiatic Society.

Bailey, Harold W. 1979. *Dictionary of Khotan Saka.* Cambridge University Press.

Berger, Hermann. 1998. *Die Burushaski-Sprache von Hunza und Nager. Teil III. Wörterbuch Burushaski-Deutsch, Deutsch-Burushaski. Unter Mitarbeit von Nasiruddin Hunzai.* Wiesbaden: Harrassowitz Verlag.

Buddruss, Georg. 1960. *Die Sprache von Wotapur und Katarqala.* Bonn: Selbstverlag des orientalischen Seminars der Universität Bonn.

Burrow, Thomas. 1975. A new look at Brugmann's law. *Bulletin of the School of Oriental and African Studies,* Vol. 38, No. 1: 55–80.

Cheung, Johnny. 2007. *Etymological Dictionary of the Iranian Verb. Leiden Indo-European etymological dictionary series.* Leiden / Boston: Brill.

Francke, August H. 1907. *A history of Western Tibet: one of the unknown empires.* London: S. W. Partridge & Co.

Gamkrelidze, Tamaz V., Vycheslav V. Ivanov. 1984. *Indoevropjskij jazyk i indoevropejcy: Rekonstrukcija i istoriko-tipologieskij analiz prajazyka i protokultury.* Tbilisi: Izdatel'stvo tbilisskogo universiteta.

Gershevitch, Ilya. 1954. *A Grammar of Manichean Sogdian. (Publications of the philological society).* Oxford: Basil Blackwell.

Grierson, George A. 1906. *The Piśāca languages of North-Western India.* London: Royal Asiatic Society.

Grierson, George A. 1915–32. *A Dictionary of the Kashmiri Language. Vol. I–IV ("Bibliotheca Indica". Work No. 229).* Calcutta: Royal Asiatic Society of Bengal.

Grierson, George A. 1919. *Linguistic Survey of India. Vol. VIII, pt. 2. Specimens of the Dardic or Piśāca Languages (including Kāshmīrī).* Calcutta: Superintendent Goverment Printing, India.

Jettmar, Karl. 1975. *Die Religionen des Hindukusch (= Die Religionen der Menschheit. Bd. 4, 1). Mit Beiträgen von Schuyler Jones und Max Klimburg.* Stuttgart: Kohlhammer Verlag.

Klima, Otakar. 1970. Das altind. kāṣṭhám. *Archiv orientální* 38: 26–27.

Kogan, Anton I. 2015. Nekotoije voprosy geneticheskoj klassifikatsii dardskix jazykov po dannym istorichekoj fonetiki. *Journal of Language Relationship* 13/1: 1–21.

Kogan, Anton I. 2016. *Problemy sravnitel'no-istoricheskogo izuchenija jazyka kashmiri.* Moskva: "Fond razvitija fundamental'nyx lingvisticheskix issledovanij".

Kogan, Anton I. 2019. On possible Dardic and Burushaski influence on some Northwestern Tibetan dialects. *Journal of Language Relationship* 17/4: 263–284.

LIV = Kümmel, Martin and Helmut Rix, eds. 2001. *Lexikon der indogermanischen Verben: Die Wurzeln und ihre Primärstammbildungen. 2nd ed.* Wiesbaden: Reichert.

Liljegren, Henrik, Naseem Haider. 2011. *Palula vocabulary.* Islamabad: Forum for Language Initiatives.

Livshits, Vladimir A. 2003. Drevneje nazvanije Syrdar'ji. *Vestnik drevnej istorii* 1(244): 3–10.

Lorimer, David L. R. 1939. *The Burushaski language. Vol. III. Vocabularies and index.* Oslo: H. Aschehoug & Co. (W. Nygaard).

Mayrhofer, Manfred. 1996. *Etymologisches Wörterbuch des Altindoarischen. II Band.* Heidelberg: Universitätsverlag C. Winter.

Morgenstierne, Georg. 1940. Notes on Bashkarik. *Acta Orientalia* XVIII(3/4): 206–257.

Morgenstierne, Georg. 1941. *Notes on Phalura, an Unknown Dardic Language of Chitral.* Oslo: J. Dybwad.

Morgenstierne, Georg. 1973. Sanskritic Words in Khowar. In: Georg Morgenstierne. *Irano-Dardica:* 256–272. Wiesbaden: Dr. Ludwig Reichert Verlag.

Norman, Rebecca. 2010. *A Dictionary of the Language Spoken by Ladakhis.* Ms. (unpublished draft).

Pokorny, Julius. 1959. Indogermanisches etymologisches Wörterbuch. Bern / München: Francke Verlag.

Rastorgueva, Vera S., Dzhoy I. Edelman. 2007. *Etimologicheskij slovar' iranskix jazykov. T. 3 (f—h).* Moskva: «Vostochnaja literatura» RAN.

Sprigg, Richard K. 2002. *Balti-English, English-Balti dictionary.* New York / London: Routledge.

Trail, Ronald L., Gregory R. Cooper. 1999. *Kalasha Dictionary with English and Urdu compiled by Ronald L. Trail, Gregory R. Cooper.* Islamabad: National Institute of Pakistan Studies, Quaidi-Azam University, Summer Institute of Linguistics.

Turner, Ralph L. 1966. *A Comparative Dictionary of the Indo-Aryan Languages.* Oxford University Press.

Wackernagel, Jacob. 1954. *Altindische Grammatik. Band II, 2. Die Nominalsuffixe von Albert Debrunner.* Göttingen: Vandenhoeck & Ruprecht.

Zemp, Marius. 2018. *A grammar of Purik Tibetan.* Leiden / Boston: Brill.

Zoller, Claus P. 2005. *A Grammar and Dictionary of Indus Kohistani. Volume 1: Dictionary.* Berlin / New York: Mouton de Gruyter.

А. И. Коган. Об историко-фонетических особенностях индоиранских заимствований в северо-западных тибетских диалектах

Как показали недавние исследования, в лексике ряда северо-западных тибетских диалектов имеется значительное количество индоиранских заимствований. Остается, однако, неясным, были ли эти заимствования усвоены из одного или нескольких источников и является ли их наличие следствием воздействия субстрата или адстрата. Не установлено и точное генетическое положение языка-донора внутри индоиранской языковой общности. Изучение данного круга вопросов, несомненно, должно опираться на факты исторической фонетики. В статье представлена попытка выявить наиболее яркие историко-фонетические особенности индоиранских элементов в северо-западных диалектах тибетского языка. Кроме того, делается ряд предварительных выводов, касающихся лингвистической географии исследуемого региона в дотибетскую эпоху, а также направления и относительной хронологии миграций тибетцев.

Ключевые слова: историческая фонетика; языковые контакты; языковой субстрат; индоиранские языки; дардские языки; тибетские диалекты; язык ладакхи; язык балти; язык пурик.

Andrey Shatskov
Institute for Linguistic Studies, St. Petersburg; andrey.shatskov@gmail.com

Hittite causative markers in a diachronic Anatolian perspective[1]

Hittite is a heavily transivitizing language, and there are several morphological markers of causativisation in Hittite. Two of them, namely suffixes *-nu-* and *-ahh-*, were productive in the history of Hittite. Other markers are either no longer productive or primarily not causative. In Old Hittite *-nu-* and *-ahh-* still had some kind of complementary distribution, although there already was some overlapping, but in Middle and especially New Hittite these markers became nearly interchangeable. The coexistence of *-nu-* and *-ahh-* derivatives from the same bases can be attributed, at least partially, to an imperfect knowledge of Hittite by Luwian scribes.

Keywords: Hittite language; causative markers; verbal morphology; iterative forms; intensive forms; Luwian language.

1. Introduction. Causative markers in Hittite

Hittite is a heavily transitivizing language (see a detailed discussion in Luraghi 2012), which features a large number of causative verbs. Nevertheless, the number of causative markers found in Hittite, both productive and fossilized, is surprising. In total, five markers, namely the suffixes *-nu-*, *-ahh-* and *-e-*[2] as well as reduplication and the nasal infix, can have causative semantics. These causative markers differ in frequency and diachronic distribution, but all of them were inherited from Proto-Indo-European and have counterparts in other Indo-European languages. The two most common causative markers in Hittite are the suffixes *-nu-* and *-ahh-*, which could be added to verbs as well as adjectives and nouns.[3] Overall, there are about 130 *nu*-verbs and 75 *ahh*-verbs; this makes them two of the most prominent verbal classes in Hittite. In de-adjectival derivation, these suffixes were often used next to fientives in *-ē-* and its enlarged variant *-ess-* (see Oettinger 1979: 238ff.), and could be added virtually to any descriptive adjective in Hittite. The other three causative markers are attested only in a few verbs each and are either no longer productive in Hittite (the suffix *-e-* and the infix) or primarily not causative (reduplication). In the following section I shall first discuss each of the causative markers, including their origin, and then address their distribution.

2. Causative suffixes in Hittite

2.1. The suffix *-nu-*

The suffix *-nu-*, the most frequent causative marker, can be found added to both verbs and adjectives already in the earliest texts. There are two denominal *nu*-verbs as well: *ēsharnu-*

[1] This research has been supported by the Russian foundation for basic research (project № 19-012-00581 "Valency alternations in Hittite, Old Armenian and Ossetic in diachronic and comparative perspective"). I am thankful to D. Sasseville, who made his dissertation available to me ahead of publication.

[2] This suffix is found in *lukke-* 'to kindle, set on fire' and *wasse-* 'to dress (someone)', see Eichner 1969: 31f. and Watkins 1973: 68f.

[3] The suffix *-ahh-* could also be added to the numerals, see Kronasser 1966: 430.

'to make bloody, red', derived from *ēshar* 'blood',[4] and KAL-*tarnu-* (**hatugatarnu-*) 'to make formidable', derived from *hatugātar* 'terribleness, terror'. This suffix is well attested in many other Indo-European languages and was clearly present already in PIE (cf. LIV: 17f.). Yet there is only one Hittite *nu*-verb that has lexical cognates in non-Anatolian IE languages: Hitt. *arnu-* 'to make go, stir, remove, deliver' vs. Gr. ὄρνυμι 'to stir up, make to arise, incite', Skt. *ṛṇóti* 'to move, put in motion'. Another *nu*-verb that was allegedly inherited from PIE is *tepnu-* 'to diminish'. It has been compared to Skt. *dabhnóti* 'deceives' (see e.g. LIV: 132), while Koch (1980: 235f.) suggested that the suffix *-nu-*, seen in *tepnu-* 'to diminish' and Skt. *dabhnóti* 'to damage', is in fact a reanalysis of the infix **-n-* inserted into the adjectival stem **dhebhu-* 'little'. This is, however, hardly possible, since *tepnu-* is mostly found in New Hittite texts as well as a few Middle Hittite texts, see Hoffner 1977: 152ff. Given that the suffix *-nu-* was extremely productive and that the meaning of Hitt. *tepnu-* 'to diminish' is based on *tēpu-* 'small' and is quite different from that of the Sanskrit verb ('to deceive, heart'), *tepnu-* is likely to represent a Hittite de-adjectival formation, which could also explain the full grade of the root, see Shatskov 2017: 197f. Besides, there seems to be no special connection between *u*-stem adjectives and verbal *nu*-stems in other IE languages. For instance, in Sanskrit *u*-adjectives often occur beside causatives in *-áya-* or infixed stems, e.g. Skt. *svādú-* 'sweet' and Skt. *svadáyati* 'to season sweeten' or Skt. *pr̥thú-* 'broad wide' and Skt. *pratháyati* 'to make spread out', cf. further Rau 2009: 170ff. and 183f. In late PIE the suffix **-neu/nu-* and the infix **-ne/n-* were clearly different morphemes, even though they may ultimately be cognate.[5]

The suffix *-nu-* is also attested in Luwian as well as in Lycian and Lydian, see Sasseville 2018: 91ff. However, despite the fact that this stem was common both in Hittite and Luwian, the cognates are few: CLuw. *asharnu-* 'to make red, bloody' vs. Hitt. *ēsharnu-* 'id.', CLuw. and HLuw. *huinu-* 'to make run' vs. Hitt. *huinu-* 'id.', CLuw. and HLuw. *marnuwa-* 'to make disappear' vs. Hitt. *mernu-* 'id.', perhaps also HLuw. *asunu-* 'to make good?' vs. Hitt. *as*(*sa*)*nu-* 'to take care of'. HLuw. *zahhanuwa-* 'to attack' seems to have a parallel in Hitt. *zahhiyanu-* 'to attack', but the latter is attested only once and apparently based on Hitt. *zahhiya-* 'to attack', so it must be an inner-Hittite formation, if this form is real at all.[6] The low number of correspondences indicates that most Hittite and Luwian *nu*-verbs were formed after the breakup of Proto-Anatolian.

2.2. The suffix *-ahh-*

In Old Hittite *-ahh-* was used to derive causatives (factitives) from adjectives and nouns, but later this suffix started to be added to verbs as well, cf. e.g. *tarup*(*iy*)*ahh-* from *tarupp-* 'to unite, collect' or *kartimmiyahh-* 'to make angry' from *kartimmiye/a-* 'to be angry', see Shatskov 2017: 226f.

Hittite causatives in *-ahh-* have cognates in Luwian, Lycian and Lydian stems in *-a-* (Melchert 1997: 332, Sasseville 2018: 15ff.). This class is well attested in both Luwian and Lycian, but correspondences with Hittite are few, the only secure example being Hitt. *marsahh-* 'to desecrate' (next to the better attested *marsanu-*) and Luw. *marsa-* 'to commit a treachery (?)', Lyc. *mrssxa-* 'to desecrate' (Sasseville 2018: 25f., 41). Another possible example is Hitt. *dannattahh-* 'to devastate' and Luw. *tannatta-* 'to devastate', though the Hittite adjective *dannatta-*

[4] Since *ēsharnu-* is cognate to Luw. *asharnu-* 'to make bloody', denominal *nu*-verbs must be of Proto-Anatolian age.

[5] For various theories concerning the origin of the infix in PIE and its relation to the suffix **-neu-* cf. e.g. Milizia 2004 and Steer 2013-14.

[6] The verbs *zahhiya-* and *zahhiyanu-* are used in the same text and in the same context, cf. KBo 3.4 II 25 *n=an* ^{D}UTU^{ŠI} *zahhiyanun* and II 60 *n=an* ^{D}UTU^{ŠI} MÈ-*yanunun* 'and I, My Majesty, attacked him', and the extra *-nu-* sign in MÈ-*ya-nu-nu-un* may well be a scribal error.

may be a Luwian borrowing, and in that case its derivative *dannatahh-* is a late and independent formation (Sasseville 2018: 35).

The suffix *-ahh-* has cognates in other IE languages (cf. e.g. Latin *novāre* 'to make new, renew', Gk. νεάω 'to plough up a fallow land' and Hitt. *newahh-* 'to renew') and was evidently employed in formation of deadjectival verbs, cf. e.g. Fortson 2010: 99f. Sasseville (2014–15, 2018: 56ff.) argued that causative verbs in **-eh₂-* are derived from nouns in **-eh₂-* via conversion, so the Hittite verbal suffix *-ahh-* results from reanalysis of the nominal stem in **-eh₂-* (Sasseville 2018: 60f.). However, in that case the predominantly de-adjectival derivation of such stems in Hittite and Latin[7] must be an independent innovation, whereas the denominal derivation was preserved as its basic function only in Lycian. If the verbal suffix is indeed cognate with nominal stems in **h₂*, one could perhaps assume that the reanalysis of the nominal stem as the verbal factitive stem occurred already in PIE.

2.3 The suffix *-e-*

A common PIE causative marker was the suffix **-éye/o-*, accompanied by the vocalism **o* in the root (e.g. Fortson 2010: 99, LIV: 22f.).[8] The Hittite reflex of this suffix, *-e-*, is preserved only in two verbs: *wasse-* 'to put on' next to *wass-* 'to be dressed' and *lukke-* 'to set fire to' next to *lukk-* 'to become light'. The causative stem in *-e-* clearly was no longer morphologically transparent, and eventually it was replaced with other stems: cf. *luk-* in *lukkun* (KBo 12.38 III 9, NH) or *wassiya-* in *wassiyanzi* (e.g. KUB 9.31 II 11, MH/NS). PIE causative **-éye/o-* stems are likely to be continued by several Hittite *hi*-conjugation verbs including *lak-* to 'fell' and *kank-* 'to hang'; in the prehistory of Hittite these formations presumably lost the suffix **-éye/o-* and shifted to the *hi*-conjugation, see Schulze-Thulin 2001 and Kloekhorst 2008: 437f, 514f. Kloekhorst (2008: 532, 1006f.) also argues that *wasse-* and *lukke-* have the suffix **-ye/o-* rather than causative-iterative **-éye/o-*. Note, however, that both *lukke-* and *wasse-* have **-éye/o-*counterparts in other Indo-European languages: Skt. *rocáyati*, Latin *lucēo* 'to cause to shine', Skt. *vāsáyati*, Goth. *wasjan* 'to dress' (LIV: 419, 692). Besides, Kloekhorst's examples of *mi*-conjugation **-ye/o-* stems next to unextended middle stems (*huett-*$^{\text{tta(ri)}}$ and *huttiye-*$^{\text{zi}}$ 'to pull', *hatt-*$^{\text{a(ri)}}$ and *hazzie-*$^{\text{zi}}$ 'to pierce') show no semantic distinction between middle and *ye/o*-stems, whereas *lukke-* and *wasse-* clearly differ from the respective middle stems and do have a causative meaning.

The reflexes of PIE causatives-iteratives in **-éye/o-* are better preserved in other Anatolian languages, see Sasseville 2018: 184ff., 208ff., 211f., though there are no exact lexical matches with Hittite.

2.4 The infix

Infixed verbs are well attested in Hittite, although they are not very numerous, cf. *hark-* 'to perish' vs. *harnink-* 'to destroy' or *istark-* 'to ail, become ill' vs. *istarnink-* 'to make ill'. Two other Hittite verbs with the infix *-ni(n)-* do not have the infixless counterparts preserved: *sarnink-* 'to compensate' (cf. Lat. *sarciō* 'to repair')[9] and *ninink-* 'to set in motion, disturb' (cf. Lith. *ap-ninkù, -nìkti* 'to assault, beset', *į-nìkti* 'to get down to, engage, attack', *ap-nìkti* 'to energetically get down to; beset', Gr. νεῖκος 'quarrel'). Finally, *hunink-* 'to scar, crack'[10] is not causative as it is

[7] On this type in Latin, cf. e.g. Leumann 1977: 546.

[8] Stems in **-éye/o-* may also have the iterative-intensive value. On the polysemy of this suffix see 3.1 below.

[9] Kloekhorst (2008: 734f.) argues that the underlying stem *sark-* is attested in *sarkiske-* which he translates as 'to be good, eminent' rather than 'to ascend', as e.g. in CHD Š: 267.

[10] On the meaning of *hunink-*, see Shatskov 2017: 27ff.

based on *huek-* 'to slaughter'.[11] As argued in Shatskov 2017: 62, the infix *-nin-* was a productive way of deriving causatives in proto-Hittite; however, in late Hittite the infixed stem was apparently no longer transparent and could be replaced with a *-nu*-stem, as in the case of *harninink-* 'to destroy', which was substituted in some texts with *harganu-* 'id.' (Ünal 1984: 77f.).

While there are no infixed verbs in Luwian, the infix is preserved in both Palaic and Luwian as a part of the suffix *-īna-* (Sasseville 2018: 455f).

2.5 Reduplication

A secure example of the Hittite causative reduplicated formation is *asās-/ases-* 'to install, settle' derived from *ēs-* 'to sit'. Two more likely examples are *titta-* 'to install'[12] and *lilakk-* 'to fell, cut down'. Dempsey (2015: 99) considers *lilakk-* to have imperfective semantics since the verbs that stand next to *lilakk-* in the only preserved context have the imperfective suffixes *-ske/a-* and *-anna/i-*.[13] However, the active forms of the parent verb *lak-* are not attested with the meaning 'to fell' and mean rather 'to knock out (a tooth)' or 'to turn (ear or eyes)', see CHD L-N: 17f. In my opinion, *lilakk-* is a causative to middle intransitive forms of *lak-* that mean 'to fall' and in that respect is similar to *laknu-* 'to fell' derived from the same verb or *karsanu-* 'to stop (tr.)', which must have been derived from the middle intransitive stem *kars-* 'to stop (intr.), withhold' rather than from its active counterpart that means 'to cut' (see Shatskov 2017: 148). Nevertheless, the majority of reduplicated verbs clearly have the iterative or intensive semantics, see Dempsey 2015: 331f.

The reduplicated verbs in Luwian likewise generally have iterative, durative or similar meanings (Dempsey 2015: 255). Reduplication is also well attested in other PIE branches, and one of its functions is causativization, see Kölligan 2004: 195ff., 223ff.

2.6 Summary

All in all, while all the Hittite causative markers are of PIE origin, there are only a few lexical matches between causative stems in Hittite and other Anatolian and, more broadly, Indo-European languages; besides there is also at least one mismatch: Hitt. *parkunu-* 'to cleanse' and Luw. *paparkuwa-* 'to cleanse' (Sasseville 2018: 27). Therefore, the majority of Hittite causatives must have been formed after the breakup of Proto-Anatolian.

3. Polysemy of the causative markers

The majority of Hittite causative markers show a causative-iterative polysemy.[14] Cross-linguistically, causative and iterative-intensive meanings are very often expressed by the same

[11] Therefore, the infix may have non-causative semantics, similarly to other historical causative markers in Hittite, cf. Section 3.1 below.

[12] If one follows the suggestion of Melchert 2018 and Dempsey 2015: 132ff., who set apart *titta/i-* 'to cause to stand, install' derived from *tiya-* 'to step, stand' and *titta/i-* 'to place' derived from *dai/tiya-* 'to put'.

[13] KUB 24.8 I 2-6: [*h*]*an*[*dand*]*us* LÚ^{MEŠ}*-us* [*k*]*uis* [(*sar*)*l*]*iskezzi hūwappas*[=*a*=*k*]*an* LÚ^{MEŠ}*-us* [(GIŠ-*ru m*)]*ān lilakki hūwapp*[*us*]=*a*=*kan* LÚ^{MEŠ}*-as* [(*tarn*)]*as*=*smas saksakilus walhannai* [*t*]=*us harnikzi* 'who vindicates just men, who fells evil men like trees, who strikes evil men on their skulls like s. and destroys them.' The forms *sarliskezzi* and *walhannai* are indeed imperfective, whereas *harnikzi* is not.

[14] The suffix *-e-* is causative in both of its preserved instances. Note, however, that in other Indo-European languages the reflexes of PIE **-éye/o-* also show the same polysemy (Kölligan 2004, 2007, Kulikov 2013).

morphemes, see e.g. Nedyalkov, Silnitsky 1973: 19f., Kittilä 2009 and Aikhenwald 2011. As I shall argue elsewhere (Shatskov forthc.), the Hittite morphemes *-nu-* and *-ahh-* show the same distribution of causative and iterative meanings as the reflexes of PIE suffix **-éye/o-* in Vedic and Latin. To be more precise, if derived from verbs of low semantic transitivity, *-nu-* and *-ahh-* have causative meanings, and if derived from verbs of high semantic transitivity, they have iterative or intensive meanings. The difference between *-nu-* and *-ahh-*, on the one hand, and reduplication, on the other hand, is that the former are mainly causative suffixes, whereas the latter is mostly iterative-intensive in Hittite.

4. Distribution of the causative markers

4.1. The suffix *-e-*, the reflex of the very common PIE causative marker **-éye/o-*, is sparsely attested in Hittite. The infix, clearly causative in Hittite and likely so in Proto-Indo-European, enjoyed limited productivity in the prehistory of Hittite but was no longer productive in the historical period (Shatskov 2017: 62f.). Reduplication is productive, but mostly used to forms imperfectives.

The only productive causative markers during the recorded period of Hittite were the suffixes *-ahh-* and *-nu-*. In the oldest texts there are causatives in *-nu-* derived from verbs and adjectives in *-u-* (*dassanu-* 'to make strong' from *dassu-* 'strong') and *-i-* (*parkunu-* 'to cleanse' from *parkui-* 'clean', *sallanu-* 'to bring up' from *salli-* 'big'), while causatives in *-ahh-* were derived from adjectives in *-ant-* (*dasuwahh-* 'to blind' from *dasuwant-* 'blind') and *-i-* (*suppiyahh-* 'to purify' from *suppi-* 'sacred, pure').[15] Thus, already in the oldest texts there was a certain overlap in the distribution of *-nu-* and *-ahh-*. In later periods *-ahh-* started to be added to adjectives in *-u-* as well as verbs, whereas *-nu-* started to be added to adjectives in *-a-* and *-ant-*.

4.2. There are several roots that have causatives in both *-nu-* and *-ahh-* with little, if any, difference in meaning. Note that in most cases one (or both) stems are attested only once or twice, often in damaged contexts. Cf. the following examples:

kardimiyanu- and *kardimiyahh-* 'to make angry'

mān=ma=šta ZI^{TUM} DINGIR$^{LI(IM}$*-ma ku*)*i*]*š* TUKU.TUKU-*ya**nu**zi* 'If [som]eone angers the soul of go[d]' KUB 13.4 I 34 (MH/NS)

*kardimiy**ahh**anzi=an=kan kuyēs* 'those who anger him' KUB 35.146 II 13 (MS)

dankuwanu- and *dankuwahh-* 'to make black'

[LUG]AL-*us=wa kuit ēsharwah*[*heskitta hahlahheskitta*] [*da*]*nkuwahheskitta* [*harganusk*]*i*[*tta nu ap*]*āt* EGIR-pa *ANA* KUR ^{LÚ}KÚ[R 'and what the king [has made red, green,] dark [and white], it back to the enemy land …' KBo 15.1 I 27–9 (NS).[16]

[(*nu=war=a*)*n*] *tankunu*[*sker kuē*(*s=war=an* SA$_5$-*nusker*)] *kuēs=wa*[*r=an* ZA.GÌN-*nusker*] *kunun=an* [(EGIR-*pa dankunuske*)*r*] 'they make it black. They then made black those they had made red and those they had made blue.' KUB 9.34 I 4–7 (NH)

taruppiyanu- and *tarupp*(*iy*)*ahh-* 'to gather, collect'

[15] As for Hittite adjectives in *-a-*, the causatives in **-eh$_2$-* (Hitt. *-ahh-*) could be derived from the predecessor of this type, the **o*-stem adjectives, already in PIE (see 2.2 above).

[16] Kümmel (1967: 113) translates this as "Womit der [Kön]ig blutrot [ge]mach[t, grün gemacht,] schwarz gemacht [(und) Weiss gemacht ist], [d]as [soll] züruck ins Feindland …".Yet the passive interpretation of the middle form [*da*]*nkuwahheskitta* is not obligatory, and *kuit,* coordinated with [*ap*]*āt* in the next clause, is likely to be a direct object here. Then [*da*]*nkuwahheskitta* is transitive despite its middle voice (unless it shall be interpreted as *dankuwahheskitt=a*, an active form with a connective particle *-a*/*-ya*).

LÚDAM.GÀR=*ma=za=kan kuin ēpta nu* KÙ.BABBAR *nawi daruppiyanuzi* '(but (as for) the merchant who took it for himself, he has not yet collected the silver' IBoT 2.129 I 21–22 (NH)

[(G)]U_4-*un* UDU-*an* LÚ.U_{19}[.L]U *taruppiyahhas* […EGIR-*p*]*a taruppiahhis* 'gathered cattle, sheep, people … gathered…' KUB 9.11 I 16–17 (OH/NS)

tepnu- and *tepawahh-* 'to diminish, demean'[17]

arahzenas=wa=mu=za KUR.KUR LÚKÚR kuiēs DUMU-*lan halzesser nu=wa=mu=za tepnusker* 'The surrounding enemy lands who called me 'a child', they humiliated me' KBo 3.4 I 23–24 (NH)

nu=ssi=za EGIR-*an ŪL memas n=an ANA PANI* KUR.KURMEŠ *tepaw*[*a*]*hta!* '(The crown prince held him by the hand), but he said "no" to him and demeaned him in the presence of the lands.' KUB 14.3 I 12–13 (NH)

wastanu- and *wastahh-* 'to sin'

GIM-*an=ma ANA* m*Arnuwand*[*a* NUMUN *ēsta*] *man wastanunun man Š*[*A* EN=*YA*] NUMUN *arha tarnahun* 'If there had been progeny for Arnuwanda, I would have sinned, I would have let away progeny of my brother (from kingship).' KBo 12.41+ II 10–12 (NH)

]*ki kuit wasta*[*hh*]*u*[*n*] 'what I sinned'[18] KUB 36.86 obv. 10 (NS)

Finally, next to *maninkuwahh-* 'to come near; to make short' and *maninkuwantahh-* 'to make short', there is *maninkuwanu-*, which is attested once in broken context (Bo 6238 7 (NH)), see Soysal, Yildiz Gülşen 2019: 4), so its meaning cannot be ascertained, but due to the productivity of the derivational model it must be something like 'to bring near' or 'to come near'.

There is no uniform chronological distribution of the suffixes in these pairs. Sometimes, the *ahh*-stem is the older and better attested one, as is the case of *maninkuwahh-*. By contrast, *tepnu-* is well attested since Middle Hittite (see Hoffner 1977: 152ff.), while *tepawahh-* is found only twice. Similarly, *wastahh-* is found twice in the same NS text (KUB 36.86), whereas *wastanu-* is attested five times, once in a NS copy of an Old Hittite text (KUB 13.3 I 6). The causatives *kartimiyanu-* and *kartimiyahh-* seem to be of the same age, both attested already in Middle Hittite, although *kartimiyanu-* is more frequent. Finally, *taruppiyanu-* and *tarupp*(*iy*)*ahh-* as well *dankuwanu-* and *dankuwahh-* are all attested one to three times in late texts.

4.3. The causative markers *-nu-* and *-ahh-* were so productive that they started to replace or reinforce causatives formed with other markers. During the reign of Hattusili III, near the end of the Hittite kingdom, an infixed causative *harnink-* started to be replaced with a parallel *nu*-formation *harganu-*, also based on *hark-* 'to perish'. The *nu*-verb *asesanu-* derived from a reduplicated causative stem *asās*/*ases-* 'to settle, to install' is attested since the times of Mursili II. There seems to be a certain difference in their semantics: *asesanu-* usually means 'to settle', while *asās*/*ases-* may also mean 'to set up' or 'to install'. The editors of HW² (A: 391f.) note that during the reign of Hattusili III the verb *asās*/*ases-* virtually ceased to be used with the meaning 'to settle', which was relegated to *asesanu-*. Similarly, *tittanu-* is derived from a reduplicated causative stem *titta*/*i-* 'to install' and is attested mostly in late Hittite texts, although it appears once in a New Hittite copy of an Old Hittite text (KUB 25.1 VI 20, cf. Dempsey 2015: 143ff.).

4.4. The addition of the causative marker to an already causative stem in case of *asesanu-* and *tittanu-* as well as the replacement of the infixed stem *harnink-* with a *nu*-stem *harganu-* must have been motivated by the desire to make these stems more recognizable as causatives. By

[17] On *tepawahh-* cf. Hoffner 2009: 302, 390, note 267.

[18] Note that *wastahh-* does not have to be transitive and might be restored as [*ŪL kuit*]*ki kuit wastahhun* 'since I have not sinned anyhow'. According to Tischler (HEG W-Z: 408f., 412), *wastanu-* is a deverbal formation whereas *wastahh-* is denominal, but in my view both verbs may be denominal as well as deverbal.

contrast, the parallel use of *nu-* and *ahh*-causatives derived from the same adjective or verb is more difficult to justify. It is clear that the causative formations with *-nu-* and *-ahh-* were very productive. Nevertheless, this does not explain why *tepawahh-* 'to demean' appears in the Tawagalawa letter (CTH 181), while *tepnu-* with the same meaning had been in use since Middle Hittite. In my opinion, the formation of new causatives instead of using the available ones could be due to imperfect language competence by the scribes. As argued by Yakubovich (2010: 307f.), by the end of the New Hittite period the native language of the majority of scribes was Luwian. If so, the scribes had to learn Hittite, and it is conceivable that they were well acquainted with the productive patterns of causative derivation but could be not aware of their distribution with regard to certain verbs.

5. Conclusions

Five markers in total, all inherited from Proto-Indo-European, could be used to form causatives in Hittite. Even though Hittite preserved the reflexes of the most frequent PIE causative markers (suffix *-e-* < PIE *-*éye*/*o-*, infix *-ni*(*n*)- < PIE *-*né*/*n-*), they were no longer productive. Instead, morphologically more transparent suffixes *-nu-* and *-ahh-* were employed in this role. The two suffixes under discussion owe their productivity partially due to the fact that they were deadjectival and could be added to virtually any descriptive adjective. Reduplication was mainly used to form imperfectives, so reduplicated causatives were rare. The coexistence of at least some parallel *-nu-* and *-ahh-* formations derived from the same root can be caused by the scribes' imperfect knowledge of Hittite.

References:

Aikhenvald, Alexandra. 2011. Causatives Which Do not Cause: Non-Valency-Increasing Effects of a Valency-Increasing Derivation. In: A. Aikhenvald, R. Dixon (eds.). *Language at Large: Essays on Syntax and Semantics*: 86–142. Leiden – Boston: Brill.

CHD = Güterbock, Hans Gustav, Harry Hoffner, Theo van den Hout (eds.). 1980–. *The* Hittite dictionary *of the Oriental Institute of the University of Chicago.* University of Chicago.

Dempsey, Timothy Richard. 2015. *Verbal Reduplication in Anatolian*. PhD thesis. University of California, Los Angeles.

Eichner, Heinrich. 1969. Hethitisch *u̯ešš-*/*u̯aššiu̯a-* "(Gewander) tragen; anziehen; bekleiden". *Münchener Studien zur Sprachwissenschaft* 27: 5–44.

Fortson, Benjamin Wynn. 2010. *Indo-European Language and Culture*: *An Introduction*. 2nd ed. Malden: Wiley-Blackwell.

HEG = Tischler, Johann. 1977-. *Hethitisches Etymologisches Glossar*. Innsbruck.

Hoffner, Harry. 1977. Studies in Hittite Vocabulary, Syntax, and Style: Hommage a M. Emmanuel Laroche. *Journal of Cuneiform Studies* 29: 151–156.

Hoffner, Harry, 2009. *Letters from the Hittite kingdom*. Atlanta: Society of Biblical Literature.

Hoffner Harry, H. Craig Melchert. 2008. *A Grammar of the Hittite Language*. Winona Lake: Eisenbrauns.

HW² = Friedrich, Johannes, Annelies Kammenhuber. 1975–. *Hethitisches Wörterbuch: Zweite, völlig neubearbeite Auflage auf Grundlage der edierten hethitischen Texte.* Heidelberg: Universitätsverlag Winter.

Kittilä, Seppo. 2009. Causative morphemes as non-valency increasing devices. *Folia Linguistica* 43: 67–94.

Kloekhorst, Alwin. 2008. Etymological Dictionary *of the Hittite Inherited Lexicon*. Leiden: Brill.

Koch, Harold. J. 1980. Indic *dabhnóti* and Hittite *tepnu-*: Etymological Evidence for an Indo-European Derived Verb Type. In: M. Mayrhofer, M. Peters, O. Pfeifer (eds.). *Lautgeschichte und Etymologie. Akten der VI. Fachtagung der Indogermanischen Gesellschaft*: 223–237. Wiesbaden: Reichert.

Kölligan, Daniel. 2004. Wenn zwei dasselbe tun: Iterativa und Kausativa. In: M. Kozianka, R. Lühr, S. Zeilfelder (eds.). *Indogermanistik – Germanistik – Linguistik. Akten der Arbeitstagung der Indogermanischen Gesellschaft, Jena 18.-20.09.2002*: 193–247. Hamburg: Dr. Kovač.

Kölligan, Daniel. 2007. Iteratives and Causatives in Latin: a Unified Approach. In: C. George et al. (eds.). *Greek and Latin from an Indo-European perspective*: 49–64. Cambridge Philological Society.

Kronasser, Heinz. 1966. *Etymologie der hethitischen Sprache*. Wiesbaden: Harrassowitz.

Kulikov, Leonid. 2013. Constraints on the causative derivation in early Vedic: Evidence for a diachronic typology of transitivity. *Poznań Studies in Contemporary Linguistics* 49: 79–101.

Kümmel, Hans Martin. 1967. *Ersatzrituale für den hethitischen König*. *StBoT 3*. Wiesbaden: Harrassowitz.

Leumann, Manu. 1977. *Lateinishen Laut- und Formenlehre*. München: Beck.

LIV = Rix, Helmut (ed.). 2001. *Lexikon der indogermanischen Verben*. *2nd ed*. Wiesbaden: Reichert.

Luraghi, Silvia. 2012. Basic Valency Orientation and the Middle Voice in Hittite. *Studies in Language* 36: 1–32.

Melchert, H. Craig. 1997. Denominative Verbs in Anatolian. In: Dorothy Disterheft, Martin E. Huld, John A. C. Greppin (eds.). *Studies in Honor of Jaan Puhvel. Part One. Ancient Languages and Philology*: 131–138. Washington D.C.: Institute for the Study of Man.

Melchert, H. Craig. 2018. Hittite *tit(ta)nu-*, *titti-*, and Lycian *stta-*. *Chatreššar* 1: 25–33.

Milizia, Paolo. 2004. Proto-Indo-European Nasal Infixation Rule. *Journal of Indo-European Studies* 32: 337–359.

Nedyalkov, Vladimir P., Georgij G. Silnitsky. 1973. The Typology of Morphological and Lexical Causatives. In: Ferenc Kiefer (ed.). *Trends in Soviet Theoretical Linguistics*: 1–32. Dordrecht: Springer.

Oettinger, Norbert. 1979. *Die Stammbildung des hethitischen Verbums*. Nürnberg: Verlag Hans Carl.

Sasseville, David. 2014–15. Luwian and Lydian agent nouns in **-é-leh$_2$*. *Die Sprache* 51/1: 105–24.

Sasseville, David. 2018. *The Pre-History of the Verbal Stem Formation of the Luwic Languages*. PhD thesis. Marburg University.

Schulze-Thulin, Britta. 2001. Zu *hi*-Konjugation von Fortsetzer urindogermanischer *-o-éi̯e/o-*Kausativa/Iterativa im Hethitischen. In: O. Carruba, W. Meid (eds.). *Anatolisch und Indogermanisch. Aktes des Kolloquiums der Indogermanischen Gesellschaft Pavia, 22.-25. September 1998*: 381–393. Innsbruck: Institut für Sprachen und Literaturen der Universität Innsbruck.

Shatskov, Andrey. 2017. *Hittite Nasal Presents*. PhD thesis. Leiden University.

Shatskov, Andrey. Ms. (forthc.). *The polysemy of the Hittite causative suffixes*.

Soysal, Oğuz, Başak Yildiz Gülşen. 2019. *Unpublished Bo-fragments in transliteration II. CHDS 3*. Chicago: Oriental Institute of the University of Chicago.

Steer, Thomas. 2013–14. Die Entstehung der indogermanischen Nasal-Infix-Präsentien. *Münchener Studien zur Sprachwissenschaft* 67/2: 197–222.

Ünal, Ahmed. 1984. Studien über das hethitische Kriegswesen II: Verba Delendi *harnink/harganu-* ‘vernichten, zugrunde richten’. *Studi Micenei et Egeo-Anatolici* 24: 71–85.

Watkins, Calvert. 1973. Hittite and Indo-European Studies: The Denominative Statives in *-ē-*. *Transactions of the Philological Society* 1971: 51–93.

Yakubovich, Ilya. 2010. *Sociolinguistics of the Luvian Language*. Leiden: Brill.

А. В. Шацков. Хеттские показатели каузатива в диахронической перспективе анатолийских языков

В хеттском засвидетельствовано несколько морфологических маркеров каузатива. Два из них, суффиксы *-nu-* и *-ahh-*, были продуктивными, а остальные маркеры либо утратили свою продуктивность, либо, как в случае с редуплицированной основой, были в первую очередь маркерами итератива. В древнехеттский период *-nu-* и *-ahh-* всё еще находились в состоянии дополнительной дистрибуции, хотя в отдельных случаях уже могли присоединяться к одним и тем же типам основ (прилагательные на *-i-*). В средне- и новохеттский периоды эти суффиксы утратили какие-либо функциональные различия между собой и могли присоединяться к одним и тем же словам. Подобное сосуществование параллельных основ на *-nu-* и *-ahh-*может частично объясняться тем, что родным языком большинства писцов был лувийский, и их знание хеттского могло быть несовершенным.

Ключевые слова: хеттский язык; лувийский язык; каузативные образования; глагольная морфология; итеративные формы; интенсивные формы.

Andrey Nikulin[†], *Rafael Andrade*[‡]
[†] University of Brasília; andre.n.guzman@gmail.com
[‡] University of São Paulo; rafael.andrade.almeida@usp.br

The rise and fall of approximants in the Tuparian languages*

This paper addresses the evolution of the approximant series in the languages of the Tuparian branch of the Tupian family, native to the region comprised between the middle course of the Guaporé/Iténez and the headwaters of the Machado/Ji-Paraná (southern Rondônia, Brazil). It is shown that in addition to the approximant series of Proto-Tuparian (which, we argue, comprised *β, *j, *w), some daughter languages created innovative approximants from a variety of sources, such as non-low vowels (*o/*i), post-oralized nasals (*mb/*nd/*$ŋg$, by the way of *b/*d/*g), and hiatus-filling glides. The evolution of these sounds is discussed in great detail; in particular, we argue that at least some approximants have been historically fortitioned in all Tuparian languages. A special attention is given to the subgrouping of the Tuparian branch.

Keywords: Tuparian languages; Tupian languages; approximants; fortition; comparative method.

1. Introduction

This paper examines the phonological development of the approximant series throughout the reconstructed history of the Tuparian languages (Tupian family), a group of indigenous languages spoken in what is now the Brazilian state of Rondônia. We will argue that Proto-Tuparian inherited a series of approximants (*β, *j, *w) from its ancestor, Proto-Tupian, which were later subject to massive fortition processes in the history of all contemporary Tuparian languages. In addition, we hypothesize that some Tuparian languages innovated at some point by creating approximants from two types of Proto-Tuparian sources: non-low vowels (*o, *i) and postoralized nasals (*mb, *nd, and *$ŋg$).

The Tupian language family is one of the most diversified and geographically disperse genetic units of South America. Its approximately 50 languages are spoken throughout a vast area which spans from the northern Amazon to the extreme south of Brazil and are classified into ten universally recognized low-level branches: Arikém, Tuparí, Mondé, Ramarama, Puruborá, Mundurukú, Juruna, Sateré-Mawé, Awetí, and Tupí-Guaraní (Rodrigues & Cabral 2012). Recent studies have shown that Ramarama and Puruborá likely constitute a valid clade (Galucio & Gabas Jr. 2002), as do Sateré-Mawé, Awetí, and Tupí-Guaraní (Awetí and Tupí-Guaraní are more closely related to each other than any of them to Sateré-Mawé; Corrêa-da-Silva 2010, Meira & Drude 2015). From a geographic point of view, the genetic diversity within the family reaches its peak in what is now the Brazilian state of Rondônia, which has therefore been identified as the likely *Urheimat* of Proto-Tupian (Rodrigues 1958: 683).

The Tuparian branch — also known in earlier literature as *Kanoé* (Rodrigues 1958: 682), *Mekens* (Hanke *et al.* 1958: 188), or *Makuráp*/*Macuráp* (Loukotka 1963: 45, 1968: 122) after different members of the branch (both *Kanoé* and *Mekens* refer to the language now known as

* We thank an anonymous JoLR reviewer for comments which helped improve the manuscript. Nikulin is grateful to Ana Vilacy Galucio for discussions on the synchronic and historical phonology of the Tuparian languages. Andrade is further grateful to Luciana Storto and Denny Moore for providing original recordings of Makurap. The usual disclaimers apply.

Mekéns or Sakurabiat) — includes the following languages. **Wayoró** (Glottocode [wayo128], ISO 639-3 [wyr]) is spoken in the Terra Indígena Rio Guaporé by three elderly speakers at the time of writing (Nogueira 2019: 3). Nogueira (2019: 4) also reports lexical and phonological differences between the varieties traditionally spoken by the Kupndiiriat and Ngwayoroiat groups. **Tuparí** (Glottocode [tupa1250], ISO 639-3 [tpr]) is spoken by 350 individuals in two reservations, Terra Indígena Rio Branco and Terra Indígena Rio Guaporé (Singerman 2018: 1). **Mekéns** (= Sakɨrabiat, Sakurabiat; Glottocode [saki1248], ISO 639-3 [skf]) is spoken by 14 individuals in the Terra Indígena Rio Mekéns (Galucio & Nogueira 2018: 96).[1] It is subdivided into three dialects, including **Sakurabiat/Guarategayat**, **Guaratira**, and **Siokweriat** (= Kampé, now spoken by just one individual). **Akuntsú** (Glottocode [akun1241], ISO 639-3 [aqz]) is spoken by three individuals near the Omerê creek (Aragon & Tavares 2019). **Makurap** (Glottocode [maku1278], ISO 639-3 [mpu]) is spoken by ca. 50 individuals in the Terra Indígena Guaporé (Galucio & Nogueira 2018: 96).

Until recently, the languages of the Tuparian branch had remained severely underdocumented. Tibor Sekelj documented short wordlists of Tuparí and Makurap during his 1948 expedition to the Rio Branco (Sekelj 1948). Emil-Heinrich Snethlage traveled around the region in 1933–4 and made notes on all Tuparian languages except Akuntsú (Snethlage 2015). Wanda Hanke visited the Mekéns in 1949 and also made some notes on the language (Hanke *et al.* 1958). Franz Caspar stayed with the Tuparí for several months in 1948 and 1955; based on his fieldnotes, a grammar sketch was prepared in 1958 (translated into Portuguese and published as Rodrigues & Caspar 2017).[2] Fortunately, the situation has improved drastically over the last 30 years due to a documentation boom in Amazonian linguistics. The following recent works have been prioritized as primary sources of lexical data used in this study. For Wayoró, we rely on the works by Nogueira (2011, 2015, 2019). For Tuparí, we give preference to Singerman's (2018) dissertation and to Alves's (2004) dictionary (especially when it comes to the position of the stress). As for Mekéns, Galucio's (1994, 2001, 2002, 2011a,b, 2014; Galucio *et al.* 2017; Alves & Galucio 2007) works have been consulted for the Sakurabiat and Guaratira dialects, whereas for the Siokweriat dialect the short appendix in Aragon (2014) was used. This latter work has also been our primary source for Akuntsú, though earlier works by the same author (Aragon & Carvalho 2007, Aragon 2008) as well as Gabas Jr. (2005) were also consulted. Finally, for Makurap we rely on Braga (1992, 2005) as well as on unpublished recordings by Denny Moore (collected in 2003 with the help of the consultant Alcides Makurap). In addition, lexical material has been extracted from the comparative works on Tuparian (presented below) whenever the relevant forms are not attested in our primary sources. For kinship terms in all Tuparian languages, we rely on Nogueira *et al.* (2019).

To this moment, however, few works have been dedicated to the phonological reconstruction of Proto-Tuparian.[3] Moore & Galucio (1994) offer a pioneering proposal of the segmental

[1] Although the language has been increasingly referred to as *Sakurabiat* in recent literature, here we reserve the label *Sakurabiat* for the dialect spoken by the Sakurabiat and Guarategayat groups. In contrast, the label *Mekéns* is used in a broader sense throughout this paper and covers the varieties spoken by the Guaratira and the Siokweriat (in addition to the one spoken by the Sakurabiat and the Guarategayat).

[2] Other premodern sources on specific Tuparian languages include Anonymous (n/d, on Wayoró, *apud* Loukotka 1963: 46–7), Becker-Donner (1955, on Mekéns, *apud* Loukotka 1963: 48), Xerez (1946, on Makurap), and Lévi-Strauss (n/d, on the *Kabišiana* variety of Mekéns, *apud* Loukotka 1963: 48 and Lévi-Strauss 1950; see Nikulin submitted for the identification of *Kabišiana* as a Mekéns variety). We were unable to access any of these works.

[3] Aragon & Cabral (2005) and Gabas Jr. (2005) also discuss the genetic relations within the Tuparian family (with special attention to the position of Akuntsú), but no claim is made with respect to the phonological reconstruction. Galucio & Nogueira (2018) reconstruct the evolution of the object focus construction in the Tuparian

phonology of Proto-Tuparian, which is based on a total of 124 cognate sets representing Wayoró, Tuparí, Mekéns, and Makurap; the respective reconstructed forms are also provided. Galucio & Nogueira's (2012) work by and large reinforces Moore & Galucio's (1994) reconstruction, differing from it mainly in that (i) the segment **nd(z)* is removed from the reconstructed inventory; (ii) the phonological status of **b* as a contrastive segment, treated as uncertain in Moore & Galucio (1994), is confirmed; (iii) the data of a fifth Tuparian language, Akuntsú, are taken into account. Furthermore, Galucio & Nogueira (2012) argue that the Proto-Tuparian segment **D* (the *ad hoc* symbol used in Moore & Galucio 1994) should be interpreted as a voiced denti-alveolar stop **d*. Galucio & Nogueira (2012) also address the reconstruction of aspects of Proto-Tuparian morphosyntax, including the person inflection and the morphosyntactic alignment, as well as derivational morphology. Due to the nature of the publication, the segment reconstructed by Galucio & Nogueira (2012) as **d* is the only one to be supported with detailed discussion and examples. The cognate sets that were used to substantiate the reconstruction of all other segments are not presented. That way, the only published work on the phonology of Proto-Tuparian in which the reconstructed phonemes are illustrated with actual linguistic data is Moore & Galucio (1994), which predates the documentation boom of the Tuparian languages by a large margin.

An in-depth study of the historical phonology of the Tuparian group, besides being an interesting subject by itself, is crucial for our understanding of the diachronic development of the entire Tupian family (cf. Galucio & Nogueira 2018: 95). Although there have been pioneering attempts at a phonological reconstruction of Proto-Tupian (Rodrigues 2002, 2005, 2007), most subgroups of Tupian have been represented in them by one single contemporary language (Tuparí for the Tuparian group, Mundurukú for the Mundurukú group, Yudjá for the Juruna group) rather than by the respective intermediate proto-languages (with the notable exception of Proto-Tupí-Guaraní). In other words, the comparative method has never been consistently applied to the Tupian family in a bottom-up manner. The situation, however, is likely to change in the near future, thanks to several recent and ongoing detailed, methodologically sound diachronic studies of low-level branches of Tupian (most recently Meira & Drude 2015 for Proto-Mawé-Guaraní, Carvalho 2019 for Proto-Juruna, Picanço 2019 for Proto-Mundurukú, Carvalho forthc. for Proto-Tupí-Guaraní). In this sense, this paper aims to contribute to the emergent field of diachronic Tupian studies in general by reconstructing parts of the consonantal system of Proto-Tuparian in some detail.

The International Phonetic Alphabet is used for representing data in this paper, with the following exceptions. The symbols *r*, *β*, *ð*, *γ*, and *e* stand for [ɾ, β̞, ð̞, ɯ̟, ɛ], respectively. The coda consonants are considered to be underspecified for features other than place of articulation in all Tuparian languages (cf. Singerman 2016 for Tuparí). Underspecified labial, dental/alveolar, palatal, and velar consonants in coda are represented in small caps: *P*, *T*, *C*, *K* (cf. a similar analytical decision for another Tupian language, Awetí, in Drude 2009). The acute accent denotes stress in Tuparí and Akuntsú and high tone in Makurap.[4]

languages. Nogueira *et al.* (2019) is an in-depth study of the Tuparian kinship terms, which also includes reconstructed forms; the phonological reconstruction in this work does not differ substantially from that of Galucio & Nogueira (2012).

[4] The evidence for two level tones in Makurap (high and low) comes from our preliminary analysis of the Makurap recordings by Moore, wherein most words are conveniently accompanied by their whistled equivalents (cf. Moore & Galucio 1994: 122). The high tone occurs at most once in polysyllables, and its position interacts with morphology in ways that are currently poorly understood (e.g. *pãrĩõ* [pə̃^{H}ˈɾ̃ĩõ^{L}] 'hawk' → *pãrĩõ-cato* [pə̃^{L}ɾ̃ĩõ^{H}ca^{L}ˈtːo^{L}] 'harpia'). We do not mark the low tone explicitly. Makurap tokens taken from printed sources which do not transcribe the tonal distinctions (such as Braga 1992, 2005) are underlined.

The remainder of this paper is structured as follows. In section **2**, we present some evidence for the subgrouping of the Tuparian group accepted in this paper. In section **3**, we present the comparative evidence which supports the reconstruction of the approximant series for Proto-Tuparian as well as for the proto-languages of shallower genetic units, such as Proto-Core Tuparian and Proto-Corumbiara, and for an earlier stage of Wayoró. Specific sound changes required by our proposal are summarized in section **4**. We conclude by a succinct discussion of our findings in section **5**, followed by a list of **abbreviations** used in this paper.

2. Internal classification of the Tuparian group

In this section, we present evidence for a specific proposal regarding the subgrouping of the Tuparian group. Namely, we claim that (i) Wayoró and Tuparí form a subgroup to the exclusion of other languages ("Wayoró–Tuparí"), (ii) Mekéns and Akuntsú likewise form a subgroup to the exclusion of other languages ("Corumbiara"), and (iii) all the aforementioned languages form a clade ("Core Tuparian") to the exclusion of Makurap.

Makurap vs. Core Tuparian. The claim regarding the binary split of Proto-Tuparian into Makurap and Core Tuparian has found extensive support in a number of published works (cf. the lexicostatistical assessment in Galucio & Nogueira 2012, Galucio *et al.* 2015: 238), even though little space has been allocated so far to the identification of innovations shared by the Core Tuparian languages. The most characteristic of them are listed below.

One such innovation appears to have affected the third person inflection pattern of the */j/-initial stems. In Makurap, a significant number of stems inflect for the third person by replacing their initial consonant (*c-* in oral environments; *ɲ-* is nasal environments) with another consonant (*t-* both in oral and nasal environments), as in *ceK* 'house.POSS', *ɲãC* 'tooth' → *t-eK* 'his/her house', *t-ãC* 'his/her tooth' (Braga 2005).[5] A plausibly cognate pattern is found in Tupian languages outside the Tuparian group, such as Mundurukú (*dək-ʔá*, *nõj* → *t-ək-ʔá*, *t-õj*; Picanço 2005), Kuruaya (*l-* → *t-*), Sateré-Mawé (*s-* → *h-*), and most Tupí-Guaraní languages (**t-*/**-r-* → **ts-*). This allows us to project the pattern attested in Makurap onto the Proto-Tuparian level (we reconstruct PTpr **j-*/**ɲ-* → **c-*, where **ɲ* is the nasal allophone of */j/). All other Tuparian languages lost the archaic prefix **c-* and now use reflexes of PTpr **i-* in this function. For example, the third person of PTpr **jaʔɨP* 'son, fraternal nephew (male ego)' (> Makurap *caɨP*) is reconstructed as **c-aʔɨP* (> Makurap *t-aɨP*). In Proto-Core Tuparian, the unpossessed form yielded **ðaʔɨP* (> Wayoró *ndaʉP*, Tuparí *haʔʉ́P*, Mekéns/Akuntsú *taɨP*); see **3.1** for PTpr **j* > Proto-Core Tuparian **ð*. However, the third person form was not preserved as **c-aʔɨP* but rather was substituted with **i-ðaʔɨP* (> Tuparí *i-aʔɨP*, Mekéns *i-taɨP*, etc.). The allomorph **i-* must have been extended through analogy from other consonant-initial stems.

Another clear innovation that identifies Core Tuparian as a valid genetic unit is the nasalization of the stops **p* and **t* in syllables with nasal nuclei, as shown in Table 1.

In turn, Core Tuparian is subdivided, in a binary manner, into Mekéns–Akuntsú and Wayoró–Tuparí. The former claim seems to be universally accepted (Gabas Jr. 2005; Galucio & Nogueira 2012), as Mekéns and Akuntsú are remarkably close to each other and are reported to be mutually intellegible (Galucio *et al.* 2015: 237–8 even suggest that they are "co-dialects of the same language"). We propose the label *Corumbiara* for the clade which comprises Mekéns

[5] Braga (1992, 2005) transcribes the palatal obstruent of Makurap as *tʃ*. In Moore's field recordings of Makurap, we found its pronunciation closer to [c] or [cç]. Throughout this paper, it will be represented as *c*.

PTpr	gloss	Wayoró	Tuparí	Mekéns	Akuntsú	Makurap
**pãT-* **pãT-ʔa-*	to be tied to tie	—	—	—	— *mãra-*	*pãT-* *pãrã-*
**pĩːT*	humming-bird	*mĩT* *mĩːT* 'penis'	*mĩːT*	—	*mĩːT*	*pĩːT*
**ãrãpĩrã*	woman	*ãrãmĩrã*	*ãrãmĩrã*	*ãrãmĩrã*	*ãrãmĩrã*	*ãrãpĩjã*[6]
**-pĩã*[7]	knee	*{kɯ̈}mĩã*	*mĩã{K-ʔã}*	*{nẽkiwa}mĩŋa* 'elbow'[8]	*{a}mĩnã*	*{ka}pĩã*
**pãrĩ(õ)*	harpia	—	*poTʔa-mãrĩ*	—	—	*pãrĩ{õ}*
**tĩ*[9]	ashamed	*nĩ-*	*nĩ-*	—	—	—
**atɨ̃P*	head	—	*anɯ̈P* 'brain'	*anɨ̃P*	*anãP*	*átɨ̃P* 'hair'
**tĩK-*[10]	to weave	*nĩK-*	*nĩK-*	*nĩ-a*	*nĩ-a*	*tĩK-*
**tĩK*[11]	spotted	*nĩK~nĩK*	*nĩK*	—	*nĩK* 'striped'	—
**tĩ(:)K*	*timbó* vine	*nĩːK*	*nĩ(:)K*	—	—	*tĩK*
**atĩK*[12]	worm	*anĩK*	*anĩK* 'leishmaniasis ulcer'	Sio *anĩK*	*anĩP*	*atĩK*

Table 1. Nasalization of **p*, **t* in the Core Tuparian languages in nasal environments

and Akuntsú. As for Wayoró and Tuparí, the special proximity between these two languages has been suggested in Galucio *et al.* (2015) based on an application of two distance-based algorithms to the 100-word Swadesh lists of the Tupian languages (83.7% confidence rate), but this result was not replicated for other datasets considered in the cited work. In what follows, we identify several shared innovations which support the validity of both branches (Corumbiara and Wayoró–Tuparí).

Corumbiara. The Corumbiara languages (Mekéns and Akuntsú) share multiple lexical innovations which are unique to these two languages. The following are some examples thereof: PTpr **ŋge* 'garden' is replaced with Mekéns/Akuntsú *tabɨT* 'garden'; PTpr **ŋgĩtaK ~ *ŋgitaK* 'night' is replaced with **matso* (Mekéns *mãtso{pi}*, Akuntsú *mãtʃo*); PTpr **mãŋĩ* 'manioc' is replaced with **taPjVT* (Mekéns *taPtsɨT*, Akuntsú *taPtoT*); PTpr **ekɨP* 'arrow' is replaced with **mãpi*

[6] The correspondence between Core Tuparian **r* and Makurap *j* is not known to be regular.

[7] Hereinafter, the curled brackets denote material which is deemed not to be cognate despite not being demonstrably segmentable in the contemporary languages.

[8] The form is tentatively phonologized based on Snethlage's (2015: 518) attestation of ‹kina kiwamínja› 'Ellbogen' (likely the first person inclusive *ki-{nẽkiwa}mĩŋa*).

[9] This root appears to have been lost in Makurap, unless *mẽtiã* 'ashamed' (Braga 2005: 191) is somehow related. A voiceless dental stop is reconstructed in light of the external cognates (Proto-Mawé-Guaraní **tĩ* 'ashamed', Meira & Drude 2015: 292).

[10] In Mekéns and Akuntsú, the only attested forms of these verbs contain the theme vowel *-a*, which triggers the deletion of the stem-final *-K*. The underlying stem is expected to have the shape *nĩK-* in both Corumbiara languages, but the forms that could prove it have not been attested in the published works.

[11] No cognate is attested in Makurap. A voiceless dental stop is reconstructed in light of the external cognates, such as Sateré-Mawé *tĩK* 'spotted' (Ribeiro 2010: 87).

[12] The Mekéns reflex is attested as *atĩK* in Galucio *et al.* (2015: 266) but not in other sources on the language that we consulted. If the existence of this form is confirmed, it could be explained as a borrowing from Makurap *atĩK* or from Karo *atĩŋ*, attested *ibidem* (the Makurap reflex is given there as *atĩ*, which must be a mistranscription, cf. Braga 2005: 184 and Moore's field data). Note that Akuntsú *-P* does not regularly continue PTpr **-K*.

'arrow' (likely from Kwaza *mãbi* or Kanoê *mapi*; cf. Voort 2005: 386). Some further examples of lexical isoglosses specific to the Corumbiara languages are **tsaro* 'yellow', **pɨ(:)K* 'black', and **kɨCpiT* 'fish', though we have been unable to provide an unequivocal Proto-Tuparian reconstruction for these specific concepts. In addition, the Corumbiara languages share multiple phonological innovations, some of which are exclusive to this subgroup (e.g. PTpr **t/*nd* > **ts* > Mekéns *ts*, Akuntsú *tʃ*; PTpr **j/*c* > **t* > Mekéns *t*, Akuntsú *t*; PTpr **i(ʔ)V* > **ijV* > Sakurabiat *itsV*, Guaratira/Siokweriat *iV*, Akuntsú *itV*). Most phonological innovations that characterize the languages of the Corumbiara branch will be discussed in more detail in section **3**.

Wayoró–Tuparí. Wayoró and Tuparí are not as tightly related to each other as Mekéns and Akuntsú, but nevertheless clear innovations shared exclusively by these two languages can be identified. For example, PTpr **mbo-ape* 'fingernail' and **ojaT* 'fire', whose reflexes are found in Makurap, Mekéns, and Akuntsú, are replaced with Proto-Wayoró–Tuparí **kĩrĩɲã* 'fingernail' and **akoP-k-aP* 'fire', respectively (the latter is evidently an **-aP* nominalization from a verbal derivative of **akoP* 'hot'). Wayoró and Tuparí are also unique in that they have high central **rounded** vowels /ʉ ʉ̃/ (Alves 2004: 41; Singerman 2016: 456; Nogueira 2019: 10), which correspond to /ɨ ɨ̃/ in Mekéns, Akuntsú, Makurap, and many Tupian languages outside the Tuparian branch. In both languages, /ʉ i/ pattern together in that they make up the environment for at least one phonological process (the diachronic assibilation **t* > *s* _/ʉ,i/ in Tuparí; the morphophonological dissimilation /e/ → *a* _/ʉ,i/ in Wayoró, cf. Nogueira 2015).

3. Proposal

In this section, we present the evidence which supports the reconstruction of the approximant series in Proto-Tuparian (**3.1**). We will also show how innovative approximants arose from various sources through multiple independent innovations in individual Tuparian languages (from non-syllabic vowels, **3.2**), in early Wayoró (from postoralized nasals, **3.3**), and in Proto-Corumbiara (**3.4**, as hiatus-filling glides). Note that in the contemporary Tuparian languages most of the segments under discussion have changed to some other sounds (either through fortition or through lenition). For example, PTpr **ejɨ* '*marico* bag' and **wawo* 'sweet potato' are reflected as Wayoró *endʉ*, *ŋgʷago*; Tuparí *énʉ*, *wáo*; Mekéns *etɨ*, *kʷa(:)ko*; Akuntsú *etɨ́*, *kʷakó*; Makurap *écɨ*, *βaβó*. For now, these examples should suffice to give the reader a flavor of the trends in the evolution of the approximants in the individual histories of the Tuparian languages. A detailed discussion thereof is deferred to section **4**.

Note that in this paper we do not discuss the consonants that arose from combinations of an underspecified consonant in the coda position followed by an onsetless or a ʔ-initial syllable. In these environments, codas are commonly resyllabified as onsets in the contemporary Tuparian languages, which is usually accompanied by lenition (Braga 1992: 63–4 for Makurap; Galucio 1994: 991–2, 2001: 23 for Mekéns; Singerman 2018: 372–3, 378–80 for Tuparí, among others). An investigation of the approximants that may have developed in some languages through resyllabification of codas lies beyond the scope of this paper.

In Table 2, we summarize our proposal regarding the development of the Proto-Tuparian onsets in oral environments. Note that *[mb], *[nd], and *[ŋg] are held here to be allophones of underlying */m n ŋ/ in oral environments (see **3.3**).

In Table 3, we summarize our proposal regarding the development of the Proto-Tuparian onsets in nasal environments. Note that *[ɲ] is held here to be the nasal allophone of an underlying */j/.

PTpr	Mak	PCT	pre-PCor	Mek/Aku	PWT	Tup	pre-Way	Way
*p	p	*p	*p	p	*p	p s-/-Ps-[A]	*p	p
*[mb]	[mb]	*b			*b/*β		*β	[mb]-/-β-
*β	p	*β	*b	b				
*t	t	*t	*t	*Mek* ts, *Aku* tʃ	*t	t, s[B]	*t	t
*[nd]	[nd]	*d			*d		*ð	[nd]
*j	c	*ð	*d	t	*ð	h-/-Ø-		
*c	t	*c	*c		*c	Ø	*tʃ	tʃ
*k	k	*k	*k	k	*k	k	*k	k
*[ŋg]	[ŋg]	*g			*g		*ɣ	[ŋg]
*w	β	*w, *ɣ[C]	*g[w], *g[C]	k[w], k[C]	*w, *ɣ[C]	w, Ø[C]	*w, *ɣ[C]	[ŋg[w]], [ŋg][C] *Kup* β
*r	r	*r	*r	r	*r	r	*r	r
*ʔ	ʔ ~ Ø	*ʔ	*ʔ ~ *Ø	ʔ ~ Ø	*ʔ	ʔ	*ʔ ~ *Ø	ʔ ~ Ø

[A] = before *i*; [B] = before *i* or *ʉ*; [C] = before a rounded vowel

Table 2. Proto-Tuparian onsets and their reflexes in oral environments

PTpr	Mak	PCT	pre-PCor	Mek/Aku	PWT	Tup	pre-Way	Way
*p	p	*m	*m	m	*m	m	*m	m
*m	m							
*t	t	*n	*n	n	*n	n	*n	n
*n	n							
*[ɲ]	ɲ	*ɲ	ɲ	ɲ	*ɲ	ɲ, Ø[A]	*ɲ	ɲ
*c	t	?[13]	?	?	?	?	?	?
*k	k	*k	*k	k	*k	k	*k	k
*ŋg	ŋg	*g			*g		*ɣ	ŋ
*w	m	*w, *ɣ[B]	*g[w], *g[B]	k[w] (*Sak* ŋ[w]), k (*Sak* ŋ)[B]	*w, *ɣ[B]	w, Ø[B]	*w, *ɣ[B]	ŋ[w], ŋ[B]
*r	r	*r	*r, *n[C]	r, n[C]	*r	r	*r	r
*ʔ	ʔ ~ Ø	*ʔ	*ʔ ~ *Ø	ʔ ~ Ø	*ʔ	ʔ	*ʔ ~ *Ø	ʔ ~ Ø

[A] = before *ĩ*; [B] = before a rounded vowel; [C] = between front vowels

Table 3. Proto-Tuparian onsets and their reflexes in nasal environments

[13] Although PTpr **c* certainly occurred in nasal environments (as in **mĩcõ* 'curassow', **c-ãC* 'his/her tooth', **c-ẽrĩ* 'his/her hammock' > Mak *mĩtő*, *t-ãC*, *t-ẽrĩ*), we have not identified reflexes of any of these word forms in any Core Tuparian language.

A note on the reconstruction of PTpr *t

Although a detailed discussion of the reconstruction of Proto-Tuparian voiceless segments is beyond the scope of this paper, we deem it appropriate to briefly comment on our interpretation of the sound correspondence between Wayoró *t*, Tuparí *t* (*s* _/i,ʉ/), Mekéns *ts*, Akuntsú *tʃ*, and Makurap *t*. Galucio & Nogueira (2012) claim that the correspondence set in question "clearly reconstructs as the affricate **ts*, which becomes [+palatal] in Akuntsú, and loses the feature [+sibilant] in Wayoro, Makurap and Tupari, except before [i] in Tupari".[14] We believe that the reconstruction should be amended to **t* for four reasons.

(i) First of all, reconstructing **ts* would imply an innovation (**ts* > *t*) shared by Makurap, Tuparí, and Wayoró, though Makurap is not known to be closely related to Tuparí and Wayoró. No such problem arises if **t* is reconstructed; in this case, we would only need to assume that PTpr **t* yielded an affricate in the Corumbiara languages.

(ii) Note that what we reconstruct as *[nd] (the oral allophone of PTpr */n/) also yielded an affricate in Mekéns/Akuntsú and *t* (*s* before non-back high vowels) in Tuparí. In our current proposal, this is straightforwardly accounted for: all PTpr postoralized nasals (**mb*, **nd*, **ŋg*) became voiced stops in Proto-Core Tuparian (**b*, **d*, **g*), which subsequently merged with PTpr voiceless stops (**p*, **t*, **k*) in the Corumbiara languages (yielding **p*, **ts*, **k*) and in Tuparí (*p*/*(P)s*, *t*/*s*, *k*). No elegant explanation of the sort is available if one accepts the reconstruction of PTpr **ts*.

(iii) There is no competing identity correspondence that could potentially involve PTpr **t* in onsets, except for two isolated etymologies: 'chicha' (Wayoró *tɨʉero*, Mekéns *tɨero*, Akuntsú *tɨeró*) and 'daughter' (Tuparí *haK*, Wayoró, Mekéns, Akuntsú, Makurap *taK*). The former item is a Wanderwort (compare Arikapú *tʃuerə*, Kanoê *tsero*; Voort 2005: 381, fn. 28, 2007: 138, fn. 4) and is thus likely to have diffused into at least some of the Tuparian languages through horizontal transmission. As for the word for 'daughter', the correspondence is unique and thus cannot back up alone the reconstruction of PTpr **t*. In fact, it is possible to reconstruct **caK* (no other examples for **c-* in the word-initial position are known, so the reflexes *h-* in Tuparí and *t-* in Wayoró could be regular).

(iv) Finally, the external correspondences of what we reconstruct as PTpr **t* are dental/alveolar stops throughout the Tupian family (Karitiana, Karo, Proto-Mawé-Guaraní *(*)t*, Puruborá *d*, as well as *t* alternating with *n* in the Mondé languages), even though admittedly affricate reflexes are also attested in non-palatalizing environments in languages such as Yudjá (*tʃ-*), Mundurukú (*tʃ-/-dʒ-*), Kuruaya (*tʃ-*) and Proto-Tupí-Guaraní (**ts* ~ **tʃ*). At least Proto-Tupí-Guaraní demonstrably innovated its affricate from Proto-Mawé-Guaraní **t*, as shown not only by external comparanda (Meira & Drude 2015) but also by the fact that the Proto-Tupí-Guaraní affricate **ts* changes to **nd* in the environment **Ṽ_* (cf. **tso* 'to go' and its causative **mõ-ndo* 'to send'). Together, all these facts point to Proto-Tupian **t* as the probable ancestor of what we reconstruct as PTpr **t*, making our reconstruction more credible. For examples and a detailed discussion of the reflexes of Proto-Tupian **t*, see Nikulin and Carvalho (2019: 276–8).

3.1. Proto-Tuparian approximants

In this section, we justify the reconstruction of three approximant phonemes for Proto-Tuparian: */β/, */j/, and */w/. All of these appear to have been straightforwardly retained from Proto-Tupian **β*, **j*, and **w*. Preceding nasal nuclei, */j/ was realized as a nasal stop *[ɲ], which

[14] In fact, the sibilant reflex in Tuparí is conditioned not only by *_i*, but also by *_ʉ* (as in *sʉT* 'peach palm', *sʉT-* 'to cook' < PTpr **tɨT*, **tɨT-*).

is retained in all contemporary languages (unlike the oral allophone [j], which underwent major changes in all daughter languages; see below). In our reconstructions, we represent the allophones of */j/ as **j* and **ɲ*, respectively, in order to highlight the fact that their default reflexes in all daughter languages are so different from each other that they are no longer synchronically analyzed as allophones of the same phoneme. As for the labiovelar approximant */w/, it is likely that it was phonetically nasalized in nasal environments (i.e., [w̃]); this is, however, not represented in our reconstructions, because in most contemporary languages the reflexes of *[w] and *[w̃] are reasonably similar to each other. The bilabial approximant */β/ is not attested in nasal environments, which may be a spurious gap, given that */β/ is an extremely rare segment in our corpus.

All Proto-Tuparian approximants, with the exception of the nasal allophone of */j/, were frequent targets of multiple fortition (and, to a lesser degree, lenition) processes, which operated in the history of each contemporary language to differing extents. These will be systematized in section 4. Nevertheless, the directionality of each sound change can be quite securely identified thanks to converging internal and external evidence, so that in each case one can be sure that the segments in question were indeed articulated as approximants in Proto-Tuparian. In subsections 3.2–4, we will discuss a number of correspondence sets for which the contrary holds: segments other than approximants (such as nasal stops or high vowels) are reconstructed for Proto-Tuparian, and it is shown that they gave rise to innovative approximants in specific languages or subgroups of Tuparian.

PTpr *β. In Table 4, we show both secure etymologies which instantiate PTpr **β*. The original articulation is preserved in Wayoró only. In Tuparí and Makurap, PTpr **β* merged with PTpr **p*, yielding Tup *p* (*-ps-* before *i*) and Mak *p*. In the Corumbiara languages, PTpr **β* is reflected as *b*. Despite the extreme scarcity of the relevant cognate sets, the reconstruction of PTpr **β* gains some credibility in light of the fact that it corresponds to Proto-Tupí-Guaraní **β*, as in PTG **toβa* 'face' and **ɨβɨtu* 'wind' (~ PTpr **jeβa*, **ɨβijo*; cf. Mello 2000: 183, 207). Note that PTG **β* is also a low-frequency segment (at least morpheme-internally): in Meira & Drude's (2015) corpus of Mawé-Guaraní etymologies, it appears only in PTG **uruβu* 'vulture' (~ PTpr **oroPʔo*), **jaβoti* 'tortoise' (no cognate in Tuparian), in the aforementioned **toβa* 'face' and **ɨβɨtu* 'wind', as well as in **aβati* 'maize', borrowed from a Cariban or other North Amazonian source (Rodrigues 1985: 389).[15]

PTpr	gloss	Wayoró	Tuparí	Mekéns	Akuntsú	Makurap
			oral			
**jeβa*	forehead	—	*épa* 'eye'	—	*eba-pé*	*cépa*
	CT only:					
**jeβa-jopaP*	eye (**jopaP* 'grain')	*eβa-paP*	—	*eba-opaP*	*eba-páP*	—
**jeβa-pi*	face (**pi* 'inner')	—	*épa-psi*	Sio *eba-pi*	*eba-pí*	—
**ɨβijo*	wind	—	*ɯPsió*	—	—	—

Table 4. Proto-Tuparian */β/ (oral **β*, unattested in nasal environments)

In our current proposal, PTpr **β* is reflected as *p* in Makurap, thus paralleling the fortition and devoicing of PTpr **j* to Makurap *c* (see below). We have also considered an alternative

[15] In addition, Meira & Drude (2015: 295) give PTG **taβa* 'village', but a more correct reconstruction would be **taP* (cf. Mello 2000: 195), of which **taβ-a* is an inflected form (the so called *argumentative case*). Its Tuparian cognate is **ja(:)P* 'village'. Meira & Drude (2015: 294) also note that PTG **ɨβõ* 'to shoot' has a cognate in Awetí and reconstruct Proto-Awetí-Guaraní **(ʔ)ɨβõ*, but no cognate is known in Sateré-Mawé.

scenario, whereby PTpr *β would have been regularly preserved in Makurap as β. This possibility is prompted by Nogueira *et al.*'s (2019: 39, 41) reconstruction of two kinship terms: PTpr *$a\beta i$ 'father (vocative)' (> Way $a\beta i$, Tup aP$s\acute{i}$, Mek *abi(-to*P*)*, Mak $\acute{a}\beta a$) and *$a\beta atso$ 'grandfather' (> Way $e\beta ato$, Mek *abatso*, Aku *abatʃó*, Mak $a\beta\acute{a}to$). Regarding the former term, note that Makurap *a* is not a regular reflex of PTpr **i*, which entails that Proto-Core Tuparian *$a\beta i$ is likely **not** to be cognate with Mak $\acute{a}\beta a$. As for the term for 'grandfather', there is evidence that the reconstruction should be amended to *joP*-ato* (literally 'father-big'), with an irregular development of the root vowel **o* in all languages except in the Tuparí compound *mẽ*P*sir-ob-ató* 'father-in-law (female ego), lit. son's grandfather' (cf. *ha*ʔ*ɯ*P*-b-ató* 'father-in-law (male ego)', in which the same vowel was irregularly lost). That way, the bilabial approximant found in Way $e\beta ato$ or Mak $a\beta\acute{a}to$ arose through resyllabification of a coda *P. As noted above, we are not concerned with such resyllabified approximants in this paper for lack of space.

PTpr */j/. The reader has already seen that we take PTpr **j* and **ɲ* to be surface realizations of PTpr */j/ in oral and nasal environments, respectively. While **ɲ* retained its articulation in all daughter languages, **j* shows more divergent reflexes. We assume that it preserved its palatal articulation in Makurap but became a voiceless affricate in this language (thus, PTpr **j* > Mak *c*). In Proto-Core Tuparian, conversely, it appears to have preserved its manner of articulation but changed its place of articulation from palatal to dental (that is, PTpr **j* > PCT **ð*).[16] In Wayoró, PCT **ð* became an underlying nasal stop /n/ (which surfaces as [nd] in oral environments), which parallels precisely other developments reconstructed for this language: PCT **w* > Way /ŋ(ʷ)/ (see below in this subsection), PCT **o-* > pre-Way **β-* > Way /m-/ (see **3.2**), and PCT **i-* > pre-Way **j-* > Way /ɲ-/ (see **3.2**). In Tuparí, one finds the reflex *h-* word-initially and *-Ø-* word-internally. In the Corumbiara languages, PCT **ð* became a homorganic stop *t*, as all other approximants (**β* > *b*; **w* > *k*$^{(w)}$; **j* > Mek *ts*/Aku *t*; see this subsection and **3.4**).

The correspondence Way *nd* ~ Tup *h*/*Ø* ~ Mek/Aku *t* ~ Mak *c* has not been previously claimed to continue the same underlying segment of Proto-Tuparian as the correspondence Way/Tup/Mek/Aku/Mak *ɲ*. Important evidence for lumping them together comes not only from the fact that they occur in a complementary distribution (in oral vs. nasal environments, respectively), but also from the fact that their reflexes in Makurap show identical behavior when they occur as the initial segments of relational stems. Namely, both Mak *c-* and *ɲ-* may derive relational stems from absolute ones (e.g. *e*K 'house', *ẽrĩ* 'hammock' → *c-e*K 'house.POSS', *ɲ-ẽrĩ* 'hammock.POSS'; Braga 2005: 48sqq.). In addition, whenever these segments occur in the beginning of a relational stem (either derived from an absolute stem or underived), they are replaced with the 3rd person prefix *t-*. External evidence unequivocally shows that the correspondences Way *nd* ~ Tup *h*/*Ø* ~ Mek/Aku *t* ~ Mak *c*, on the one hand, and Way/Tup/Mek/Aku/Mak *ɲ*, on the other hand, go back to a single consonant of Proto-Tupian. For example, in Karo, both correspond to *j* (e.g. Proto-Tuparian **jajo* 'armadillo', **jaote* 'peccary', **ja(:)ko* 'lizard', **jao* 'stingray', **ɲãc* 'tooth', **wãkɨɲã* 'agouti', **ɲõkãt* 'toucan' ~ Karo *jajo*, *jate*, *jaʔo*, *jaw*, *jãj*, *wakãja*, *jokãn*; data from Gabas Jr. 1999). In Proto-Mundurukú, the regular correspondence is **ð* (e.g. Proto-Tuparian **jajo* 'armadillo', **jaote* 'peccary', **ja(:)ko* 'lizard', **jake*K 'army ant', **ɲãc*

[16] Note that unconditional dentalization of **j* to *ð* is known from the phonological histories of many Amazonian languages, such as Shiwilu (Kawapanan; Valenzuela Bismarck 2011: 279–80) and Guarasugwe (Tupí-Guaraní < Tupian; Ramirez *et al.* 2017: 432). In Kubeo (Tukanoan), the reflex of Proto-Tukanoan **j* is realized as [ð] between non-high vowels (Chacon 2014: 65sqq). Several Cariban languages, such as Venezuelan Kari'ña, Pemón, and Makuxí, reflect Proto-Cariban **j* as [ð] at least in some environments. We know of no Amazonian language in which an opposite development (i.e., **ð* > *j*) would be claimed to have taken place in a non-palatalizing environment (an anonymous reviewer rightly remarks that outside the Americas, such a development is attested between vowels in many Turkic languages). This yields additional support for our hypothesis regarding PTpr **j* > PCT **ð*.

'tooth', **ɲõK* 'chigoe flea', **ɲẽT* 'faeces' ~ Proto-Mundurukú **ðg̱jðo(ʔ)*, **ðg̱ɟe(ʔ)*, **ðáʔo*, **ðaʔɨk*, **ðãj*, **ðõŋ*, **ðãn*; Picanço 2019). In Proto-Tupí-Guaraní, the default correspondence in the word-initial position is **t-* (e.g. PTpr **jajo* 'armadillo', **jakeK* 'army ant', **ɲãC* 'tooth', **ɲõK* 'chigoe flea', **ɲõkãT* 'toucan' ~ PTG **tatu*, **taʔoK*, **tãC*, **tũK*, **tũkãT*; cf. Mello 2000).

In Table 5, we list all known Proto-Tuparian tokens which instantiate PTpr **j* and do not show significant irregularities in the development of this consonant in the daughter languages.

PTpr	gloss	Wayoró	Tuparí	Mekéns	Akuntsú	Makurap
oral (> Proto-Core Tuparian **ð*)						
**japo: ~ *jambo:*	shimbillo (CT only)	—	*hapó*	*tapo:*	—	—
**jajo*[17]	armadillo	*ndato*[18]	—	*tato*	*tató*	—
**ja(:)ko*	lizard	—	*hakó*	*ta:ko*	—	*cáko*
**jakeK*[19]	army ant	? *akeK*	*hakék{e}*	*takeK*	—	—
**jaɨ*[20]	howler monkey	*ndaʉ*	*haʉ́*	*ta:ʔɨ*	*taɨ-kóP*	—
**jao*[21]	stingray	—	—	—	—	*cáo*
**jaote*	peccary	—	*haote-ʔirí*	*taotse*	*taotʃé*	*cáóte*
**jaP*	hair, feather	*ndaP*	*haP*	*taP*	*a-táP*	*caP*
**ja:P*	village	—	*ha:P*	*ta:P*	—	*caP*
**wejaP*	anteater (CT only)	*ŋʷẽndaP*	—	*kʷetaP*	*witáP*	—
**jaT*	snake	*ndaT*	*haT*	—	—	*caT*
**jaTʔa*	bullet ant (WT only)	*ndara*	*háTʔa*	—	—	—
**ojaT*	fire	—	—	*otat*	*otáT*	*ócaT*
**jaC*	itchiness (WT only)	*ndaC*	*peáC* *háC-ka* 'to scratch'	—	—	—

Table 5. Proto-Tuparian */j/* (oral **j*, nasal **ɲ*) (to be continued)

[17] Although no cognate in Makurap is known, PCT **ðaðo* can be securely traced back to PTpr **jajo* because precise external cognates are found all across Tupian (Karo and Puruborá *jajo*, Karitiana *sosɨ*, Proto-Mundurukú **ðg̱jðo(ʔ)*, Sateré-Mawé *sahu*, PTG **tatu*; cf. Galucio *et al.* 2015: 262). Moore & Galucio (1994: 132) give Makurap *tayto* (in our transcription, *taCto*), but we could not confirm the existence of this form in our main sources on Makurap; moreover, it does not correspond regularly to the remaning forms. Mekéns *tato* is attested by Moore & Galucio (1994: 132) and Snethlage (2015: 520, ‹tatú› 'Tatú (Gürteltier)').

[18] The expected reflex of the intervocalic PTpr **j* would be **nd*, not *t*. It is likely that that some sort of a dissimilation of the kind **NDVND… > NDVD…* applied in this word, as in **ŋgʷaŋgo > ŋgʷago* 'sweet potato' (with an additional devoicing: **nd > *d > t*; note that [d] is not part of the phonetic inventory of Wayoró).

[19] The expected reflexes in Wayoró and Tuparí would be Way **ndakeK*, Tup **hakeK*. We surmise that Way *akeK* was borrowed from an early form of Tuparí (or a variety close to it), whereas in Tuparí some sort of expressive reduplication could have applied. The Mekéns reflex is regular but is attested only in Hanke *et al.* (1958: 212) as ‹takêk›, where its regular relation to Mundurukú *daʔək* (< **ðaʔɨk*) and Tupinambá *taʔoK* is noted.

[20] Although no cognate in Makurap is known, PCT **ðaɨ* can be securely traced back to PTpr **jaɨ* because a precise external cognate is found in Karo (*jaɨ* 'howler monkey', Galucio *et al.* 2015: 267). The Wayoró reflex is also attested as *ndeʉ* (Nogueira 2015: 617). Akuntsú *-koP* stands for 'red'.

[21] Although no cognate in Core Tuparian is known, Mak *cáo* can be securely traced back to PTpr **jao* because a precise external cognate is found in Karo (*jaw* 'stringray', Gabas Jr. 1999: 13). Galucio *et al.* (2015: 272) also cite possible cognates in Puruborá, Yudjá, and Xipáya.

PTpr	gloss	Wayoró	Tuparí	Mekéns	Akuntsú	Makurap
*ja?ɨP	son, fraternal nephew (male ego)	ndaɯP	ha?ɯ́P	taɨP	taɨP	caɨP (also 'sperm')
*joP	father	ndoP	ho(:)P	toP	toP	coP
*jo(:)P[22]	to lie	ndoP-	—	to:P-	to-a	—
*jo-aP	hammock (CT)	ndo-aP	o-áP ~ w-áP	to-aP	to-aP	—
*jo?oP	that (sitting)	—	ho?oP	—	—	co:P
*jɨrɨ	two (CT only)	ndɯrɯ-T	hɯrɯ́ 'pair'	tirɨ	tírɨ	—
*j-eP	leaf.POSS[23]	ndeP	heP	teP	t-eP	ceP
*jeT	name	ndeT	heT	teT	teT	ceT
*jeT?oP[24]	rubber	—	heróP	—	tedó	coroP
*j-eK	house.POSS	nd-eK	h-eK	t-eK	t-eK	c-eK
*ji:T	flower	kɯP-ndi:T 'forest'	? hí:T 'side dish'	—	—	kɨP-cír-eT
*ji:T-?a	flower (CT only)	ndi:r-a	híT-?a	tir-a[25]	tir-á	—
*ejɨ	*marico* bag	endɯ	éɯ	etɨ	etɨ́	écɨ
nasal (> Proto-Core Tuparian *ɲ)						
*ɲã	mother (voc.)	ɲã	ɲã	ã-tsi[26]	—	ɲã
*wãkɨ̃ɲã	agouti	ŋ[w]ãkɨ̃ɲã	—	Sak mãkɨ̃ɲã Gua, Sio pakɨ̃ɲã	pakɐ̃ɲã	mãkɨ̃ɲã
*tãɲã	earring (CT only)	tãɲã	—	—	tʃãɲã	—
*kĩrĩɲã	nail (WT only)	kĩrĩɲã	kirĩɲã	—	—	—
*ɲãC	tooth	ɲãC	ɲãC	ɲãC	ɲãC	ɲãC
*ɲõkãT	toucan	—	ɲõkãT	—	—	ɲőkãT
*ɲõtaP	flea	—	ɲõtáP	—	—	ɲőtãP

Table 5. Proto-Tuparian */j/ (oral *j, nasal *ɲ) (to be continued)

[22] Although no cognate in Makurap is known, PCT *ðo(:)P can be securely traced back to PTpr *jo(:)P (as opposed to **ndo(:)P) because a precise external cognate is found in Karo (*jop-* 'to live', Gabas Jr. 1999: 127). The noun *jo-aP 'hammock' is a nominalization of this verb ('lying place').

[23] In most Tuparian languages, the only known term for 'leaf' is synchronically an underived relational noun., with no absolute equivalent attested. However, in Akuntsú one finds both an absolute (eP) and a relational (t-eP) form (Aragon 2014: 130), which is symbolized here by a hyphen.

[24] The absence of a coda consonant in Akuntsú is irregular, as is the vowel *o* in the initial syllable in Makurap.

[25] The form is tentatively phonologized based on Snethlage's (2015: 520) attestation of ‹itíra› 'Blüte' (likely the third person *i-tira*). Hanke *et al.* (1958: 212) attests ‹ôtira› 'eyebrow' (*o-* is the 1SG prefix), which is likely a semantic offshoot of the same word. The semantic development from 'flower' to 'eyebrow' has also been reconstructed for the Macro-Jê language Maxakalí, where the compound *kyC-dyT* 'eyebrow' means literally 'forehead flower' (*mĩ-dyT* 'flower', literally 'tree flower'); see Nikulin and Silva (2020: 56).

[26] The Mekéns vocative term for 'mother' is identified by Nogueira *et al.* (2019: 41) as a fossilized compound of the original vocative term for 'mother' (PTpr *ɲã* in our reconstruction) and the original referential term for 'mother' (PTpr **ti* in our reconstruction, still found in Mekéns as *tsi*), paralleling Mekéns *abi-toP* 'father (voc.)' = *abi* 'father (voc.)' + *toP* 'father (ref.)'. Deriving the Mekéns form from PTpr **ɲã-ti* is unproblematic, because the loss of stem-initial PTpr */j/ is a recurrent phenomenon regular in polysyllabic relational stems, see below.

PTpr	gloss	Wayoró	Tuparí	Mekéns	Akuntsú	Makurap
**ɲõʔõP*[27]	powder	*ɲõːP*	*ɲõʔṍP*	—	*e-ɲõP* ‘*farofa* flour’	—
**ɲõK*	chigoe flea	*ɲõK*	*ɲõK* ‘pimple’	—	—	*ɲõK*
**ɲẽT*	faeces, guts	*ɲẽT*	*ɲẽT*	*ɲẽT*	*ɲẽT*	*ɲẽT*
**ɲẽ(ː)T*	ashes	*{kɯ}ɲẽ(ː)T*	*{kɯ}ɲẽ́(ː)T*	*ɲẽːT*	*otaT-ɲẽ́T*	*{kɨ̃}ɲẽːT*
**ɲẽTʔã*	meat	*ɲẽrã*	*ɲẽ́Tʔã*	*ɲẽrã*	—	*ɲɨ̃rã*[28]
**mãɲĩ*	manioc	*mãɲĩ*	*mãC*[29]	—	—	*mãɲĩ*
**ɲĩːK*[30]	smoke	*ɲĩːK*	*{s}ĩ̃K*	*otaT-nĩːK*	*nĩ̃K*	*ocáT-ɲĩ̃ŋ-ẽ̃T*

Table 5. Proto-Tuparian */j/ (oral **j*, nasal **ɲ*)

In a subset of relational stems, an unexpected correspondence occurs between Core Tuparian vowel-initial stems and Makurap *c-* or *ɲ*-initial stems. It is worthy of note that all such stems are, at the very least, disyllabic. We reconstruct their PTpr etyma as */j/-initial and posit a shared innovation for Core Tuparian, which consists in an almost regular loss of stem-initial */j/ at the left margin of polysyllabic relational stems. We have been so far unable to explain why some polysyllabic relational stems (such as PTpr **jaʔɨP* ‘son of a male ego’, **ɲẽTʔã* ‘meat’) resisted the deletion of */j/. An anonymous reviewer suggests that the loss of */j/ should be viewed as a morphological — rather than phonological — change; however, at present we lack evidence for analyzing */j/- as a prefix (other than in the relational nouns **j-eK* ‘house.POSS’ and **j-ekɨP* ‘arrow.POSS’, which are indeed derived from the absolute nouns **eK* ‘house’ and **ekɨP* ‘arrow’). The relevant cognate sets are given in Table 6 below.

PTpr **w*. We reconstruct a labiovelar approximant **w* for Proto-Tuparian. In Makurap, it is reflected as *β* before oral nuclei and as *m* before nasal nuclei (Makurap has no /w/). In the Core Tuparian languages, it usually retains its labiovelar articulation (Way /ŋ^{w}/, Tup *w*, Mek/Aku *k*^{w}) except before rounded vowels, where one finds Way /ŋ/, Tup *∅*, Mek/Aku *k*.[31] It is thus possible that Proto-Core Tuparian */w/ had a velar allophone *[ɣ], which occurred preceding rounded nuclei. In Wayoró and in the Corumbiara languages, both allophones (PCT *[w] and *[ɣ]) underwent developments which affected most or all approximants in these languages: in Wayoró, they changed into underlying nasal stops /ŋ^{w}/ and /ŋ/ (which surface as [ŋg^{w}] and [ŋg] in oral environments), whereas in the Corumbiara languages they were

[27] Although no cognate in Makurap is known, PCT **ɲõʔõP* can be securely traced back to PTpr **ɲõʔõP* because a precise external cognate is found in Proto-Mundurukú (**ðõ̯m* ‘powder’, cf. Picanço 2005: 181).

[28] The vowel of the first syllable is attested as high in Moore’s field recordings ([ɲĩ^{H}ɾ̃ẽ^{L}]), in Galucio *et al.* (2015: 253, *ɲĩ̃ɾaʔ*), and in Sekelj (1948, ‹ñino›), but is not a regular reflex of PTpr **ẽ*. Braga (2005: 208) attests *ɲɨ̃rã*. The expected form *ɲẽrã* is in fact documented in one of our secondary sources (Moore & Galucio 1994: 133).

[29] This token instantiates the loss of PTpr **ɲ* in Tuparí preceding an *ĩ* (that way, PTpr/PWT **mãɲĩ* > **mãĩ*). The hiatus in the resulting form was resolved by reanalyzing [ĩ] as a coda offglide [j], which is the surface realization of the underspecified palatal coda *-C* in modern Tuparí (cf. Singerman 2016).

[30] The Makurap form is from Moore’s field data. The Tuparí form appars to contain the fossilized third person prefix: *s-ĩ̃K* < **j-ĩ̃K* < PCT **i-ɲĩK*; this shows that the sound change **ɲĩ* > *ĩ* preceded the fortition of PCT **j* in Tuparí. In the Corumbiara languages, the root-initial is attested as *n* (instead of the expected **ɲ*) in all sources (Aragon 2008: 53, 2014: 108; Galucio *et al.* 2015: 259) except Aragon & Cabral (2005: 1537), who attest the Akuntsú word as *otáT-ɲĩ̃T*, with the expected palatal onset but with an unexpected coronal coda.

[31] The Kupndiiriat dialect of Wayoró appears to have *β* as its default reflex, though only one reliable datum is attested, *βiri* ‘açaí’ (Nogueira 2019: 4).

			oral			
**jaɨ*	flour, crumbs	—	*haʔɯ́*	—	—	*tʃaɨ*
**ja:mbi*	crop seed	*aβi*	*a:Psí*	*a:pi*	—	*cãmbi*
**jape*	sharpened	—	*apé*	—	*apé*	*mbi-cape*
**jato*	big	—	*-ató*	*atso*	*-atʃo*	*cato*
**jato*	to bathe	*ato-*	*ató-*	*atso-*	*atʃo-*	*cato-*
**jati*	pain, to hurt; sour	*ati*	*así*	*atsi*	*atʃi*	*cáti*
**jo(ʔ)a*	brother (female ego)[32]	*ok^{w}a*	*owá ~ oʔá*	*ok^{w}a*	*ok^{w}a*	*cóa*
**jopaP*	grain	—	*opáP* 'maize'	*opaP* 'maize'	—	*copaP*
**jopiʔa*	egg	*ɯpia*[33]	*oPsíʔa*	*opitsa*	*õpíta*	*cópia*
**joaC*	tail	*ok^{w}aC*	*oáC*	*ok^{w}aC*	*ok^{w}áC*	*ɲẽT-coaC*
**jeβa*	forehead	—	*épa* 'eye'	—	*eba-pé*	*cépa*
	CT only:					
**jeβa-jopaP*	eye	*eβa-paP*	—	*eba-opaP*	*eba-páP*	—
**jeβa-pi*	face	—	*épa-Psi*	Sio *eba-pi*	*eba-pí*	—
**j-e-kɨP*	arrow.POSS	—	*e-kɯ́P*	—	—	*c-e-kɨP*
**jeɨ*	blood	*aɯ*	*éɯ*	*aɨ*[34]	*eʔɨ*	*céɨ*
			nasal			
**ɲãpi(-ʔa)*	nose	*ãpi-a*	*ãPsí*	*ãpi-tsa*	*ãpí-ta*	*ɲãpi*
**ɲãŋã*	branch	—	*ãkã*	—	—	*nãŋgã*

Table 6. Loss of stem-initial Proto-Tuparian */j/ in polysyllabic relational stems

fortitioned to /k^{w}/ and /k/, respectively. Tuparí preserved *[w] without further changes, whereas the allophone *[ɣ] yielded zero.

Unlike PTpr */j/, PTpr */w/ does not show radically different reflexes in oral and nasal environments. Only in Makurap does one consistently find different phonemes (*β* vs. *m*) as its reflxes. More marginally, the Sakurabiat dialect of Mekéns appears to sometimes have *ŋ^{w}* as the reflex of PTpr **w* in nasal environments (e.g. *ŋ^{w}ãẽ* 'pot') as opposed to *k^{w}*, which is found in oral environments in Sakurabiat and in all environments in Guaratira and Siokweriat (cf. Galucio 2001: 19). In our proposal, this is accounted for by positing a voiced stop stage in the development of the PTpr approximants in the Corumbiara languages, hence: **wãẽ* > **g^{w}ãẽ* > Sak *ŋ^{w}ãẽ*, Gua *k^{w}ãẽ*, Sio *k^{w}ãʔẽ*, Aku *k^{w}ãʔẽ*.

The Tuparian etymologies which instantiate PTpr **w* are listed in Table 7.

This concludes our presentation of the reflexes of the approximant series of Proto-Tuparian. In what follows, we present evidence for reconstructing innovative approximants for earlier stages of individual Tuparian languages and the proto-languages of low-level subgroups.

[32] In all Tuparian languages except Akuntsú, the reflexes of this kinship term also denote a woman's cousin (son of a woman's paternal uncle). In addition, Makurap *cóaC* is used for man's and woman's brothers alike (Nogueira *et al.* 2019: 42).

[33] We have no explanation for the occurrence of *ɯ* (as opposed to *o*) in this word.

[34] We have no explanation for the occurrence of *a* (as opposed to *e*) in this word. Note the similarity of this (apparently irregular) development to the dissimilation of **ei*, **eɯ* to *ai*, *aɯ* in Wayoró (Nogueira 2015).

PTpr	gloss	Wayoró	Tuparí	Mekéns	Akuntsú	Makurap
			oral			
**awa*[35]	yam	*ag^{w}a*	*awa{té}*	*ak^{w}a*	*ak^{w}á*	—
**wara(:)C*[36]	frog/toad sp. (CT only)	*ŋg^{w}ara(:)C*	*waráC-ʔa*	*k^{w}ara:C*	*k^{w}aráC*	—
**wariʔa*	bat (CT only)	*ŋg^{w}aria*	*wáriʔa*	*k^{w}aritsa*	—	—[37]
**waco*	alligator	*ŋg^{w}aCco*	*wáo*	*k^{w}ato*	*k^{w}ató*	*βáto*
**wako*	guan	*ŋg^{w}ako*	*wakó*	*k^{w}a(:)ko*	*k^{w}akó*	*βakó-pẽP*
**wakara*[38]	great egret	—	*wakara-tó* 'jabiru'	—	—	*<u>βakara</u>*
**wawo*	sweet potato	*ŋg^{w}ago*[39]	*wáo*	*k^{w}a(:)ko*	*k^{w}akó*	*βaβó*
**waʔi*	stone	*ŋg^{w}ai*	*wáʔi*	*k^{w}ai* Sio *k^{w}aʔi*	*k^{w}aʔí*	*βai*
**waT-*[40]	to go away	*ŋg^{w}aT-*	*waT-*	*k^{w}aT-*	*k^{w}aT-*	? *βaT* 'always'
**waCʔa*	labret (CT only)	—	*wáCʔa*	—	*k^{w}aCta*	—
**waK-*	to cry; sound	—	*waK-*	*k^{w}aK*	*k^{w}aK*	—
**waK-toP-*[41]	to hear	*ŋg^{w}aK-to-a*	—	*k^{w}aK-tsoP-*	*k^{w}aK-tʃoP-*	*<u>βaT-to-a</u>* 'to look'

Table 7. Proto-Tuparian */w/ (oral and nasal **w*) (to be continued)

[35] Although no cognate in Makurap is known, PCT **awa* can be securely traced back to PTpr **awa* because a precise external cognate is found in the Juruna branch (Yudjá *awa[ʔá]* 'yam', Mondini 2014: 113). The expected reflex in Wayoró would be **ãŋg^{w}a*; it is unclear why the medial consonant is oral.

[36] The vowel length is attested in the Wayoró and Mekéns reflexes by Galucio *et al.* (2015: 273). The Wayoró form is given with a short vowel in Nogueira (2011: 43, 52).

[37] The Makurap term for 'bat' is *βacáriaC*. Despite the obvious similarity to the Core Tuparian forms, there is no regular correspondence between them; Mak *βacáriaC* could go back to PTpr **wajari(ʔ)aC*. It is unclear whether we are dealing here with an irregular development or with an indirect borrowing.

[38] This token is a Wanderwort, as similar forms are found in many unrelated languages spoken all across the Amazon and even as far away as on the Caribbean coast of South America. Epps (2020, entry 'great egret') lists multiple languages of the Cariban (Carijona, E'ñepa, Wayana, Yabarana), Guahiboan (Cuiva, Sikuani, Macaguán), Arawakan (Paresí), Sáliban (Piaroa, Sáliba), Nadahup (Nadëb), and Tupí-Guaraní (Kokama, Wajãpí) families and groups as having a Wanderwort of the approximate shape %*wakara* meaning 'great egret'. One may add other Arawakan (Wapixana *wakara*, Mehinaku *wakala*, Proto-Ta-Arawakan **wak'ara*; Silva *et al.* 2013: 106, Corbera Mori 2008: 64; Nikulin & Muzykantova in prep.) as well as Tupian (Suruí-Paiter *wakár*, Zoró *wakal*, Ka'apor *wakara*; Bontkes 1978: 18, Lacerda 2014: 321, Caldas 2009: 304) languages to this list. Given the regularity of the correspondence between the Tuparí and Makurap forms, we deem it possible that **wakara* was borrowed from an unknown source into Proto-Tuparian.

[39] The expected reflex of PTpr **w* before *o* would be **ŋg*, not *g*. It is likely that that some sort of a dissimilation of the kind **NDVND… > NDVD…* applied in this word, as in **ndando > ndato* 'armadillo'. In fact, Nogueira (2019: 8, with a reference to her ongoing research) entertains the hypothesis that Wayoró [g] and [g^{w}] could be even synchronically described as allophones of /ŋ/ and /ŋ^{w}/.

[40] Even if Makurap *βaT* 'always' does not belong to this cognate set, the Core Tuparian verb cannot be regarded as a Core Tuparian innovation because precise external cognates are found across Tupian (Karitiana *hot*, Sateré-Mawé *waT* 'to go.PL').

[41] This compound can be analyzed as 'sound-see'. The development **-Kt-* > *-Tt-* in Makurap is unparalleled. The Wayoró and Makurap reflexes are attested only in their forms which contain the thematic vowel *-a-*, which triggers the deletion of the stem-final consonant. Clear cognates of PTpr **waK* are found in other Tupian languages as well (e.g. Karitiana *hok* 'to play violin', Sateré-Mawé *waK* 'to cry').

PTpr	gloss	Wayoró	Tuparí	Mekéns	Akuntsú	Makurap
**wɨ*[42]	blow	—	—	—	—	*βɨ*
**ãwɨ-*	to enter (WT only)	*ãŋgɯ-*	*ãɯ́-*	—	—	—
**wo(:)*[43]	thorn	*ŋgo:*	—	*ko*[44]	*ko* 'fish-hook'	—
**a(:)wo*	bone	—	—	*a:ko*	—	*ao*[45]
**wora*	sound, speech (?)	*ŋgora* 'music'	—	—	—	*βorá-pi* 'mouth'
**woroa-*	to look for (CT only)	*ŋgora-*	*óroa-*	*kora-*	*kóra-*	—
**aworo*[46]	parrot	—	*áwro ~ áoro*	—	—	—
**woP*	red	*ŋgoP*	*oP*	*koP*	*koP*	*βoP*
**woT-kɨP*	neck	*ŋgoT-kɯP*	*oT-kɯ́P*	*koT-kɨP* Sio *kɨT-kɨP*	*pɨT-kɨP*[47]	*βóT-kɨP*
**wetoK*	far	*ŋg*^{w}*etoK*	? *toK*	*k*^{w}*etsoK*	—	*βétoK*
**wereP*	foreigner (CT only)	*ŋg*^{w}*ereP*	—	*k*^{w}*ereP*	*k*^{w}*eréP* 'dark'	—
**wejaP*	anteater (CT only)	*ŋ*^{w}*ẽndaP*	—	*k*^{w}*etaP*	*k*^{w}*itáP*[48]	—
**weP-*[49]	to go up	*ŋg*^{w}*eP-*	—	*k*^{w}*eP-*	*k*^{w}*eP-*	—
**wi*	ax	—	*wi(:)*	*k*^{w}*i*	*k*^{w}*i*	*βi*
**ara:wi*	peanut (CT only)	*ara:g*^{w}*i*	—	*ara(:)k*^{w}*i*	*arak*^{w}*í*	—[50]

Table 7. Proto-Tuparian */w/ (oral and nasal **w*) (to be continued)

[42] Although no cognate in Core Tuparian is attested in our primary sources, Mak *βɨ* can be securely traced back to PTpr **wɨ* because precise external cognates are found across Tupian (e.g. Karitiana *he:* 'to blow'). A likely cognate in Tuparí, *ɯ-* 'to blow, to play a wind instrument' is mentioned by Rodrigues (2002: 291), but we were unable to locate this form in our primary sources on Tuparí, thus putting its existence in doubt.

[43] Although no cognate in Makurap is known, PCT **wo(:)* can be securely traced back to PTpr **wo(:)* because precise external cognates are found across Tupian (Karitiana *hɨ* 'thorny tree', Sateré-Mawé *hu*, PTG **ju* 'thorn').

[44] Attested by Wanda Hanke only (Hanke *et al.* 1958: 212) as ‹kú› 'thorn, needle'.

[45] The absence of *β* is unclear. The word is frequently attested as *céβ-ao* ([ce^{H}.βao:?^{L}] in Moore's data; cf. also Braga 2005: 162), with the relationalizer prefix *ceP-* (Braga 2005: 42–3).

[46] Tup *áwro ~ áoro* can be securely traced back to PTpr **aworo* because precise external cognates are found across Tupian (Proto-Mundurukú **áro*, PTG **ajuru*).

[47] The development of PCT **wo* into Akuntsú *pɨ* is not known to be regular. Note that the unrounding of **o* is also attested in Siokweriat.

[48] The development of PCT **e* into Akuntsú *i* is not known to be regular. The word is attested as [wiˈtapˀ] ~ [wiˈttapˀ] ~ [wiˈtdapˀ] in Aragon (2008: 57), but is phonologized here with a /k^{w}/ because this is almost certainly the same word as the one found in the hydronym *K*^{w}*itaP kɨ* (Aragon 2014: 14), plausibly interpretable as 'anteater river'; the optional realization of /k^{w}/ as [w] is independently attested by Aragon (2014: 57sqq.).

[49] Although no cognate in Makurap is known, PCT **weP* can be securely traced back to PTpr **weP* because precise external cognates are found across Tupian (Karitiana *hap* 'to rise (of sun)', Awetí *teP* 'to go up'; Landin 2005: 10, Reiter 2011: 205).

[50] The Makurap term for 'peanut' is attested as *araβoɨK* (Braga 2005) or *aráβɨ:K* (Moore's data). Despite the obvious similarity to the Core Tuparian forms, there is no regular correspondence between them. It is unclear whether we are dealing here with an irregular development or with an indirect borrowing.

PTpr	gloss	Wayoró	Tuparí	Mekéns	Akuntsú	Makurap
**ewiT*[51]	honey, bee	*ẽŋgʷiT*	*ewíT*	*ekʷir-itsa* ‘bee sp.’ Sio *ekʷiT*	*ekʷíT*	—
**wiTʔi*	*açaí*	Ngw *ŋgʷiri* Kup *βiri*	*wíTʔi*	*kʷiri*	*kʷirí*	*βirí{ca}*
			nasal			
**wãmõʔã*	shaman	—	*wãmõʔã*	*kʷãmõã*	*kʷãmőã*	*mãmõã*
**wãkɨ̃ɲã*[52]	agouti	*ŋʷãkɯ̃ɲã*	—	Sak *mãkɨ̃ɲã* Gua/Sio *pakɨ̃ɲã*	*pakɨ̃ɲã*	*mãkɨ̃ɲã*
**wãẽ*[53]	pot	*ŋʷãẽ*	*wãẽ-tóP-ʔa*	Sak *ŋʷãẽ* Gua *kʷãẽ* Sio *kʷaʔẽ*	*kʷãʔẽ*	—
**wãCkɨT*	plate (WT only)	*ŋʷãCkɯT*	*wãCkɯ́t*	—	—	—
**wɨ̃-*	to enter	*ŋɯ̃-*	—	—	—	*mɨ̃-*
**wõ*	pet (CT only)	*ŋõ*	? *õ[ãkiT]*	*ŋõ*	—	—
**wĩʔĩK*	leafcutter ant	*ŋʷĩK*	*wĩʔĩK*	—	—	*mĩK*

Table 7. Proto-Tuparian */w/ (oral and nasal **w*)

3.2. Loss of syllabicity in high vowels

This subsection deals with the sound change whereby the high vowels */i o/ (and possibly their nasal equivalents) of PTpr became approximants when adjacent to vowels. Note that in the phonological systems of all Tuparian languages /o/ is analyzed as a high vowel, as there is no /u/. This type of sound change arguably recurred multiple times in the histories of the Tuparian languages, which is quite unsurprising given its naturalness. Its operation is most easily seen in the allomorphy patterns of the 3NCRF prefix (PTpr **i-*) and of the 1SG prefix (PTpr **o-*).

PTpr */i/ and */o/ were not affected by this process in the same fashion in the individual languages: while only Wayoró and Tuparí show traces of the desyllabification of the reflexes of PTpr */o/, the front high vowel */i/ has been affected in all daughter languages. In Makurap, the PTpr 3NCRF prefix */i-/ before vowels yielded /ɲ-/ (which surfaces as *ndʒ-* in oral environments and as *ɲ-* in nasal environments), as shown in 1. Phonetically, this development must have proceeded through the stage **j* (hence, PTpr **iV- > *jV- > */ɲV-/*) and evidently postdates the specifically Makurap sound change **j > c*.

(1) MAKURAP: /ɲ-/ *ndʒ-* oral, *ɲ-* nasal (Braga 2005: 50, 204; the glosses are ours)

a.	***ndʒ****-akáre-T*	b.	***ndʒ****-apiter-eT*
	/ɲ-akare-eT/		**/ɲ-**apiteT-eT/
	3NCRF-head-POSS		**3NCRF-**sadness-POSS
	‘his/her head’		‘his/her sadness’

[51] Although no cognate in Makurap is known, PCT **ewiT* can be securely traced back to PTpr **ewiT* because precise external cognates are found all across Tupian: Karitiana *e:t* (< **ahit*), Proto-Mundurukú **eit*, Sateré-Mawé *ewɨT*, Awetí *ekɨT*, PTG **eiT*.

[52] The reflexes of this word in the Corumbiara languages show labial consonants (Sak *m*, Gua/Sio/Aku *p*) instead of the expected labiovelars (Sak *ŋʷ*, Sua/Sio/Aku *kʷ*). That is, Proto-Corumbiara innovated by replacing **gʷãkɨ̃ɲã* with **bãkɨ̃ɲã*. Currently we have no explanation for this (apparently idiosyncratic) development.

[53] Although no cognate in Makurap is known (unless *ɨ̃ẽ* ‘pot’ is related), PCT **wãẽ* can be securely traced back to PTpr **wãẽ* because precise external cognates are found all across Tupian (e.g. Awetí *taʔẽ*, PTG **jaʔẽ*).

c. *etetenã* *kite* *pe* ***ɲ-***õ-ã *βɨT*
/etetenã kite pe **ɲ-**õP-a βɨT/
after_that people LOC **3NCRF-**give-TH all
'After that, he gave it to everyone.'

In the Core Tuparian languages, the PTpr 3NCRF prefix */i-/ was desyllabified before vowels as well, yielding PCT */j/. This must have happened **after** PTpr **j* became PCT **ð*, because the reflexes of PCT **ð* and **j* in oral environments are distinct in all Core Tuparian languages except Akuntsú (PCT **ð* > Way *nd*, Tup *h-/-Ø-*, Mek *t*, Aku *t*, whereas PCT **j* > Way *ndʒ*, Tup *s*, Mek *ts*, Aku *t*). In nasal environments, however, there is no distinction between the reflexes of PTpr */j/ and those of PTpr */i/ before vowels in PCT, as both merge in PCT **ɲ*.[54] That way, we believe that the inventory of approximants was augmented by one phoneme in PCT as compared to PTpr: first, PTpr **j* was dentalized to PCT **ð*, leaving room for PTpr **i* > PCT **j* (before vowels). In 2–5, we show the prevocalic allomorphs of the 3NCRF prefix in each Core Tuparian language.

(2) WAYORÓ: /ɲ-/ *ndʒ-* oral, *ɲ-* nasal (Nogueira 2019: 18)
ndʒ*-aʉ-βa,* ***ɲ****-ĩndiak^{w}a* *aʉ-βa,* ***ndʒ****-ʉβape* *aʉ-βa,*
/ɲ-aʉ-βa **ɲ-**ĩniak^{w}a aʉ-βa **ɲ-**ʉβape aʉ-βa
3NCRF-heal-VZR **3NCRF-**food heal-VZR **3NCRF-**beverage heal-VZR
ndʒ*-ato-a-P* *aʉ-βa*
ɲ-ato-a-P aʉ-βa/
3NCRF-bathe-TH-NMLZ heal-VZR
'He is healing it, healing her food, healing her drink, healing her bath.'

(3) TUPARÍ: *s-* oral, *ɲ-* nasal (Singerman 2018: 60–2)[55]
a. ***s****-opé* b. ***s****-aT* c. ***ɲ****-õpé* d. ***ɲ****-õpó*
3NCRF-thigh **3NCRF-**grab **3NCRF-**tongue **3NCRF-**kill
'his/her thigh' 'to grab him/her/it' 'his/her tongue' 'to kill him/her/it'

(4) MEKÉNS: *ts-* oral and nasal (Galucio 2001: 35–7, 191)
a. ***ts****-akoP* b. ***ts****-anɨ̃P* c. ***ts****-õ-k^{w}e-a-T*
/ts-akoP **ts-**anɨ̃P **ts-**(m)õ-k^{w}eP-a-T/
3NCRF-be_hot **3NCRF-**head **3NCRF**-CAUS**-**climb-TH-PST
'hot (it)' 'his/her/its head' 'he made him climb'

d. *ãrãmĩrã,* *aotse* ***ts****-õpo* *ka:T* *i-tser-a-T*
/ãrãmĩrã aotse **ts-**õpo ka:T i-tseT-a-T/
woman man **3NCRF-**beat and 3NCRF-leave-TH-PST
'The woman, the man beat her and she left.'

[54] One could claim that Mekéns has no merger: PTpr/PCT **ɲ* yields *ɲ* in Mekéns, whereas the allomorph of the 3NCRF prefix which occurs before nasal vowels is *ts-* (Galucio 2001: 35, fn. 6; 225, fn. 24) and not **ɲ*. We believe that in Mekéns the allomorph *s-*, originally restricted to stems which start with oral vowels, has been analogically extended to **all** vowel-initial stems. That way, Mekéns forms such as *ts-õpo* 'to beat him/her/it' (4d) are probably not cognate with Akuntsú *ɲ-õp-a* (5d) or Tuparí *ɲ-õpo* (3d), but rather arose through analogy.

[55] At different occasions, Singerman (2018) analyzes *ɲ-* as a realization of /i-/ (p. 60–2) or of /j-/ (p. 371). For our current purposes, the choice between these two analytical options is irrelevant. Also note that in Tuparí the allomorph *i-*, which was historically restricted to consonant-initial stems, may synchronically occur before vowels, as in *i-eT* 'his/her name' (Singerman 2018: 56), as a result of the elision of PCT **ð* (**i-ðeT*).

(5) AKUNTSÚ: *t-* oral, *ɲ-* nasal (Aragon 2014: 46, 138, 177, 279)[56]

a. ***t**-akoP* *te*
3NCRF-be_hot FOC
'It is hot.'

b. *tatʃo, tatʃe tɨrɨ **t**-ajtʃi*
Tatʃo Tatʃe two **3NCRF-**wife
'Tatʃo and Tatʃe were his two wives.'

c. ***t**-anãP* *etʃe* *kaP*
3NCRF-head DIFF wasp
'A wasp is on his head.'

d. ***ɲ**-õp-a*
3NCRF-beat-TH
'to beat him'

In some Core Tuparian languages, the innovative PCT **j* has merged with segments whose ultimate source is different from **i*. In subsection **3.4**, we will show that Proto-Corumbiara innovated by creating transitional glides (as in PCT **piʔa* > Proto-Corumbiara **pija*), which have the same reflexes as PCT **j* < PTpr **i* (that is, Mekéns *ts*, Akuntsú *t*). Similarly, the word for 'spider monkey' can be reconstructed as Proto-Corumbiara **jakɨraP* (> Mekéns *tsakɨraP*, Akuntsú *takɨraP*).

Now we turn to the desyllabification of PTpr **o*. This process is synchronically attested in Tuparí, in which the 1SG prefix occurs as *o-* before consonants, but as *w-* before vowels (Singerman 2018: 42); it also coalesces with a following /o/ or /õ/, yielding a long vowel. In Wayoró, the 1SG prefix also occurs as *o-* before consonants; before unrounded vowels, however, one finds the allomorph /m-/ (*mb-* in oral environments, *m-* in nasal environments), whereas before rounded vowels the zero allomorph occurs (Nogueira 2019: 11, 15, 150–1). This is shown in Table 8.

	Wayoró (Nogueira 2019)		Tuparí (Singerman 2018)	
before a consonant	*o-* /o-/	*o-ʔʉβa* 'my pot' *o-piti:K* 'I feel cold' *o-ŋgora* 'to seek me'	*o-*	*o-si* 'my mother' *o-kẽPk-a:* 'I nursed' *o-karãP* 'toward me'
before a rounded vowel	*Ø-* /Ø-/	*Ø-ʉpipe* 'my port' *Ø-õmb-a:* 'hit me!'	*o-o… → o:-*	*o:P* (*o*+*oP*) 'my father' *õ:jaora* (*o*+*õjaoT*+*a*) 'to answer me'
before an oral unrounded vowel	*mb-* /m-/	*mb-apiteP* 'my ear' *mb-ato-a-P* 'my bath' *mb-e-tʃʉ:P-k^{w}a-T* 'I got wet'	*w-*	*w-apaPʔa* 'my head' *w-e-kiaraP-k-a* 'I became happy' *w-e-pak-a* 'I woke up'
before a nasal unrounded vowel	*m-* /m-/	*m-ẽŋgʉ* /m-ẽŋʉ/ 'my chicha' *m-ãmõC-k^{w}a-T* 'I dance fast'		*w-ẽkẽT~ʔẽkẽT-k-aP* 'my throwing up'

Table 8. The allomorphy of the 1SG prefix in Wayoró and Tuparí

While the allomorphy pattern attested in Tuparí can be explained away as a consequence of a recent natural sound change (**o* > *w* _V$_{[-rounded]}$), the pattern found in Wayoró requires a more elaborate diachronic account: positing a one-step sound change such as */o/ > /m/ would be an entirely implausible solution on typological grounds. Fortunately, there is independent comparative evidence which shows that at some point in the history of Wayoró all inherited word-initial (and some word-medial) approximants have become homorganic (underlying) nasals. We have already seen in **3.1** that PTpr **j* and **w* (> PCT **ð*, **ɯ*/*[ɣ]) are reflected in Wayoró as *nd*, *ŋg^{w}*/*ŋ^{w}*, and *ŋg*/*ŋ* (underlying /n/, /ŋ^{w}/, and /ŋ/); earlier in this subsection, it has been shown that the innovative PCT **j* has yielded Wayoró *ndʒ* /ɲ/. That way, it appears quite

[56] Aragon (2014: 46, fn. 28) analyzes *ɲ-* as an allophone of /i-/ but is explicit regarding its phonetic realization. For example, the example 5d is transcribed as [ʃũ.ˈba] ~ [ɲõ.ˈba] in Aragon (2014: 46).

plausible that the allomorph /m-/ 1SG in Wayoró also continues an earlier approximant, which we reconstruct as pre-Wayoró *β- and derive from PTpr/PCT *o- in the prevocalic position. Therefore, forms such as *β-*apiteP* 'my ear' and *β-*ato-a-P* 'my bath' are posited for the pre-Wayoró stage. Later on, *β- would have undergone nasalization word-initially (in the intervocalic position, as we have seen in **3.1**, it was preserved, as in PCT *$e\beta a$-*opaP* > Wayoró $e\beta a$-*paP* 'eye').

Note that the desyllabification of *o- before vowels **cannot** be considered an innovation shared by Wayoró and Tuparí, even though it occurred in both languages in comparable contexts. First of all, the outcome of this process is different in pre-Wayoró (*β-) and Tuparí (w-). The second piece of evidence for positing two independent innovations is that the allomorph w- in Tuparí occurs not only in originally vowel-initial stems, but also in stems which have diachronically lost their initial consonant (PCT *$\eth$, preserved as *nd* in Wayoró). For example, the consonantal allomorph occurs in Tuparí *w-eK* 'my house' (Singerman 2018: 43), which goes back to PCT *o-$\eth$*eK* (apparently by the way of pre-Tuparí **o-eK*). Its Wayoró cognate *o-ndeK* 'my house' (Nogueira 2019: 145, 165), which has not been affected by any process of consonantal loss, expectedly shows the vocalic allomorph *o*-.

The allomorphy patterns examined in this subsection are decisive in establishing the directionality of the sound changes which underlie the correspondence sets involving Wayoró nasal stops and non-nasal segments in other Tuparian languages. If one were to derive them from something other than approximants, it would be quite difficult to explain why Wayoró has /m-/ and /ɲ-/ as the prevocalic allomorphs of /o-/ and /i-/, respectively. In our account, this is unproblematically attributed to a combination of two processes: the desyllabification of high vowels in the environment #_V (*i > *j in PCT, *o > *β in pre-Wayoró) and the nasalization of approximants in Wayoró (*β-, *$\eth$, *j-, *w, *[ɣ] > /m-/, /n/, /ɲ-/, /ŋʷ/, /ŋ/).

3.3. Proto-Tuparian postoralized nasals and their development in Wayoró

It is possible to reconstruct three phonemic nasals for Proto-Tuparian: */m/, */n/, and */ŋ/. In oral environments, they likely acquired an oral phase before an oral nucleus and thus surfaced as *mb, *nd, *$ŋg$ (Galucio & Nogueira 2012). In nasal environments, */m/ and */n/ appear to have surfaced as *m and *n, which have been preserved as such in all contemporary Tuparian languages (in contrast, */ŋ/ was likely postoralized even before nasal nuclei). Wetzels & Nevins (2018) classify the allophonic pattern of this type, which is known from many Amazonian languages, as nasal shielding. We call the allophones *mb, *nd, *$ŋg$ **postoralized** in what follows. For our current purposes, it is essential that the Wayoró reflexes of the postoralized allophones are identical or similar to those of the Proto-Tuparian approximants in oral environments. More specifically,

- PTpr *β and *mb merge in Wayoró as *mb*- (word-initially) or -β- (between vowels);
- PTpr *j (> PCT *$\eth$) and *nd merge in Wayoró as *nd*;
- PTpr *w before rounded vowels (> PCT *[ɣ]) and *$ŋg$ merge in Wayoró as $ŋg$.

Based on the contemporary Wayoró reflexes, one may be tempted to attribute these mergers to a single sound change from the Proto-(Core) Tuparian approximants to Wayoró underlying nasals. Indeed, in **3.1–2** we have seen that most approximants of Proto-Core Tuparian became homorganic nasals in Wayoró: PTpr *j > PCT *$\eth$ > Wayoró *nd*; PTpr *w > PCT *w/*[ɣ] > Wayoró $ŋ(g)^w$/$ŋ(g)$; PTpr *o-, *i- before vowels > PTpr *β-, *j- > Wayoró *m(b)-*, *ndʒ*-. In contrast, the Wayoró reflexes of PTpr *mb, *nd, and *$ŋg$ are identical to their reconstructed states, as in PTpr *mbo 'hand', *ndeT- 'to grind', *ŋgaP 'wasp' > Wayoró *mbo*, *ndeT-*, *ŋgaP*. At first glance, these sounds would appear to have been preserved intact in Wayoró all the way from Proto-Tuparian.

In this paper, however, we advance an alternative proposal. Namely, we hypothesize that the postoralized allophones of PTpr nasals (i.e., **mb*, **nd*, **ŋg*) have been affected by a series of sound changes in Wayoró, which came full cycle to the initial state. The suggested evolution pathway of PTpr **mb*, **nd*, **ŋg* in the Core Tuparian languages is as follows: (i) in PCT, they lose the nasal phrase and become **b*, **d*, **g*; (ii) in the Corumbiara languages and in Tuparí, they merge with PCT **p*, **t*, **k* and yield Mek/Aku *p*, *ts*/*tʃ*, *k*, Tup *p* (*s-*/*-ps-* before *i*), *t* (*s* before *ɨ*/*i*), *k*; (iii) in pre-Wayoró, they lenite to **β*, **ð*, **ɣ* (and merge, therefore, with PCT **β*, **ð*, **[ɣ] from PTpr **β*, **j*, **w*; the bilabial approximant in pre-Wayoró may also come from **o-* as seen in **3.2**); (iv) in contemporary Wayoró, they have been affected by the independently established nasalization process (**β-*, **ð*, **ɣ* > /m/, /n/, /ŋ/). That way, the development from PTpr **mb*, **nd*, and **ŋg* to Wayoró *mb*, *nd*, and *ŋg* is assumed to have proceeded in three steps (**mb*/**nd*/**ŋg* > **b*/**d*/**g* > **β*/**ð*/**ɣ* > *mb*/*nd*/*ŋg*), as opposed to a straightforward retention. It also entails that PTpr postoralized nasals and approximants first merged as pre-Wayoró approximants (and not as modern Wayoró underlying nasals).

Crucial evidence for our proposal comes from the development of PTpr **mb* after an oral vowel in Wayoró: in this position, it is reflected as *β*. Nogueira (2011: 45–6) documents forms such as *o-βo* 'my hand' and *o-βi* 'my foot' (< PTpr **o-mbo*, **o-mbi*), of which the uninflected forms are *mbo* and *mbi*, respectively (Moore & Galucio 1994: 133). Note that the environment which conditions the development of PTpr **mb* in Wayoró is precisely the same that the one we have seen above for pre-Wayoró **β* from other sources (PTpr **β* or **o*): it is reflected as /m/ word-initially (as in PCT **o-apiteP* > **β-apiteP* > *mb-apiteP* 'my ear'), but is retained as /β/ after an oral vowel (as in PCT **eβa-opaP* > *eβa-paP* 'eye'). It is, therefore, conceivable that PTpr **mb* (as in **mbo* 'hand' and **mbi* 'foot') merged with other segments as pre-Wayoró **β* (as in **βo* 'hand', **βi* 'foot', **o-βo* 'my hand', **o-βi* 'my foot'), which was subsequently reverted to /m/ word-initially (and after a nasal vowel) by means of an independently reconstructed process (see **3.2**), as in *mbo* 'hand' and *mbi* 'foot', but suffered no further changes after an oral vowel, as in *o-βo* 'my hand' and *o-βi* 'my foot'. For PCT, we reconstruct **b* based on the fact that neither pre-Wayoró nor Tuparí or Corumbiara show any traces of a nasal phase.

We find it likely that PTpr **nd* and **ŋg* have undergone in Wayoró a cycle of sound changes comparable to the one described for PTpr **mb* in the preceding paragraph. For PCT, we reconstruct **d* and **g*: in Tuparí and in both Corumbiara languages they merge with PCT **t* and **k* (thus paralleling the merger of PCT **b* and **p* in these languages), whereas in Wayoró they merge with PCT **ð* and **[ɣ] as pre-Wayoró **ð*, **ɣ* > Wayoró /n/, /ŋ/ (thus paralleling the merger of PCT **b* and **β*) in Wayoró. That way, PTpr **ndeT-* 'to grind', **ŋgaP* 'wasp' are hypothesized to have developed into PCT **deT-*, **gaP* > pre-Wayoró **ðeT-*, **ɣaP* > Way *ndeT-*, *ŋgaP*.

PTpr */m/. In nasal environments, PTpr */m/ surfaced as **m* and was preserved as such in all daughter languages. In oral environments, it likely had the allophone **mb*, which was preserved in Makurap but suffered some changes in the Core Tuparian languages. As stated above, we believe it yielded PCT **b*. In Tuparí and in the Corumbiara languages, it merged with the reflexes of PTpr/PCT **p* as Tuparí *p* (assibilated to *s-*/*-ps-* before *i*) and Mekéns/Akuntsú *p*. In pre-Wayoró, PCT **b* merged with the reflexes of PTpr/PCT **β* and PTpr/PCT **o* (before vowels) as pre-Wayoró **β*, which yielded /m/ word-initially or after a nasal vowel and /β/ after an oral vowel. The Tuparian etymologies which instantiate PTpr */m/ are listed in Table 9.

PTpr */n/. In nasal environments, PTpr */n/ surfaced as **n* and was preserved as such in all daughter languages. In oral environments, it likely had the allophone **nd*, which was preserved in Makurap but suffered some changes in the Core Tuparian languages. As stated

PTpr	gloss	Wayoró	Tuparí	Mekéns	Akuntsú	Makurap
			oral			
**mbo*	hand	*mbo / -βo*	*po*	*po(-pi)*	*po*	*mbo*
**(ɲ)õmbo-*	to kill, to beat (CT only)	*õmbo-*	*õpó-*	*õpo-*	*õp-á*	—
**mboejoP- ~ *mboec/to-*[57]	to know	*mõẽndoP-*	*pɯ́(ʔ)oP-*	*poetoP-*	*poetóP-*	*mboeto-* ~ *mbɨeto-*
**mboKʔa*	tortoise (CT only)	*mboga*	*póKʔa*	*poga*	*pogá*	—
**mbi(-to)*	foot	*mbi / -βi*	*sitó*	*pitso*	*pi*	*mbi*
**ja:mbi*	crop seed	*aβi*	*a:Psí*	*a:pi*	—	*cãmbi*
**mbiro*	to have (CT only)	*mbiro*	*-(P)síro*	*piro*	—	—
**mbiriʔa*[58]	trahira fish	*mbirija*	*siriʔá*	Sio *piritsa*	*biritá*	*mbíria*
**mbiʔo*	horsefly	—	*siʔó*	—	—	*mbio*
**mbiP*	to be afraid	—	—	—	*piP*	*mbiP*
			nasal			
**mã-*	to put	*mã-*	*mã-*	*mã-*	*mã-*	*mã-*
**mãɲĩ*	manioc	*mãɲĩ*	*mãc*	—	—	*mãɲĩ*
**ãmãnã*	tayra	*ãmãnã*	—	—	—	*ãmãnã*
**mẽT*[59]	husband	*mẽT*	*mẽT*	*mẽT*	*mẽT*	—
**mẽpiT*	child, sororal nephew/niece (female ego)	*mẽpiT*	*mẽPsíT*	*mẽpiT*	*mẽpíT*	*mẽpiT*
**mẽpir-ẽpiT*	grandchild (female ego, CT only)	*mẽpir-ẽpiT*	*mẽPsĩr-ẽPsĩT*	*mẽpir-ẽpiT*	*mẽpir-ẽpíT*	—
**mẽɲõP*	son-in-law (daughter's husband)	*mẽɲõP*	*mẽɲṍP*	—	—	*mẽɲõP*
**amẽko*	jaguar, dog	*amẽko*	*amẽkó*	*amẽko*	*amẽkó*	*ãmẽko*
**mĩcõ*[60]	curassow	—	—	—	—	*mĩtṍ*

Table 9. Proto-Tuparian */m/ (oral **mb*, nasal **m*)

[57] The Core Tuparian languages unequivocally point to PCT **boeðoP-* (in Wayoró, **ð* > *nd* nasalized the preceding vowels; in Tuparí, **poeoP-* was apparently simplified to *pɯ́(ʔ)oP-*). However, the expected Makurap correspondence would be **mboecoP-* and not *mboeto- ~ mbɨeto-*.

[58] The Wayoró reflex is from Galucio *et al.* (2015: 274), where it is given as *mbirija, mbiriʤa* (with a transitional *j* /ɲ/). The expected reflex in Akuntsú would contain a /p/ and not a /b/; in fact, Galucio *et al.* (2015: 274) do give Akuntsú *pirita* (in our transcription, *pirita*), but our primary source has *biritá* (Aragon 2014: 109).

[59] Although no cognate in Makurap is known, PCT **mẽT* can be securely traced back to PTpr **mẽT* because precise external cognates are found all across Tupian (Karitiana *mãn*, Karo *mẽn*, PTG **mẽT*; Landin 2005: 16, Gabas Jr. 1999: 13, Mello 2000: 178).

[60] Although no cognate in the Core Tuparian languages is known, Makurap *mĩtṍ* can be securely traced back to PTpr **mĩcõ* because precise external cognates are found all across Tupian (Karitiana *mbisɨ̃*, Proto-Mundurukú **wítõ*, PTG **mɨ̃tũ*; Landin 2005: 16, Picanço 2019: 140, Mello 2000: 182).

PTpr	gloss	Wayoró	Tuparí	Mekéns	Akuntsú	Makurap
			oral			
**ndɨʔa*	lip color	—	*sɯ́ʔa*	—	—	*ndɨa*
**(ɲ)ẽndɨ*	mortar (CT only)	*endɯ{dʒa}* ‘pestle’	*ẽTsɯ́-ʔa*	*ẽtsɨ*[61]	—	—
**ndɨrɨ*	collared anteater	—	*sɯ́rɯ~sɯ́rɯ*	—	—	*ndɨrɨ*
**ndo:*	mound, hill	*ndo:*	*to-téT*	*tso(:)*	—	*ndó-a*
**ndeT-*	to grind	*ndeT-*	*teT-*	—	—	*ndeT-*
**(j)aindi*[62]	wife	*aindi*	—	*aitsi*	*ajtʃí*	—
			nasal			
**ãmãnã*	tayra	*ãmãnã*	—	—	—	*ãmãnã*
**nãko*	man	—	—	*nãko*[63]	*nãkó*	*nãko-βɨT* ‘boy’
**nõ*	other	*nõ*	*nõ*	*nõ*	*nõ*	*nõ-T*
**ãnõrẽ*[64]	barred *sorubim* fish	*ãnõrẽ*	*anő́re*	—	—	*ãnő́re*
**nẽ*[65]	arm	*nẽ-* (in compounds)	*{a}nẽ-tó* ‘shoulder’	*nẽ* ‘shoulder’	—	*nẽ*
**nẽ-*	to make	*nẽ-*	*nẽ-*	—	—	*nẽ-*
**nẽcɨK*	horsefly	—	*nẽɯK*	—	—	*nẽtɨK*

Table 10. Proto-Tuparian */n/ (oral **nd*, nasal **n*)

above, we believe it yielded PCT **d*. In Tuparí and in the Corumbiara languages, it merged with the reflexes of PTpr/PCT **t* as Tuparí *t* (assibilated to *s* before *i*/*ɯ*), Mekéns *ts*, and Akuntsú *tʃ*. In pre-Wayoró, PCT **d* became **ð* (merging with the reflexes of PTpr **j* > PCT **ð*), which yielded *nd* /n/ in modern Wayoró. The Tuparian etymologies which instantiate PTpr */n/ are listed in Table 10.

PTpr */ŋ/. PTpr */ŋ/ quite probably surfaced as **ŋg* before nasal and oral vowels alike. This contrasts with the pattern we reconstruct for */m/ and */n/, whereby the postoralized realization is found in oral environments only. One piece of evidence comes from Tuparí,

[61] Attested as ‹enzê› in Hanke *et al.* (1958: 204) and as *ẽsɨ* in Moore & Galucio (1994: 134).

[62] PTpr also likely had the compound **jaʔɨP-ti* ‘wife’ (literally, ‘son’s mother’) preserved as Wayoró *ndaɨP-ti* and Mekéns *taɨP-si*. In Tuparí and Makurap, **(j)aindi* and **jaʔɨP-ti* appear to have contaminated: the former language has *aʔɯ́si* ‘wife’ (instead of the expected reflexes **aisi* or **haʔɯP-si*); the latter has *caɨp-ndi* (instead of the expected reflexes **(c)aindi* or **caɨP-ti*). The irregularities in the correspondences have been noted by Nogueira *et al.* (2019: 46), where the reconstruction **ai(+)tsi* is given. Although PCT **aindi* has no exact cognate in Makurap, it can be securely projected to PTpr **(j)aindi* because it influenced the shape of *caɨP-ndi* and because cognates are also found elsewhere in Tupian (Proto-Mundurukú **tajtʃi*; Picanço 2019: 138).

[63] Attested in Hanke (1958: 206) as ‹nankú› ‘man’. Galucio *et al.* (2015: 251) give *nakop* ‘man’ instead.

[64] Alves (2004: 145) claims that the Tuparí form is borrowed from Makurap, but there would appear to be no formal reason to believe so.

[65] The Wayoró and Mekéns forms are attested in Snethlage (2015: 518, 686) as Wayoró ‹onäntǫ› ‘Schulter’, ‹unämí á› ‘Ellbogen’ (likely *o-nẽ-to*, *o-nẽ-mĩã*, with the 1SG prefix *o-*) and Mekéns ‹kiná, kinã› ‘Schulter’ (likely *ki-nẽ*, with the 1INCL prefix *ki-*). The Mekéns form is also attested as ‹unᵉᵃ› ‘bras’ (likely 1SG *o-nẽ*) by Claude Lévi-Strauss in his *Kabišiana* wordlist (*apud* Loukotka 1963: 48).

Mekéns and Akuntsú, which reflect PTpr */ŋ/ as *k* regardless of whether the nucleus of the syllable is oral or nasal (unlike what we saw above for PTpr */m/ and */n/, which show a conditioned split in these three languages). Based on the correspondence between Wayoró /ŋ/, Tuparí /k/ and Mekéns/Akuntsú /k/, we may safely reconstruct PCT **g*, which therefore differs from PCT **b* and **d* in occurring in oral and nasal environments indiscriminately. The second piece of evidence for reconstructing PTpr **ŋg* as the only realization of PTpr */ŋ/ comes from Braga's (2005) transcriptions of Makurap words, in which /ŋ/ is transcribed as [ⁿg] even in nasal environments: [ⁿgẽm] 'breast', [ⁿgẽˈɾẽj] 'to shut up, to be silent', [ⁿgɜ̃ˈĩ] ~ [ⁿgə̃ˈĩ] 'knife' (Braga 2005: 195–6).[66] That way, PTpr **ŋg* (the only allophone of PTpr */ŋ/) would have been preserved in Makurap. In PCT, it would have yielded **g*, which was further devoiced to *k* in Tuparí and in the Corumbiara languages (and merged with PCT **k*, paralleling the merger of PCT **p*/**b*, **t*/**d* in these languages). In Wayoró, **g* was probably lenited to pre-Wayoró **ɣ* (by means of the process which also lenited PCT **b*/**d* to pre-Wayoró **β*/**ð*) and later nasalized to modern Wayoró /ŋ/. Unlike in PTpr (in our reconstruction), however, the fully nasal realization of Wayoró /ŋ/ in nasal environments is compulsory (Nogueira 2011: 50–1). The Tuparian etymologies which instantiate PTpr */ŋ/ are shown in Table 11.

3.4. Innovative approximants in Proto-Corumbiara

There is good reason to think that the proto-language of the Corumbiara branch acquired innovative approximants via hiatus resolution, whereby glides were inserted in the environment **i,o,ɨ_V*. This includes both original hiatuses, retained from Proto-Tuparian, and new hiatuses, which arose as a result of elision of a PTpr glottal stop. Although the epenthesized segments are not phonetically approximants in the contemporary languages —rather, consonants such as *t*, *s*, or k^w are found— we believe that these go back to erstwhile transitional glides, **j* (inserted in the environment **i_V*) and **w* (**o,ɨ_V*), which were subsequently fortitioned. That way, the epenthesis in pre-Proto-Corumbiara could be characterized as a natural sound change, *glide epenthesis* (Blevins 2008: 84sqq.), which fed another natural sound change, *approximant fortition* (independently established in subsections **3.1–2** above). The development pathway advanced in this subsection is thus essentially identical to Blust's (1994: 112–5) account of certain sound changes in a number of Austronesian languages (such as Chamorro), in which not only inherited approximants but also transitional/epenthetic glides have been historically fortitioned, as in Chamorro *pugwaʔ* 'betel nut', *gwidza* '3SG' (from **buaq*, **ia*; cf. also Blevins 2008: 92).

As was already mentioned above, we suggest that the consonant originally epenthesized in the environment **i_V* in pre-Proto-Corumbiara was **j*. In **3.2**, we saw that PCT **j-* (from PTpr **i-* before vowels) yielded Mekéns *ts*, Akuntsú *t*. In the environment **i_V*, however, a slightly more complicated situation is found: at least in some words, the Guaratira and the Siokweriat dialects of Mekéns have *∅* corresponding to *ts* in the Sakurabiat dialect (and to *t* in Akuntsú). For example, Galucio *et al.* (2017: 338, fn. 6) report that *pia* 'to wait' is the form used in the Guaratira and Siokweriat dialects, which corresponds to *pitsa* in the Sakurabiat dialect. The Tuparian etymologies which instantiate Proto-Corumbiara **j* are shown in Table 12. The Proto-Corumbiara reconstructions themselves are not given in the table for reasons of space; we reconstruct **kije* 'one', **pija* 'liver', **ãpija* 'nose', **opija* 'egg', **pirija* 'trahira fish', **g^warija* 'bat', **(a)mĩjã* 'knee' (in the last word, Mekéns and Akuntsú have *ɲ*/*n* rather than *ts*/*t*, quite possibly due to the nasal environment).

[66] We assume that these transcriptions supersede Braga's earlier claim, according to which /ŋ/ is obligatorily postoralized before oral vowels only (Braga 1992: 45–7).

PTpr	gloss	Wayoró	Tuparí	Mekéns	Akuntsú	Makurap
			oral			
*ŋgaP	wasp	ŋgaP	kaP	kaP	kaP	ŋgaP
*ŋgapi(-ʔa)[67]	bullet ant	—	—	—	—	ŋgápia
*ŋgɨ̰[68]	liquid, saliva	ŋgʉ̰	kʉ̰	kɨ̰	kɨ̰	—
*ŋgɨ̰P	louse	ã-ŋgʉ̰P	kʉ̰P	kɨ̰P	kɨ̰P	ŋgɨ̰P
*ŋgɨ̰ʔɨ̰T	salt	ŋgʉ̰:T[69]	kʉ̰ʔʉ́T	kɨ̰:T	? kɨ̰C[70]	ŋgɨ̰T
*ŋgoPʔiT[71]	termite	ŋgʉ̰i	kóPʔi	kobi	kopí	? ŋgóβ-a
*ŋgɨ̰piʔa[72]	tick (CT only)	ŋgʉ̰piʔa	—	—	kɨ̰píta	—
*ŋgoT	palm larva	—	koT	—	—	ŋgoT
*ŋge	garden	ŋge	—	—	—	ŋge
*ŋgeaT	sun, sky	ŋgiaT ‘sky’	kiáT ‘up’	—	—	ŋgéaT
*ŋgeK[73]	caterpillar	—	—	—	—	ŋgeK
*ŋgi-akoP	sun (CT only)	ŋgi-akoP	ki-akóP	ki-akoP	ki-akóP	—
			nasal			
*ɲãŋgã	branch	—	ãkã̰	ãkã̰[74]	—	ɲãŋgã
*kʉ̃nĩŋga[75]	scorpion	kʉ(ʉ)nĩŋã	kwĩnĩká	kɨ̃nĩŋã	—	—
*ŋgẽP	breast	ŋẽP	kẽP	kẽP	kẽP	ŋgẽP
*ŋgẽT- *ŋgẽT-ŋga-	to sink to swallow (WT only)	ŋẽT- ŋẽT-ŋga-	kḛ̃T-ka	—	—	—

Table 11. Proto-Tuparian */ŋ/ (oral and nasal *ŋg)

[67] Although no cognate in Core Tuparian is known, Mak *ŋgápia* can be securely traced back to PTpr *ŋgapi(-ʔa)* because an external cognate is found in Karitiana (*nõpi* ‘bullet ant’, Landin 2005: 19). Karitiana *n* is a regular reflex of Proto-Tupian *ŋ in nasal environments, but the mismatch between the nasality values of the first syllable in Makurap and Karitiana awaits further explanation.

[68] Although no cognate in Makurap is known, PCT *ŋgɨ̰ can be securely traced back to PTpr *ŋgɨ̰ because an external cognate is found in Karitiana (*ŋge* ‘blood’, Landin 2005: 9).

[69] Form attested in Moore & Galucio (1994: 134). Nogueira (2011: 40) documents a form with an initial *k-*, which could be a mistranscription or a borrowing from another Tuparian language.

[70] It is unclear if this is an irregular reflex of *ŋgɨ̰ʔɨ̰T or a semantic extension of kɨ̰C ‘earth’ (< PTpr *kɨ̰C).

[71] The Wayoró and Mekéns forms are cited after Galucio *et al.* (2015: 272); the expected Wayoró reflex would actually be *ŋgoβi*. It is uncertain if the Makurap word is a precise cognate because of the vowel mismatch; it is possible that continues a derivative close to *ŋgoPʔi-ʔa* (compare Proto-Mundurukú *kópiɾ̃ ‘blood’, Sateré-Mawé *nupiʔa*; Picanço 2019: 138, Ribeiro 2010: 76).

[72] The Wayoró form is from Galucio *et al.* (2015: 273).

[73] Although no cognate in Core Tuparian is known, Mak *ŋgeK* can be securely traced back to PTpr *ŋgeK* because an external cognate is found in Karitiana (*ŋgak* ‘caterpillar’, Landin 2005: 9).

[74] The form is tentatively phonologized based on Snethlage’s (2015: 520) attestation of ‹zānka› ‘Zweig’ (likely the third person *ts-ãkã*).

[75] The Wayoró and Mekéns are given as *kɨ̰nĩjãʔ* and *kɨ̃niŋã* in Galucio *et al.* (2015: 270) and as *kɨ̰nĩŋã* and *kɨ̰tnĩŋã* in Moore & Galucio (1994: 134); our phonologization is tentative. In Tuparí, the vowel of the first syllable is irregularly diphthongized (*ʉ̃ > *wĩ*), assuming our primary source (Alves 2004: 204) records the word correctly; in Mekéns, we would expect a /k/ rather than a /ŋ/. The stem is likely inherited from Proto-Tuparian (as opposed to a Core Tuparian innovation), because there is a probable cognate in Karitiana: *kennõn* ‘scorpion’ (Landin 2005: 13) < pre-Karitiana *kɨ̰nVŋãT.

PTpr	gloss	Wayoró	Tuparí	Mekéns	Akuntsú	Makurap
**kie*	one (CT only)	*kie-T*	*kíe*	*kitse, kie{ka}* Sio *gɨe*[76]	*kite*	—
**piʔa*	liver	*pia*	*siʔá*	*pitsa*	*bíta*	*pía*
**ɲãpi(ʔa)*	nose	*ãpia*	*ãPsí*	*ãpitsa*	*ãpíta*	*ɲãpi*
**jopiʔa*	egg	*ʉpia*	*oPsíʔa*	*opitsa*	*õpíta*	*cópia*
**mbiriʔa*[77]	trahira fish	*mbirija*	*siriʔá*	Sio *piritsa*	*biritá*	*mbíria*
**wariʔa*	bat (CT only)	*ŋg^{w}aria*	*wáriʔa*	*k^{w}aritsa* Sio *k^{w}aritsa ~ g^{w}-*	—	*(βacáriaC)*
**-pĩã*	knee	*{kʉ̃}mĩã*	*mĩã{K-ʔã}*	*{nẽkiwa}mĩɲa* ‘elbow’	*{a}mĩnã*	*{ka}pĩã*

Table 12. Proto-Corumbiara **j* (> Sakurabiat *ts*, Guaratira/Siokweriat *∅*, Akuntsú *t*)

It is interesting to observe that in the variety of Mekéns attested by Wanda Hanke the consonant in question is consistently transcribed as ‹z› (more rarely, ‹s›), which implies that she must have made her notes with speakers of the Sakurabiat dialect: ‹kitzê› ‘one’, ‹-piza› ‘liver’, ‹ampiza› ‘nose’, ‹kurakura-rupiza› ‘egg’ (*korakora* ‘hen’), ‹kwarisa, kwarisasu› ‘bat’ (Hanke *et al.* 1958). In contrast, Emil-Heinrich Snethlage’s notes on Mekéns typically have zero: ‹-piá› ‘liver’, ‹-ãmpiá› ‘nose’, ‹-upiá› ‘egg’; one exception is ‹kízäkạ̊tn› ‘one’ (Snethlage 2015: 518–20).

Curiously, not all instances of Sakurabiat *ts* ~ Guaratira/Siokweriat *∅* reflect an erstwhile epenthetic glide. At least in two words, this correspondence continues PTpr/PCT **t* > Proto-Corumbiara **ts*. One example is the word for ‘foot’, which has the shape *pitso* in Sakurabiat, but *pio* in Guaratira and Siokweriat (Galucio 2001: 19); the Tuparí cognate *sito* ‘foot’ shows that all these words continue PCT **bito* (ultimately derived from PCT **bi* < PTpr **mbi* ‘foot’). Similarly, Galucio (2011b:7, fn. 11) states that the 1INCL pronoun is *kitse* in Sakurabiat but *okie* in Guaratira and Siokweriat (< PTpr **kite*).

Now we turn to the environment **o,ɨ_V*. We believe that such hiatuses were resolved in pre-Proto-Corumbiara by means of the insertion of **w* (merging with PTpr/PCT **w*), which yielded Mekéns/Akuntsú *k^{w}*. Curiously enough, the Wayoró cognates of these words also have *k^{w}*, which must be attributed to an independent innovation. The Tuparian etymologies which instantiate pre-Proto-Corumbiara **w* are shown in Table 13.

PTpr	gloss	Wayoró	Tuparí	Mekéns	Akuntsú	Makurap
**jo(ʔ)a*	brother (female ego)	*ok^{w}a*	*owá ~ oʔá*	*ok^{w}a*	*ok^{w}á*	*cóa*
**joaC*	tail	*ok^{w}aC*	*oáC*	*ok^{w}aC*	*ok^{w}áC*	*ɲẽT-coaC*
**oʔe*[78]	to wash	*ok^{w}-á*	*oʔé*	*ok^{w}-a*	*õk^{w}-a*	*ó-a*
**ɨa:C*	tapir	*ʉk^{w}a:C*	—	*ɨk^{w}a:C*	*ɨk^{w}áC*	*ɨáC*

Table 13. Pre-Proto-Corumbiara **w* > Mekéns/Akuntsú *k^{w}*

[76] The forms *kitse* and *kieka* are cited after Galucio *et al.* (2015: 250); we would expect *kitse* to be from the Sakurabiat dialect and *kieka* from the Guaratira dialect. Siokweriat [gɨe] is cited after Áragon (2014: 310); we would expect it to be *kie*.

[77] The Wayoró reflex is from Galucio *et al.* (2015: 274), where it is given as *mbirija, mbiriʤa* (with a transitional *j* /ɲ/). The expected reflex in Akuntsú would contain a /p/ and not a /b/; in fact, Galucio *et al.* (2015: 274) do give Akuntsú *pirita* (in our transcription, *pirita*), but our primary source has *biritá* (Aragon 2014: 109).

[78] The stem without the thematic vowel is attested only in Tuparí. Braga (2005: 196) gives the stem without the thematic vowel as *oP*, but it is not attested in any linguistic example and does not match the Tuparí form.

Even though only one example is known for the environment *ɨ_V, the *w-epenthesis must have been regular. No reflexes of PTpr *ɨaT 'grass, lawn' (> Tuparí ʉáT 'lawn', Makurap ɨaT 'grass') are attested in Wayoró or in the Corumbiara languages (we would expect them to have the shape *ʉk^{w}aT in Wayoró and *ɨk^{w}aT in Mekéns/Akuntsú). The correspondence between Wayoró ʉβeko:P, Akuntsú ɨekó, Mekéns ɨeko, and Makurap ɨéko, all meaning 'king vulture', appears to be, therefore, irregular (especially regarding the Wayoró form), suggesting that this word has been diffused via horizontal transmission in the region.

4. Evolution in the daughter languages

In this section, we outline the evolution of the consonantal onsets of Proto-Tuparian in the daughter languages (except those which arose from erstwhile codas via resyllabification).

Proto-Tuparian to Makurap. The evolution of the Proto-Tuparian onsets in Makurap involves as few as four innovations: (i) fortition of the approximants *β and *j to Mak *p*, *c* (**3.1**); (ii) loss of the velar articulation in *w, which became Mak *β*/*m* (as per nasality of the nucleus); (iii) fronting of *c into Mak *t*; (iv) consonantization of the prevocalic instances of *i into *ndʒ*/*ɲ* (as per nasality of the nucleus), as in the third person prefix *i- (**3.2**). Note that PTpr *ɲ remained in Makurap as *ɲ*, but its phonological status appears to have changed: it is analyzed as the nasal allophone of */j/ in Proto-Tuparian, but as the nasal allophone of */ɲ/ in Makurap due to the influx of *ndʒ* (from earlier *i before vowels). This is shown in Figure 1.

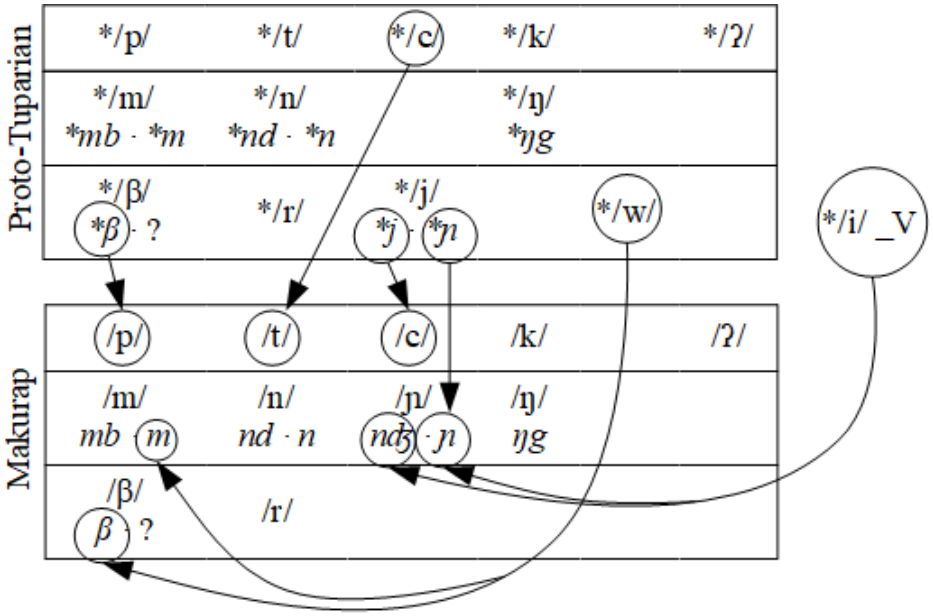


Figure 1. Evolution of the Proto-Tuparian onsets in Makurap

Proto-Tuparian to Proto-Core Tuparian. The Proto-Tuparian onsets have evolved in the following way in Proto-Core Tuparian: (i) the postoralized allophones of PTpr underlying nasal stops (*mb /m/, *nd /n/, *ŋg /ŋ/) lose their nasal phase and become PCT *b, *d, *g (**3.3**); (ii) PTpr *p and *t nasalize to PCT *m and *n preceding nasal nuclei; (iii) *j dentalizes to PCT *ð (**3.1**); (iv) in some polysyllabic relational stems, the stem-initial */j/ disappears; (v) the prevocalic instances of *i consonantize to *j*/*ɲ* (as per nasality of the nucleus), as in the third person prefix *i- (**3.2**); (vi) *w loses its labial articulation preceding rounded vowels and becomes *ɣ

(phonologically, still an allophone of */w/). Note that in PCT the voiced stops **b* and **d* occurred in oral environments only and were thus in a complementary distribution with PCT **m* and **n*. We tentatively analyze PCT **b*/**m* and **d*/**n* as allophones of underlying */b/ and */d/ conditioned by the nasality of the nucleus, but other solutions cannot be at present ruled out. The evolution of the Proto-Tuparian onsets in Proto-Core Tuparian is schematized in Figure 2.

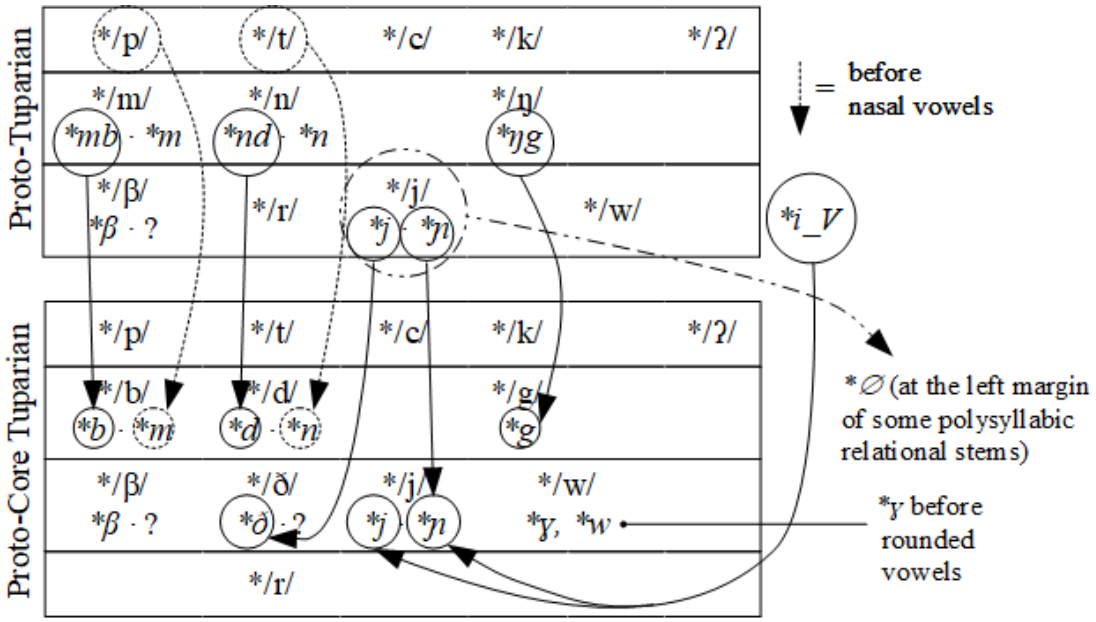


Figure 2. Evolution of the Proto-Tuparian onsets in Core Tuparian

Proto-Core Tuparian to Wayoró. For an accurate description of the evolution of the PCT onsets in Wayoró, it is useful to distinguish between two chronological stages, which we dub *pre-Wayoró* and *Wayoró*. In pre-Wayoró, PCT **b*, **d*, and **g* were lenited to **β*, **ð*, and **ɣ*, thus merging with pre-existing PCT **β*, **ð*, and **ɣ* (**3.3**). Note that **ɣ* is analyzed as an allophone of */w/ in PCT, where its occurrence was conditioned by a following rounded vowel. In pre-Wayoró, however, */ɣ/ and */w/ are contrastive thanks to the influx of **ɣ* < PCT **g* (compare pre-Wayoró **we*P- 'to climb' and **ɣe* 'garden'). Another minor source of pre-Wayoró **β* is PCT **o* before vowels (notably in the 1SG prefix **o-*; **3.2**).

After the completion of these processes, the approximant-rich sound system of pre-Wayoró suffered further changes. Namely, pre-Wayoró **β*, **ð*, **j*, **ɣ*, **w* were fortitioned to *mb*, *nd*, *ndʒ*, *ŋg*, *ŋg*^{w} in oral environments and to *m*, *ŋ*, *ŋ*^{w} in nasal environments (no examples are available for pre-Wayoró **ð* or **j* in nasal environments); exceptionally, **-β-* and **-j-* between vowels yielded *-β-* and *-dʒ-* in modern Wayoró. Synchronically, Nogueira (2019) analyzes *mb*, *nd*, *ndʒ*, *ŋg*, *ŋg*^{w} as allophones of underlying /m n ɲ ŋ ŋ^{w}/, which are thus derived via nasal shielding (Wetzels & Nevins 2018).[79]

[79] Although we provisionally accept Nogueira's (2019) analysis in this paper, there appear to be serious arguments for alternatively analyzing Wayoró *mb*, *nd*, *ndʒ*, *ŋg*, *ŋg*^{w} as allophones of voiced stops /b d dʒ g g^{w}/ derived via nasal venting (Wetzels & Nevins 2018). Note that *mb*, *ndʒ*, *ŋg*, and *ŋg*^{w} may only occur word-initially and after nasal vowels; between oral vowels, [β dʒ g g^{w}] occur (Nogueira 2019: 8, 62, fn. 1). This distribution matches perfectly all four predictions made by Wetzels & Nevins (2018: 842) for nasal venting, including (i) the absence of prenasalized voiceless consonants; (ii) the absence of prenasalized fricatives; (iii) the decreasing susceptibility of consonants to nasal venting as one moves from velars to labials (as in pre-Wayoró **aɣɯ*P 'louse', **ewi*T 'honey', **eðɯ* '*marico* bag' > *ãŋgɯ*P, *ẽŋg*^{w}*i*T, *endɯ*, but **aβi* 'seed' > *aβi*); and (iv) greater prominence of the nasal venting in prosodic domain-initial position as compared to the intervocalic position (compare **βo* 'hand' > *mbo*, but **o-βo* 'my hand' >

The epenthesis of *k*^{w} in the environment *o,ʉ_V* must have proceeded through the stage **owV/*ʉwV*; that way, a glide was inserted and subsequently fortitioned. Both processes must have applied **after** the sound change **w* > *ŋ(g)*^{w}, because otherwise we would expect tokens such as PTpr **ɨa:C* > pre-Wayoró **ʉa:C* 'tapir' to have participated in them, yielding **ʉwa:C* > **ʉ̃ŋg*^{w}*a:C* (rather than the attested *ʉk*^{w}*a:C*).

Finally, the change from PCT **c* to Wayoró *tʃ* has not involved any phoneme splits or mergers and cannot be straightforwardly attributed to any specific chronological level. In fact, it is not even clear whether any sound change has been involved at all, because almost nothing is known about the articulation of PTpr/PCT **c* other than that it was a voiceless coronal segment capable of changing to *t*, *tʃ*, or *Ø*. In Figure 3 below, it has been provisionally attributed to the pre-Wayoró stage, but it should be kept in mind that other scenarios are also possible.

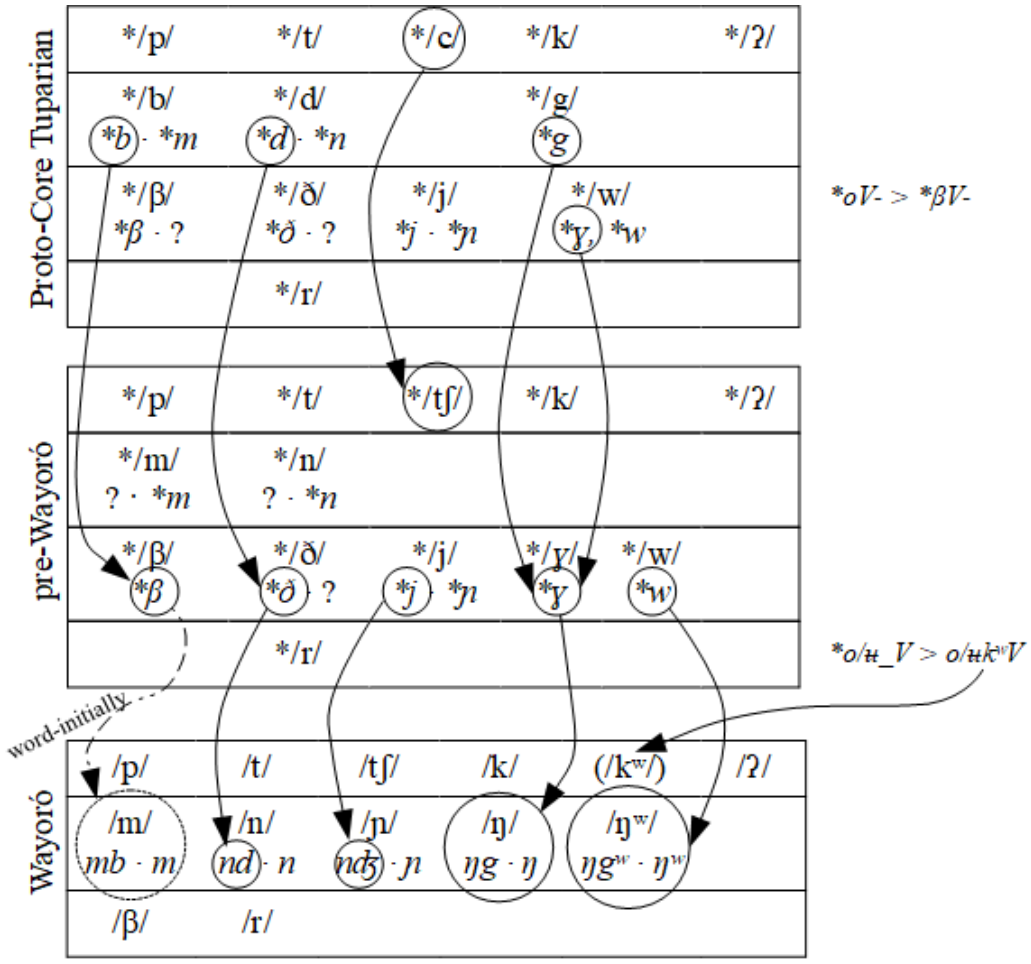


Figure 3. Evolution of the Proto-Core Tuparian onsets in Wayoró

Proto-Core Tuparian to Tuparí. The phonological history of Tuparí is characterized by multiple mergers, which resulted in the loss of the original voice distinctions reconstructed for Proto-Core Tuparian, as well as by a number of palatalization processes. A full list of the innovations we could identify follows: (i) mergers of **p/*b/*β* > **p*, **t/*d* > **t*, **k/*g* > **k* (**3.1**, **3.3**); (ii) **p* is palatalized before *i* to *-ps-* (between vowels) or *s-* (elsewhere); (iii) **t* is palatalized to *s*

o-βo). Future studies in Wayoró phonology should determine whether it is possible to posit the phonemes /b d dʒ g gʷ/ for the language instead of /m n ɲ ŋ ŋʷ β/.

before *i*/*ɨ*; (iv) **j* is fortitioned to *s* (**3.2**) but its nasal counterpart *ɲ* stays except before *ĩ*, where it is deleted (**3.1**); (v) **ð* is debuccalized to *h* word-initially and disappears between vowels (**3.1**); (vi) **ɣ* (an allophone of PCT */w/) is deleted (**3.1**); (vii) **c* is deleted at least between vowels; (viii) PCT **o* is consonantized to *w* before unrounded vowels and fuses with following rounded vowels yielding long vowels (notably in the 1SG prefix **o-*; **3.2**). This is schematized in Figure 4 below.

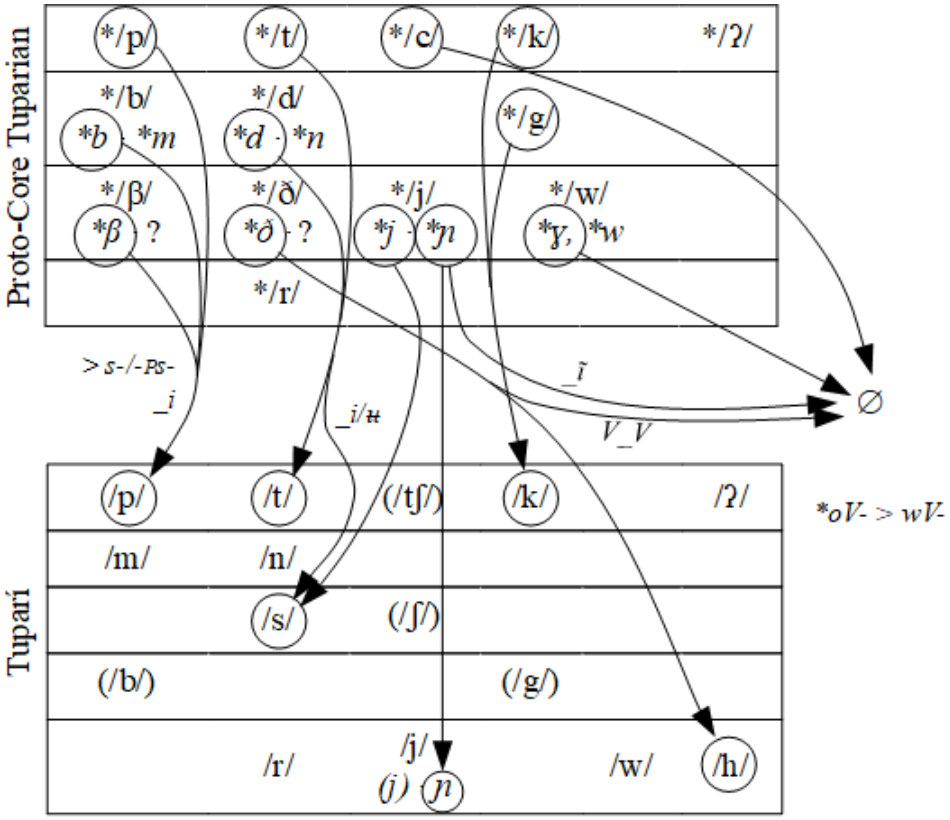


Figure 4. Evolution of the Proto-Core Tuparian onsets in Tuparí

Proto-Core Tuparian to Mekéns/Akuntsú. It is convenient to present the evolution of the Proto-Core Tuparian onsets in the Corumbiara languages Mekéns and Akuntsú by dividing the phonological history of these languages into two stages. At stage 1, a chain shift affects the voiced oral segments of PCT: the approximants **β*, **ð*, **ɣ*, **w* become voiced stops (**b*, **d*, **g*, **g*^{w}; **3.1**), whereas the PCT voiced stops lose voice (**b*, **d*, **g* > **p*, **t*, **k*; **3.3**). These processes are also fed by the glide insertion (PCT **i_V*, **o*/*ɨ_V* > **ijV*, **o*/*ɨwV*; **3.4**). At stage 2, **t* becomes an affricate (Proto-Corumbiara **ts* > Mek *ts*, Aku *tʃ*). Innovative Proto-Corumbiara **t* (> Mek/Aku *t*) comes from two sources: fronting of **c* > **t* and devoicing of **d*/**g*/**g*^{w} > **t*/**k*/**k*^{w}. The latter process occurred in all Corumbiara varieties, with the proviso that in the history of the Sakurabiat dialect it was bled by the nasalization of voiced stops in nasal environments (as in **bãkɨɲã*, **g*^{w}*ãẽ*, **gõ* > *mãkɨɲã* 'agouti', *ŋ*^{w}*ãẽ* 'pot', *ŋõ* 'pet'; **d* is not known to have occurred in nasal environments). In other dialects of Mekéns and in Akuntsú, there was no nasalization of voiced stops, and the devoicing of **d*/**g*/**g*^{w} applied categorically.

Fortition and devoicing also affected pre-Proto-Corumbiara **j*, which yielded *ts* in Sakurabiat and *t* in Akuntsú. Based on these reflexes, one could be tempted to reconstruct Proto-Corumbiara **ʝ*, which would have been subsequently devoiced to **c* > Sak *ts*, Aku *t*. However, the fact that the Guaratira and Siokweriat dialects of Mekéns have *∅* corresponding to Sak *ts*, Aku *t* in the environment *i_V*, implies that the fortition of **j* must have occurred only recently

in the history of the Corumbiara languages: while the sound change **ijV* > *iV* is easily conceivable, it would be more difficult to account for a sound change such as **icV*, **iɟV*, or **itsV* > *iV*. A detailed reconstruction of the evolution of **j* in the Corumbiara languages awaits further investigation.

Figure 5 summarizes our proposal regarding the development of the PCT onsets in Mekéns and Akuntsú.

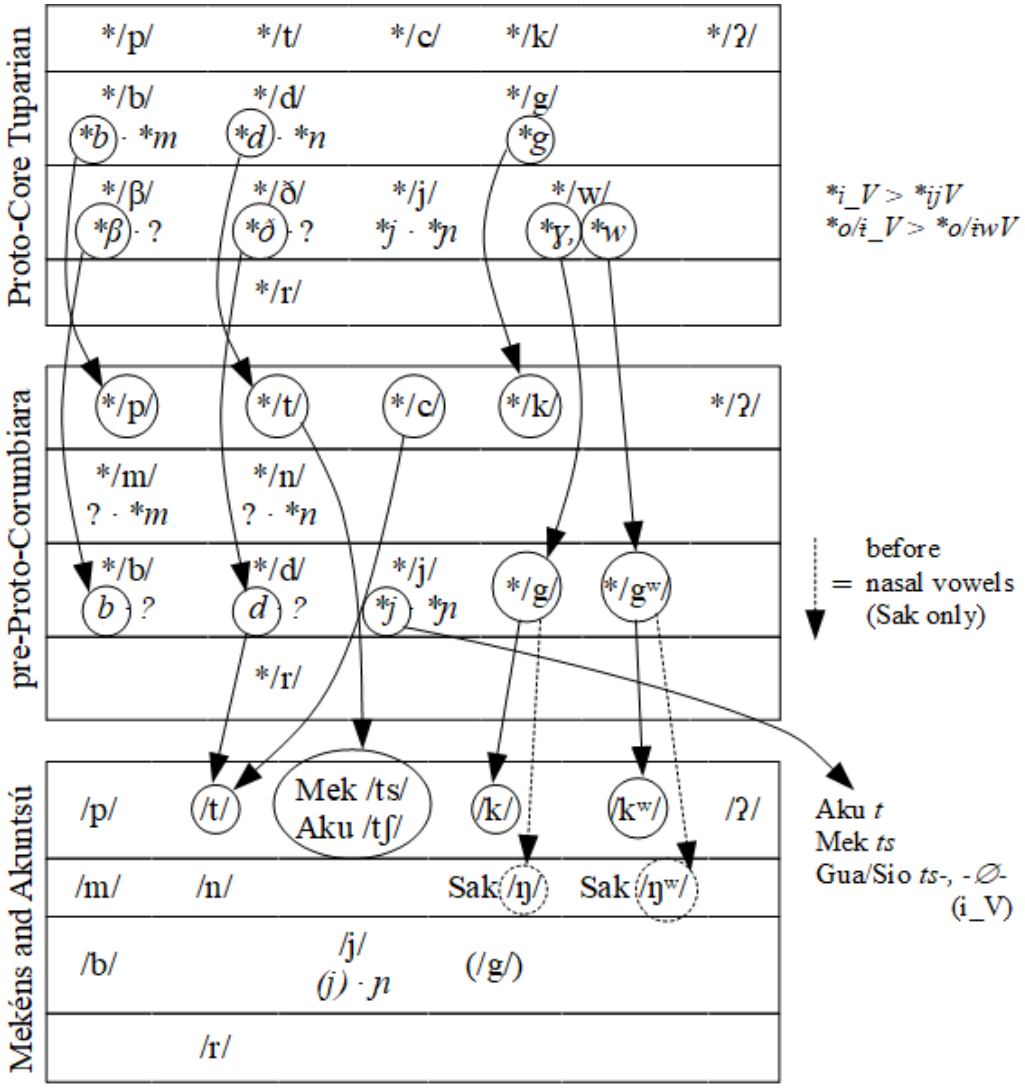


Figure 5. Evolution of the Proto-Core Tuparian onsets in Mekéns and Akuntsú

5. Conclusions

This paper has advanced the reconstruction of the evolution of a number of Proto-Tuparian segments in daughter languages, with a focus on the Proto-Tuparian approximants as well as on sounds which went through an approximant stage at least in some Tuparian languages. Compared to earlier works on the historical phonology of this genetic unit, our proposal stands out in taking into account its subgrouping, thus allowing us to identify crucial innovations restricted to specific clades (such as Core Tuparian and Corumbiara), as well is in considering important data sources which have become available only recently (notably Aragon 2014; Singerman 2016, 2018; Nogueira 2019).

Future research on the historical phonology of Proto-Tuparian should address issues such as the evolution of the Proto-Tuparian glottal stop (which is known to frequently disappear in the daughter languages under unknown conditions), the coda resyllabification patterns, the reconstruction of vowel length in Proto-Tuparian, and the reconstruction of the Proto-Tuparian prosody (which should minimally account for the contrastive stress in Tuparí and Akuntsú and for the tonal patterns of Makurap).

It is hoped that this study will inform further comparative research of the entire Tupian family, and we believe that at least some of our results are of interest to the typology of sound change (cf. Blevins 2008) and phonological theory (cf. Wetzels & Nevins 2008).

Abbreviations

Aku = Akuntsú
Gua = Guaratira
Kup = Kupndiiriat
Mak = Makurap
Mek = Mekéns
Ngw = Ngwayoroiat
Sak = Sakurabiat
Sio = Siokweriat
Tup = Tuparí
Way = Wayoró

PCor = Proto-Corumbiara
PCT = Proto-Core Tuparian
PTG = Proto-Tupí-Guaraní
PTpr = Proto-Tuparian
PWT = Proto-Wayoró–Tuparí

1/3 = first/third person
CAUS = causative
FOC = focus
DIFF = diffuse locative
INCL = inclusive
LOC = locative
NCRF = non-coreferential
PL = plural
POSS = possessed
PST = past
SG = singular
TH = thematic vowel

References

Alves, Ana Carolina, Ana Vilacy Galucio. 2007. Análise fonética acústica das vogais orais da língua Sakurabiat. *Letras de Hoje* 42(3): 27–42.

Alves, Poliana Maria. 2004. *O léxico do Tuparí: proposta de um dicionário bilíngüe.* Ph.D. dissertation, Universidade Estadual Paulista "Júlio de Mesquita".

Anonymous. No date. *Palavras e frases da lingua Wyarú.* Rio de Janeiro, manuscript (personal archive of Darcy Ribeiro).

Aragon, Carolina Coelho. 2008. *Fonologia e aspectos morfológicos e sintáticos da língua Akuntsú.* MA thesis, Universidade de Brasília.

Aragon, Carolina Coelho. 2014. *A grammar of Akuntsú, a Tupían language.* Ph.D. dissertation, University of Hawai'i at Mānoa.

Aragon, Carolina Coelho, Ana Suelly Arruda Câmara Cabral. 2005. A posição da língua Akuntsú na família lingüística Tuparí. In: Lúcia Maria Pinheiro Lobato; Stella Maris Bortoni-Ricardo; Ana Suelly Arruda Câmara Cabral; Heloisa Maria Moreira Lima Salles; Maria Marta Pereira Scherre; Daniele Marcelle Grannier (orgs.). *Anais do IV Congresso Internacional da ABRALIN*: 1533–9. Brasília: Universidade de Brasília.

Aragon, Carolina Coelho, Fernando O. de Carvalho. 2007. Análise acústica das vogais orais da língua Akuntsú. *Revista da ABRALIN* 6(2): 41–55. doi:10.5380/rabl.v6i2.52623

Aragon, Carolina Coelho, Luciana Keller Tavares. 2019. *Aspectos da cultura material do povo Akuntsú.* Paper presented at the 3° Congresso Internacional Povos Indígenas da América Latina (CIPIAL). Brasília: Universidade de Brasília. Retrieved from: <https://www.academia.edu/39865327/Aspectos_da_cultura_material_do_povo_Akunts%C3%BA>, accessed on 02/07/2020.

Becker-Donner, Etta. 1955. Notizen über einige Stämme an den rechten Zuflüssen des Rio Guaporé. *Archiv für Völkerkunde* 10: 275–343.

Blevins, Juliette. 2008. Consonant epenthesis: natural and unnatural histories. In: Good, Jeff (ed.). *Linguistic Universals and Language Change*: 79–107. Oxford University Press.

Blust, Robert. 1994. Obstruent epenthesis and the unity of phonological features. *Lingua* 93(2–3): 111–139. doi:10.1016/0024-3841(94)90001-9

Bontkes, Willem. 1978. *Dicionário preliminar Suruí-português, portugués-Suruí.* Porto Velho: Summer Institute of Linguistics.

Braga, Alzerinda. 1992. *A fonologia segmental e aspectos morfofonológicos da língua Makurap (Tupí).* MA thesis, Universidade Estadual de Campinas.

Braga, Alzerinda. 2005. *Aspects morphosyntaxiques de la langue Makurap/Tupi.* Ph.D. dissertation, Université de Toulouse–Le Mirail.

Caldas, Raimunda Benedita Cristina. 2009. *Uma proposta de dicionário para a língua Ka'apór.* Ph.D. dissertation, Universidade de Brasília.

Carvalho, Fernando O. de. 2019. Revisitando o Proto-Jurúna: A reconstrução das oclusivas orais. In: Edna dos Santos Oliveira, Eduardo Alves Vasconcelos, Romário Duarte Sanches (eds.). *Estudos linguísticos na Amazônia*: 215–236. Campinas: Pontes Editores.

Carvalho, Fernando O. de. Forthc. *The Historical Phonology of the Tupi-Guarani Language Family* (working title). Berlin: Language Science Press.

Chacon, Thiago Costa. 2012. *The phonology and morphology of Kubeo: The documentation, theory, and description of an Amazonian language.* Ph.D. dissertation, University of Hawai'i at Mānoa.

Corbera Mori, Ángel. 2008. Aspectos da fonologia segmental Mehináku (Aruák). *Estudos Lingüísticos* 37(1): 63–72.

Corrêa-da-Silva, Beatriz Carretta. 2010. *Mawé/Awetí/Tupí-Guaraní: relações lingüísticas e implicações históricas.* Ph.D. dissertation, Universidade de Brasília.

Drude, Sebastian. 2009. Nasal harmony in Awetí: a declarative account. *ReVEL – Revista Virtual de Estudos Da Linguagem* 3.

Epps, Patience. 2020. South American Languages. In: Claire Bowern, Patience Epps, Jane Hill, Patrick McConvell. *Languages of Hunter-Gatherers and their Neighbors: Database.* Retrieved from <https://huntergatherer.la.utexas.edu>, accessed in 19/07/2020.

Gabas Jr., Nilson. 1999. *A Grammar of Karo, Tupi (Brazil).* Ph.D. dissertation, University of California, Santa Barbara.

Gabas Jr., Nilson. 2005. Classificação da língua Akuntsú. *Estudos Lingüísticos* 36: 105–10.

Galucio, Ana Vilacy. 1994. Fonologia segmental da língua Mekens. In: *Anais do IX Encontro Nacional da ANPOLL (Caxambu, junho de 1994). Área de Lingüística, v. 2*: 988–97. João Pessoa: Associação Nacional de Pós-Graduação e Pesquisa em Letras e Lingüística, 1994.

Galucio, Ana Vilacy. 2001. *The morphosyntax of Mekens (Tupi).* Ph.D. dissertation, University of Chicago.

Galucio, Ana Vilacy. 2002. Word order and constituent structure in Mekens. *Revista da ABRALIN* 1(2): 51–74. doi:10.5380/rabl.v1i2.52693

Galucio, Ana Vilacy. 2011a. Nominalization in the Mekens language. *Amerindia* 35: 237–60.

Galucio, Ana Vilacy. 2011b. Subordinate adverbial constructions in Mekens. In: Rik van Gijn, Katharina Haude, Pieter Muysken (eds.). *Subordination in native South American languages*: 25–44. Amsterdam: John Benjamins.

Galucio, Ana Vilacy. 2014. Estrutura argumental e alinhamento gramatical em Mekéns. In: Luciana Storto, Bruna Franchetto, Suzi Lima (orgs.). *Sintaxe e semântica do verbo em línguas indígenas do Brasil*: 167–96. Campinas: Mercado de Letras.

Galucio, Ana Vilacy, Antônia Fernanda de Souza Nogueira. 2012. Comparative study of the Tupari branch of the Tupi family: contributions to understanding its historical development and internal classification. In: *Memorias del V Congreso de Idiomas Indígenas de Latinoamérica, 6–8 de octubre de 2011, Universidad de Texas en Austin*. Austin: Center for Indigenous Languages of Latin America (CILLA). Retrieved from <https://ailla.utexas.org/sites/default/files/documents/Galucio_Nogueira_CILLA_V.pdf>, accessed on 02/07/2020.

Galucio, Ana Vilacy, Antônia Fernanda de Souza Nogueira. 2018. From object nominalization to object focus: the innovative A-alignment in the Tuparian languages (Tupian family). *Journal of Historical Linguistics* 8(1): 95–127. doi:10.1075/jhl.16025.gal

Galucio, Ana Vilacy, Mercedes Guaratira Sakyrabiar, Manoel Ferreira Sakyrabiar, Rosalina Guaratira Sakyrabiar, Olimpio Ferreira Sakyrabiar. 2017. Sakurabiat. In: Kristine Stenzel, Bruna Franchetto (eds.). *On this and other worlds: Voices from Amazonia*: 331–51. Berlin: Language Science Press. doi:10.5281/zenodo.1008787

Galucio, Ana Vilacy, Nilson Gabas Júnior. 2002. Evidências de agrupamento genético Karo-Puruborá, tronco Tupi. In: *XVII Encontro Nacional da ANPOLL: boletim informativo n° 31*: 163. Porto Alegre: Universidade Federal do Rio Grande do Sul.

Galucio, Ana Vilacy, Sérgio Meira, Joshua Birchall, Denny Moore, Nilson Gabas Júnior, Sebastian Drude, Luciana Storto, Gessiane Picanço, Carmen Reis Rodrigues. 2015. Genealogical relations and lexical distances within the Tupian linguistic family. *Boletim do Museu Paraense Emílio Goeldi. Ciências Humanas* 10(2): 229–74. doi:10.1590/1981-81222015000200004

Hanke, Wanda, Morris Swadesh, Arion Dall'Igna Rodrigues. 1958. Notas de Fonologia Mekens. In: Juan Comas (ed.). *Miscellanea Paul Rivet octogenario dicata, vol. II*: 187–217. Mexico City: Universidad Nacional Autónoma de México.

Landin, David. 2005. *Dicionário e léxico Karitiana / português.* 2nd ed. Cuiabá: Sociedade Internacional de Lingüística.

Lacerda, María Conceição de. 2014. *Bekã Pamakube (lugar de aprender). Aprendendo com os Zoró: análise da identidade indígena através da experiencia das escolas nas aldeias do povo indígena Zoró*. Ph.D. dissertation, Universidad de Salamanca. doi:10.14201/gredos.124106

Lévi-Strauss, Claude. No date. *Vocabulaires de trois langues: Kabišiana, Kepkiri-uat et de São Pedro*. Ms. Paris.

Lévi-Strauss, Claude. 1950. Documents Rama-Rama. *Journal de la Société des Américanistes* 39: 73–84. doi:10.3406/jsa.1950.2381

Loukotka, Čestmír. 1963. Documents et vocabulaires inédits de langues et de dialectes sud-américains. *Journal de la Société des américanistes* 52: 7–60. doi:10.3406/jsa.1963.2001

Loukotka, Čestmír. *Classification of South American Indian languages.* Los Angeles: Latin American Center, University of California; Caracas: Sucre, 1968. (Reference Series, v. 7.)

Meira, Sérgio, Sebastian Drude. 2015. A summary reconstruction of Proto-Maweti-Guarani segmental phonology. *Boletim do Museu Paraense Emílio Goeldi. Ciências Humanas* 10(2): 275–96. doi:10.1590/1981-81222015000200005

Mello, Antônio Augusto Souza. 2000. *Estudo histórico da família lingüística Tupi-Guarani: aspectos fonológicos e lexicais.* Ph.D. dissertation, Universidade Federal de Santa Catarina.

Mondini, Juliana Nazatto. 2014. *Yudja Utaha: a culinária juruna no Parque Indígena Xingu – uma contribuição ao dicionário bilíngue juruna-português.* MA thesis, Universidade Estadual Paulista "Júlio de Mesquita Filho" (Araraquara campus).

Moore, Denny, Ana Vilacy Galucio. 1994. Reconstruction of Proto-Tupari consonants and vowels. In: Margaret Langdon, Leanne Hinton (eds.). *Proceedings of the Meeting of the Society for the Study of the Indigenous Languages of the Americas, July 2–4, 1993, and the Hokan-Penutian Workshop, July 3, 1993, both held at the 1993 Linguistic Institute at Ohio State University in Columbus, Ohio*: 119–37. Berkeley: Survey of California and other Indian Languages.

Nikulin, Andrey. Submitted. *O Kabišiana de Lévi-Strauss é um codialeto do Sakurabiat (Mekéns).*

Nikulin, Andrey, Fernando O. de Carvalho. 2019. Estudos diacrônicos de línguas indígenas brasileiras: um panorama. *Macabéa – Revista Eletrônica do Netlli* 8(2): 255–305.

Nikulin, Andrey, Mário André Coelho da Silva. 2020. As línguas Maxakalí e Krenák dentro do tronco Macro-Jê. *Cadernos de Etnolingüística* 8(1): 1–64.

Nikulin, Andrey, Serafima Muzykantova. In preparation. *A phonological reconstruction of Proto-Ta-Arawakan.*

Nogueira, Antônia Fernanda de Souza. 2011. *Wayoro ẽmẽto: fonologia segmental e morfossintaxe verbal.* MA thesis, Universidade de São Paulo. doi:10.11606/D.8.2011.tde-22102012-093617

Nogueira, Antônia Fernanda de Souza. 2015. Dissimilação em Wayoro (família Tuparí, tronco Tupí). In: Larissa Lima, Marília Ferreira, Eunice Braga, Jaqueline Reis, Zerben Aguiar (orgs.). *Anais do IX Congresso Internacional da ABRALIN*: 611–18. Belém: ABRALIN, PPGL/UFPA.

Nogueira, Antônia Fernanda de Souza. 2019. *Predicação na língua Wayoro (Tupi): propriedades de finitude.* Ph.D. dissertation, Universidade de São Paulo. doi:10.11606/T.8.2019.tde-14082019-101006

Nogueira, Antônia Fernanda Souza, Ana Vilacy Galucio, Nicole Soares-Pinto, Adam Roth Singerman. 2019. Termos de parentesco nas línguas Tuparí (família Tupí). *Boletim do Museu Paraense Emílio Goeldi. Ciências Humanas* 14(1): 33–64. doi:10.1590/1981.81222019000100004

Picanço, Gessiane Lobato. 2005. *Munduruku: phonetics, phonology, synchrony, diachrony.* Ph.D. dissertation, University of British Columbia. doi:10.14288/1.0092991

Picanço, Gessiane Lobato. 2019. *A fonologia diacrônica do Proto-Munduruku (Tupí).* Curitiba: Appris.

Ramirez, Henri, Valdir Vegini, Maria Cristina Victorino de França. 2017. O warázu do Guaporé (tupi-guarani): primeira descrição linguística. *LIAMES* 17(2): 411–506. doi:10.20396/liames.v17i0.8647468

Reiter, Sabine. 2011. *Ideophones in Awetí.* Ph.D. dissertation, Christian-Albrechts-Universität zu Kiel.

Ribeiro, Maria de Jesus Pacheco. 2010. *Dicionário Sateré-Mawé/português.* MA thesis, Universidade Federal de Rondônia (Guajará-Mirim campus).

Rodrigues, Aryon Dall'Igna. 1958. Die Klassifikation des Tupi-Sprachstammes. In: *Proceedings of the Thirty-Second International Congress of Americanists, Copenhagen 1956*: 679–84. Copenhagen: Munksgaard.

Rodrigues, Aryon Dall'Igna. 1985. Evidence for Tupi-Carib Relationships. In: Harriet E. Manelis Klein, Louisa R. Stark (eds.). *South American Indian languages: retrospect and prospect*: 371–404. Austin: University of Texas Press.

Rodrigues, Aryon Dall'Igna. 2002. Correspondências lexicais e fonológicas entre Tupí-Guaraní e Tuparí. In: Ana Suelly Arruda Câmara Cabral, Aryon Dall'Igna Rodrigues (orgs.). *Línguas indígenas brasileiras: Fonologia, gramática e história. Atas do I Encontro Internacional do Grupo de Trabalho sobre Línguas Indígenas da ANPOLL. Tomo I*: 288–97. Belém: Editora da Universidade Federal do Pará.

Rodrigues, Aryon Dall'Igna. 2005. As vogais orais do Proto-Tupí. In: Aryon Dall'Igna Rodrigues, Ana Suelly Arruda Câmara Cabral (orgs.). *Novos estudos sobre línguas indígenas*: 35–46. Brasília: Editora UnB.

Rodrigues, Aryon Dall'Igna. 2007. As consoantes do Proto-Tupí. In: Ana Suelly Arruda Câmara Cabral, Aryon Dall'Igna Rodrigues (orgs.). *Línguas e culturas Tupí, vol. I*: 167–203. Campinas: Curt Nimuendajú; Brasília: LALLI/UnB.

Rodrigues, Aryon Dall'Igna, Ana Suelly Arruda Câmara Cabral. 2012. Tupían. In: Lyle Campbell, Veronica Grondona (eds.). *The Indigenous languages of South America: a comprehensive guide*: 495–574. Berlin/Boston: Mouton de Gruyter. (The World of Linguistics, 2). doi:10.1515/9783110258035.495

Rodrigues, Aryon Dall'Igna, Franz Caspar. 2017. *Esboço da gramática da língua tuparí.* Brasília: LALLI/UnB.

Sekelj, Tibor. 1948. *Wordlist Aruá-Makurap-Zaboti-Arikapú-Tupari.* Ms. Retrieved from http://etnolinguistica.wdfiles.com/local--files/sekelj%3A1/Sekelj_Wordlist_Aru%C3%A1_Makurap_Zabot%C3%AD_Arikap%C3%B3_Tupar%C3%AD.pdf, accessed on 28/08/2020.

Silva, Bazilio da, Nilzimara de Souza Silva, Odamir de Oliveira (orgs.). 2013. *Paradakary urudnaa: dicionário Wapichana/português, português/Wapichana.* Boa Vista: EDUFRR.

Singerman, Adam Roth. 2016. Nasal harmony and phonotactic well-formedness in Tupari. *International Journal of American Linguistics* 82(4): 453–85. doi:10.1086/688603

Singerman, Adam Roth. 2018. *The morphosyntax of Tuparí, a Tupían language of the Brazilian Amazon.* Ph.D. dissertation, University of Chicago. doi:10.6082/ns2a-cv36

Snethlage, Emil Heinrich. 2015. *Die Guaporé-Expedition (1933–1935): in Forschungstagebuch* (aus dem Nachlass herausgegeben von Rotger M. Snethlage, Alhard-Mauritz Snethlage und Gleice Mere). Wien/Köln/Weimar: Böhlau Verlag. doi:10.7788/9783412218591

Valenzuela Bismarck, Pilar. 2011. Contribuciones para la reconstrucción del proto-cahuapana: comparacion léxica y gramatical de las lenguas jebero y chayahuita. In: Willem F. H. Adelaar, Pilar Valenzuela Bismarck, Roberto Zariquiey Biondi (eds.). *Estudios en lenguas andinas y amazónicas. Homenaje a Rodolfo Cerrón-Palomino*: 271–304. Lima: Fondo Editorial Pontificia Universidad Católica del Perú.

Voort, Hein van der. 2005. Kwaza in a comparative perspective. *International Journal of American Linguistics* 71(4): 365–412. doi:10.1086/501245

Voort, Hein van der. 2007. Proto-Jabutí: um primeiro passo na reconstrução da língua ancestral dos Arikapú e Djeoromitxí. *Boletim do Museu Paraense Emílio Goeldi. Ciências Humanas* 2(2): 133–68. doi:10.1590/S1981-81222007000200007

Wetzels, W. Leo, Andrew Nevins. 2018. Prenasalized and postoralized consonants: The diverse functions of enhancement. *Language* 94(4): 834–66. doi:10.1353/lan.2018.0055

Xerez, Sebastião Moacyr. 1946. *Pequeno vocabulário dos dialetos Macurape, Mondé e Caritiana*. Manaus.

А. Никулин, Р. Андради. Развитие и падение аппроксимантов в языках тупарийской группы

В статье рассматривается развитие серии аппроксимантов в языках тупарийской группы (тупийская семья), изначально распространённых между средним течением реки Гуапоре́/Ите́нес и верховьями Маша́ду/Жи-Парана́ (юг штата Рондония, Бразилия). Показано, что помимо исходной серии пратупарийских аппроксимантов (в нашей ре-

конструкции, **β*, **j* и **w*) в некоторых языках-потомках развились аппроксиманты и из других источников — из гласных ненижнего подъёма (**o*/**i*), посторализованных носовых (**mb*/**nd*/**ŋg*, через ступень **b*/**d*/**g*) и глайдов, заполняющих зияние. Детально обсуждается эволюция этих звуков; в частности, показывается, что во всех тупарийских языках хотя бы некоторые бывшие аппроксиманты подверглись фортиции. Особое внимание уделяется внутренней классификации тупарийской группы.

Ключевые слова: тупарийская группа; тупийская семья; аппроксиманты; фортиция; сравнительно-исторический метод.

М. Е. Васильев[†], *М. Н. Саенко*[‡]
[†] Институт языкознания РАН (Россия, Москва); mvhumanity@gmail.com
[‡] Институт славяноведения РАН (Россия, Москва); veraetatis@yandex.ru

Анализ топологии и оценка точности лексикостатистических классификаций (на примере славянских языков)*

Благодаря своей простоте и универсальности лексикостатистика остается одним из самых популярных методов для установления языкового родства и построения генеалогических классификаций. Среди российских компаративистов наибольшее распространение получило приложение Starling, использующее видоизменённую методику «присоединения соседей» при реконструкции филогенетических деревьев. Применение данной методики на материале разных языковых семей показывает хорошие или правдоподобные результаты в большинстве случаев. В то же время исследователи отмечают ряд недостатков в построенных классификациях, наиболее существенными из которых являются неустойчивость структуры древа даже к минимальным изменениям в составе идиомов, а также наличие в ней фиктивных таксонов и узлов, трудно объяснимых или противоречащих существующим представлениям.

В данной статье проводится детальное рассмотрение отмеченных проблем на примере лексической классификации 25 славянских идиомов. При этом показано, что главной причиной обоих явлений является несовершенство процедуры построения древа, используемой в Starling. По результатам исследования была предложена методика, позволяющая минимизировать влияние установленных недостатков путем выявления в топологии древа недостоверных узлов (на основе статистических расчетов) и их последующего исключения. Особенности предложенной методики делают ее применимой для анализа любых лексикостатистические классификаций, а также легко реализуемой в виде дополнительного компонента Starling или отдельного приложения.

Ключевые слова: лексикостатистика; метод присоединения соседей; генеалогическая классификация; среднее абсолютное отклонение.

I. Введение

Несмотря на то, что вопрос о надежности и эффективности лексического критерия при изучении языкового родства регулярно становится предметом обсуждения[1], лексикостатистические классификации по-прежнему остаются востребованными и одними из наиболее популярных среди исследователей. Главными причинами привлекательности лексикостатистики, очевидно, являются ее универсальность и относительная простота по сравнению с другими существующими методами. Так, например, если выявление общих инноваций, которое, как правило, считается более надежным классификационным критерием, требует диахронического исследования фонетических, морфологических и лек-

* Авторская работа М. Е. Васильева осуществлена при поддержке Российского научного фонда (проект № 20–18–00159); организация, осуществлявшая финансирование, — Институт языкознания Российской академии наук (ИЯз РАН). Авторская работа М. Н. Саенко выполнена в рамках работы по теме НИР «Славянская межкультурная и межъязыковая коммуникация в диахронии и синхронии».

[1] См., например, Бурлак, Старостин 2005: 148; Starostin 2010: 194; Поздняков 2014: 221–222; Грунтов, Мазо 2015: 211–216.

сических изменений в каждом из сравниваемых языков на протяжении всего рассматриваемого периода, то для лексикостатистического анализа достаточно сведений о составе базисного словаря этих языков всего на один произвольно выбранный момент времени, причем для каждого языка этот момент может быть разным. Это обстоятельство не только значительно упрощает получение генеалогической классификации, но также позволяет использовать лексикостатистику для установления родственных связей между малоизученными языками, в случае с которыми применение других методов часто оказывается затрудненным из-за недостатка данных.

Еще одним немаловажным фактором является формализованный характер лексикостатистических расчетов, благодаря которому процедуру построения древа можно выполнить автоматически с помощью компьютерной программы. Такая возможность была реализована в приложении Starling[2], разработанном С. А. Старостиным в 1985–2000 гг. и получившим распространение как среди российских, так и зарубежных компаративистов[3]. Для определения степени родства между языками программа рассчитывает процент этимологически совпадающих лексем в их 100-словных (или 110-словных) списках и по итогам расчетов формирует таблицу с долями совпадений между всеми идиомами попарно. Непосредственно построение генеалогического древа осуществляется на основе полученной таблицы с помощью методики[4], представляющей собой несколько видоизмененный и адаптированный для лингвистического материала метод «ближайших соседей»[5], широко применяемый в биологии.

За последнее десятилетие представителями Московской школы компаративистики был накоплен значительный опыт применения лексикостатистического метода для построения генеалогических классификаций различных языков мира, включающих как близкородственные малые группы, так и крупные языковые общности с большой временной глубиной. При этом построенные деревья повсеместно используются в ходе исследовательской работы, регулярно приводятся в научных публикациях, а также демонстрируются во время конференций.

Анализируя полученные результаты[6], исследователи, как правило, подчеркивают полезность классификации в целом и акцентируют внимание на ее особенностях, значимых для целей исследования, но в то же время указывают на отдельные несоответствия в структуре древа, плохо поддающиеся объяснению или противоречащие известным данным. Среди таких странностей особо следует выделить два наиболее характерных недостатка, которые проявляются независимо от выбора рассматриваемых языков и, как можно предположить, обусловлены самой методикой формирования деревьев:

---

[2] StarLing for Windows, v. 2.6.10: computerized system for multilingual database processing, (c) 1985–2005 by S. A. Starostin, StarLing Software Inc. Текущая версия программы доступна на сайте проекта «Вавилонская башня» по адресу: https://starling.rinet.ru/downl.php?lan=ru#soft.

[3] В отличие от многочисленных существующих программ для построения филогенетических деревьев, предназначенных в первую очередь для классификации биологических видов, Starling изначально создавался как специализированное приложение для сбора, обработки и анализа именно лексических данных, благодаря чему и завоевал свою популярность у лингвистов.

[4] Суть данной методики подробно разбирается в учебнике (Бурлак, Старостин 2005: 163–167). Некоторые из ее особенностей мы более подробно рассмотрим далее.

[5] Метод «ближайших соседей» или «присоединения соседей» (Neighbor-Joining Method) — алгоритм построения филогенетических деревьев, в основу которого положен принцип последовательного попарного объединения «ближайших» (т.е. имеющих наибольшее сходство) таксонов. Первоначально метод предназначался для классификации нуклеотидных последовательностей (см. Saitou, Nei 1987), однако в дальнейшем стал широко применяться также за пределами генетики.

[6] См., например, обсуждение полученных классификаций в работах Kogan 2016: 235–238; Vydrin 2009: 112–114.

1. Конфигурация древа является неустойчивой и крайне чувствительна к изменению количества или состава идиомов. В частности, нередки случаи, когда исключение или добавление одного языка приводит к абсолютно неожиданным и радикальным изменениям в топологии, затрагивающим не только таксон с новым или исключенным элементом, но также ветви, максимально от него удаленные.
2. Древо содержит большое количество незначимых узлов, интерпретация которых крайне проблематична или невозможна на основании имеющихся сведений об истории развития языков и их взаимной дивергенции. Как правило, такие узлы располагаются в непосредственной близости или на незначительном временном расстоянии от других узлов, а иногда образуют непрерывные цепочки в виде характерной ступенчатой структуры.

Несмотря на то, что обе указанные особенности очень распространены (и хорошо знакомы всем пользователям Starling), они крайне редко удостаиваются отдельного обсуждения: в большинстве случаев авторы ограничиваются констатацией несовершенства методики, с которым неизбежно приходится мириться. При этом большинство исследователей признают, что наличие подобных «артефактов» существенно снижает практическую ценность построенных деревьев, а также ставит под сомнение их достоверность. Таким образом, мы сталкиваемся с необходимостью анализа выявленных методических погрешностей, а также поиска способов их оценки и минимизации.

В рамках нашей статьи мы рассмотрим возможный подход к решению данной задачи на примере лексикостатистической классификации 25 славянских языков и диалектов, уделяя особое внимание вышеупомянутым проблемам вариативности и незначимой кластеризации в строении деревьев.

II. Исходные данные

Дадим краткое описание идиомов, списки базисной лексики которых задействованы в исследовании[7]. В соответствии с принципами проекта «Глобальная лексикостатистическая база данных», в рамках которого были собраны списки, предпочтение отдавалась диалектным данным, поскольку базисная лексика литературных языков, предположительно, отличается бóльшим консерватизмом.

1. *Банатский болгарский*. Переселенческий говор, на котором говорят болгары-католики в румынском и сербском Банате. Переселение состоялось в двух волнах. Сначала в 1688 г. после неудачного восстания в Валихию сбежали жители города Чипровци и окрестностей. В 1720-е гг. к ним присоединились так называемые павликиане из-под Свиштова и Николпола. Обе группы встретились и смешались в Банате, где восточноболгарский говор более многочисленных свиштовцев и николполцев почти полностью вытеснил западноболгарский говор чипровцев (Стойков 2002: 193). Словарь составлен выдающимся болгарским диалектологом С. Стойковым на материале, собранном начиная с 1953 г., преимущественно в румынских селах Стар-Бешенов (рум. Dudeştii Vechi) и Винга (рум. Vinga) (Стойков 1968).

2. *Македонский д. Горно-Каленик*. Говор деревни Горно-Каленик, которая находится в Греции (греч. Ἀνω Καλλινίκη), недалеко от города Лерин (греч. Φλώρινα). Материал был собран П. Хиллом преимущественно в Австралии у македонцев, сбежавших из Греции

[7] Большая часть списков с описанием и аннотацией доступна по ссылке https://starling.rinet.ru/cgi-bin/response.cgi?root=new100&basename=new100\ier\slv.

Рисунок 1. География идиомов, списки которых используются в исследовании

во время гражданской войны (Hill 1991). Относится к леринскому говору юго-западного диалекта македонского языка.

3. *Штокавский сербохорватский племени Загарач*. Говор черногорского племени Загарач (местное произношение – *Загàрач*), населяющего территорию вокруг горы Гарач (Ћупић, Ћупић 1997). Относится к зетско-южносанджакскому диалекту штокавского наречия, согласно классификации П. Ивича.

4. *Чакавский сербохорватский о. Вргада*. Говор острова Вргáда. Словарь составлен хорватским лингвистом, носителем говора, Б. Юришичем на основе записей 1908–1960 гг. (Jurišić 1973). По классификации Брозовича и Ивича, говор относится к южночакавскому диалекту.

5. *Чакавский сербохорватский д. Орлец*. Говор деревни О́рлец (местное произношение – *Õrlec*), расположенной на острове Црес. Словарь составлен Х.П. Хоутзагерсом на основе полевых записей 1980–1982 гг. (Houtzagers 1985). По классификации Брозовича и Ивича, говор относится к северночакавскому диалекту.

6. *Чакавский сербохорватский д. Орбаничи*. Говор деревни Орбáничи (местное произношение — *Orbânići*), находящейся в двух километрах от города Жминь в центральной Истрии. Словарь составлен нидерландской исследовательницей Я. Калсбек на основе материала, собранного в 1980–1984 гг. (Kalsbeek 1998). Согласно классификации Брозовича и Ивича, говор относится к юго-западному истрскому диалекту чакавского наречия.

7. *Чакавский сербохорватский д. Девинска-Нова-Вес*. Переселенческий говор деревни Девинска-Нова-Вес в Словакии (по-словацки *Devínska Nová Ves*, в говоре — *Nụ̑ovo sèlo*; в настоящее время — район Братиславы). Носители говора — градищанские хорваты, поселившиеся на этой территории в XVI в. Словарь составлен чешским исследователем В. Важным на основе полевых записей 1923–1926 гг. Помимо основного материала из деревни Девинска-Нова-Вес часть была записана в соседних деревнях Дубравка (*Dúbravka*; *Dubráva*) и Ламач (*Lamač*; *Làmụ̑oč*) (Vážný 1927).

8. *Градищанский кайкавский сербохорватский*. Переселенческий говор, на котором говорят в двух деревнях в Венгрии — Хидегшег (венг. *Hidegség*, произношение в говоре — *Hęḍęšîn* / *Hęḍęšîn*) и Фертёхомок (*Fertőhomok*; *Hòmok*). Предки носителей говора переселились в начале XVI века из Славонии, предположительно из населенных пунктов Кралева-Велика (*Kraljeva Velika*) и Меджурич (*Međurić*), которые находятся значительно восточнее современной границы кайкавского наречия. Словарь составлен Х. П. Хоутзагерсом на основе полевых записей 1985–1994 гг. (Houtzagers 1999).

9. *Чабарский словенский*. Говор окрестностей города Ча́бар (схр. *Čàbar*) в Хорватии. Словарь составлен С. Малнаром (Malnar 2008). Говор относится к костельскому диалекту доленьского наречия словенского языка.

10. *Костельский словенский*. Говор деревни Де́лач (произношение в говоре — *ˈDėlaːč*) и окрестностей составляет южную часть костельского диалекта доленьского наречия. Словарь был составлен Й. Грегоричем, уроженцем Делача (Gregorič 2014).

11. *Чрновршский словенский*. Диалект плато Чрни-Врх (слвн. *Črni Vrh*) относится ровтарскому наречию. Словарь был составлен И. Томинцем, носителем диалекта, преимущественно на основе говора его родной деревни Ломе (слвн. *Lome*) (Tominec 1964).

12. *Словенский д. Затолмин*. Говор деревни Затолми́н (слвн. лит. *Zatolmin*, в говоре — *Zatˈmin*), лежащей в 1 км от города Толмин в западной Словении, недалеко от границы с Италией. Материал собирался носительницей говора Х. Чуец-Стрес свыше десяти лет, начиная с 1996 г. (Čujec Stres 2011, 2014). Говор относится к толминскому диалекту ровтарского наречия.

13. *Словенский д. Била*. Говор деревни Би́ла (ит. *San Giorgio*, в говоре — *Bíla*) в Резьянской долине в Италии. Материал записан в 1987–1991 гг. Х. Стенвейком (Steenwijk 1992). Говор относится к резьянскому диалекту приморского наречия.

14. *Словенский д. Брдо*. Говор Зильской долины в Австрии. Материал был собран Т. Пронком в 2001–2006 гг. преимущественно у одной информантки, которая родилась и выросла в деревне Эгг (нем. *Egg bei Hermagor*, в говоре — *Bŕdo*), а после замужества проживала в Почахе (нем. *Potschach*, в говоре — *Pətǫ́čani*) (Pronk 2009). Говор относится к зильскому диалекту каринтийского наречия.

15. *Прлекийский словенский*. Говор деревень Бре́нгова (слвн. лит. *Brengova*, в говоре — *Bˈŕę̣ŋgọva*) и Це́нкова (слвн. лит. *Cenkova*, в говоре — *ˈCę̣ŋkọva*), входящий в северо-западную часть прлекийского диалекта паннонского наречия. Словарь составлен Б. Райхом (Rajh 2010).

16. *Подкрконошский чешский*. Диалект чешского Подкрконошья (чеш. *Podkrkonoší*), территории к югу от Крконошских гор (Bachmannová 2016).

17. *Моравский чешский д. Мистршице*. Говор деревни Мистршице (чеш. *Mistřice*), находящейся в 7 км от города Угерске-Градиште в Моравии. В словарь, составленный И. Малиной, включены также некоторые лексемы, записанные в соседних населенных пунктах (Malina 1946).

18. *Словацкий деревни Пилишсанто*. Переселенческий говор деревни Пилишсанто (венг. *Pilisszántó*), расположенной недалеко от Будапешта. Словаки поселились там предположительно в начале XVIII века. Большая часть пришла с территории Малых Карпат и говорила на западнословацком диалекте. Материал был собран Ф. Грегором в 1950-е гг. (Gregor 1975).

19. *Малопольский д. Венцюрка*. Говор деревни Венцюрка (пол. *Więciórka*), расположенной в мысленицком повяте Малопольского воеводства. Словарь, составленный уроженцем Венцюрки М. Куцалой, включает лексику, собранную в трех деревнях: Венцюрка (основной материал словаря), Сидзина-Гурна (*Sidzina Górna*) и Фацимех (*Facimiech*) (Kucała 1957).

20. *Коцевский великопольский*. Говоры региона Коцеве (*Kociewie*), относящиеся к великопольскому диалекту. Материал был записан Б. Сыхтой в 1930–1970-е гг. (Sychta 1–3).

21. *Белорусские говоры Гродненской области*. Говоры Гродненской области Белоруссии, входящей в ареалы юго-западного и центральнобелорусского диалектов. Материал собран Т. Ф. Стешкович в 1948–1960 гг. (Сцяшковіч 1972; Сцяшковіч 1983).

22. *Белорусские говоры Турова и окрестностей*. Говоры города Турова и 33 деревень в его окрестностях. Относятся к юго-западному диалекту белорусского языка. Материал записан коллективом исследователей в экспедициях 1967–1981 гг. (ТС 1–5).

23. *Украинский д. Торунь*. Говор села Торунь (местное произношение — *Tǫ́run*) в Закарпатской области. Относится к восточнобойковской группе юго-западных украинских говоров. Материал собран экспедицией под руководством С. *Л*. Николаева в 1990 г. (Николаев, Толстая 2001).

24. *Русский д. Деулино*. Говор деревни *Д*еулино (Рязанская область), относящийся к рязанской группе говоров южнорусского наречия. Материал записан в 1960–1963 гг. (ССРНГ 1969).

25. *Русский д. Островцы*. Говор деревни Островцы (Псковская область), относящийся к гдовской группе говоров среднерусского наречия. Материал собран в 1995–1998 гг. З. Хонселааром и опубликован в виде монографии, включающей словарь (Хонселаар 2001).

Как мы видим, имеющийся материал несколько неоднороден: какие-то словари предоставляют в наше распоряжение материал говора лишь одного населенного пункта, какие-то — нескольких, какие-то — целого большого региона. Некоторые словари являются дифференциальными, то есть дают лишь ту лексику, которая отличается от лексики литературного языка (и тогда собирать списки базисной лексики приходится по большей части из примеров, имеющихся в словаре), другие же — недифференциальными, то есть описывают словарный состав говора во всей его полноте. Кроме того, данные были записаны были в разное время и исследователями с отличающимися подходами. Неоднородность исходных данных значительно усложняет задачу для исследователя, желающего построить лексикостатистическое древо, и оказывает непосредственное влияние на качество полученной в итоге классификации.

Определенные искажения в структуру древа вносят случаи заимствований (в базе данных Starling им присваивает значение «-1») и синонимов или супплетивизма, когда одной строке в базе соответствует два или более корней. Также, к сожалению, не для всех списков удалось собрать полные 110-словные списки, иногда искомый материал отсутствует в словаре. В таблице 1 мы приводим краткие сведения о подобных изъянах в материале.

Оговорим сразу, что на основании собранных нами 25 списков нельзя построить репрезентативную классификацию славянских языков, поскольку они покрывают славянский мир неравномерно, и для создания качественного лексикостатистического древа требуется значительно больший объем данных. Однако в рамках данной работы мы и не ставим перед собой такой задачи. В нашем случае мы планируем использовать имеющийся славянский материал для проверки и уточнения некоторых аспектов современной лексикостатистической теории.

III. Анализ лексикостатистической классификации

*Д*ля расчета долей совпадений между 110-словными списками славянских идиомов и построения их генеалогической классификации использовалось приложение Starling («стандартный» метод). В результате проведённых вычислений была получена исходная лексикостатистическая матрица — Таблица 3 (см. Приложение), а также генетическое древо, представленное на рис. 2 ниже. Рассмотрим его более подробно.

№	Название	Лакуны	Заимствования	Синонимы или супплетивизм
1	Банатский	0	2 (belly, rain)	4 (bird, I, smoke, go)
2	Горно-Каленик	3 (bark, fat, swim)	1 (liver)	4 (come, I, person, say)
3	Загарач	1 (cloud)	0	7 (ashes, dog, I, leaf, person, say, we)
4	Вргада	0	2 (liver, sand)	4 (belly, I, person, we)
5	Орлец	0	1 (round)	3 (I, person, we)
6	Орбаничи	0	3 (dog, man, sand)	6 (fat, I, many, worm, person, we)
7	Девинска-Нова-Вес	0	3 (road, tree, snake)	3 (I, person, we)
8	Градищанский кайкавский	2 (moon, snake)	0	5 (belly, I, person, say, we)
9	Чабарский	0	0	4 (burn, person, road, we)
10	Костельский		1 (belly)	4 (I, louse, person, we)
11	Чрновршский	0	0	4 (I, many, person, we)
12	Затолмин	1 (worm)	0	3 (I, person, we)
13	Била	1 (warm)	5 (bark, fat, feather, mouth, tree)	4 (I, kill, person, we)
14	Брдо	0	2 (fly, neck)	3 (I, person, we)
15	Прлекийский	2 (bark, round)	1 (belly)	3 (I, person, we)
16	Подкрконошский	0	0	4 (I, many, person, we)
17	Мистршице	4 (cloud, fat, mountain, sand)	0	3 (I, person, we)
18	Пилишсанто	2 (bark, root)	1 (sand)	4 (cloud, I, person, we)
19	Венцюрка	0	5 (bark, feather, heart, red, skin)	4 (big, I, person, we)
20	Коцевский	1 (yellow)	2 (heart, red)	4 (I, many, person, we)
21	Гродненский	0	6 (heart, red, see, seed, skin, worm)	3 (hair, person, we)
22	Туровский	0	3 (dog, red, see)	5 (cloud, I, liver, person, we)
23	Торунь	1 (tooth)	3 (one, seed, short)	3 (I, many, we)
24	Деулинский	0	2 (cloud, dog)	3 (I, person, we)
25	Островцы	1 (horn)	3 (cloud, say, what)	5 (ashes, I, long, person, we)

Таблица 1. Лакуны, заимствования и синонимы в используемых списках

В целом полученную классификацию можно охарактеризовать как удовлетворительную. На древе отчетливо выделяются болгаро-македонский, восточно-славянский, словенско-сербохорватский и западнославянский таксоны. Отсутствие объединения болгаро-македонского и сербохорватско-словенского таксонов в южнославянскую подгруппу само по себе не является критическим: ряд исследователей не поддерживает выделение такой подгруппы в составе славянских языков (см. обзор в Blažek 2017). Намного важнее то, что на самом нижнем уровне древо выглядит неверно: болгаро-македонский таксон объединен с восточнославянским, а словенско-сербохорватский — с западнославянским. Сербохорватские и словенские идиомы, представленные наибольшим числом списков, в рамках своих таксонов ведут себя не вполне корректно: словенские говоры выстроились «лесенкой» без какого-либо выраженного диалектного деления. Из шести сербохорватских списков четыре являются чакавскими, однако они не выделились в особую подгруппу, а объединились попарно и разбились штокавским и кайкавским списками.

Таким образом, положительно можно оценить «среднюю» часть древа (то есть объединение идиомов в четыре подгруппы), и отрицательно – «нижнюю» и «верхнюю» части, в которых мы наблюдаем фантомные корневые узлы и ступенчатое членение словенских и сербохорватских говоров, не соответствующее лингвистической действительности.

Перейдем теперь к анализу внутренних свойств классификации и выясним, насколько ее структура зависит от изменений в составе рассматриваемых языков. Для этого сформируем из них 25 дополнительных выборок, поочередно исключая из полного списка по одному идиому, а затем сравним деревья, построенные для каждого нового набора, с исходным. В результате сопоставления[8] были найдены три идиома, исключение которых из классификации привело к значимым изменениям в топологии древа:

1. Македонский говор д. Горно-Каленик;
2. Чакавский сербохорватский говор д. Орлец;
3. Градищанский кайкавский сербохорватский.

Начнем наше рассмотрение с первого случая. На рис. 3 приведено генеалогическое древо 24-х славянских языков с исключенным македонским говором. Сравнивая его с исходной классификацией (рис. 2), мы обнаруживаем, что все основные таксоны, соответствующие восточнославянской, западнославянской, сербохорватско-словенской подгруппам, полностью сохранили свою целостность и внутреннее строение. В то же время сокращение выборки привело к неожиданным и весьма существенным изменениям в корневой части древа: после исключения македонского восточнославянские идиомы образовали единую общность с сербохорватскими, словенскими и западнославянскими языками, тогда как банатский болгарский оказался обособленным от всех остальных идиомов[9].

Последнее отличие выглядит особенно проблематично, так как раннее отделение болгарского от основного массива славянских говоров не подтверждается никакими лингвистическими данными.

На первый взгляд, примеры такой вариативности в конфигурации древа при минимальных изменениях в составе языков заставляют усомниться в практической ценности лексикостатистических классификаций и перспективах использования методики в целом. Однако, прежде чем делать подобные неутешительные выводы, следует принять во внимание, что доли совпадений, определяющие последовательность объединения таксонов и взаимное расположение узлов древа, в действительности являются не детерминированными[10], а *статистическими* величинами[11], которые обусловлены случайным характером процесса лексических замен и могут отклоняться от расчетных значений в большую или меньшую сторону. Это означает, что проценты совпадений, соответствующие узлам древа, имеют некоторый разброс — погрешность, которую необходимо учитывать как при построении, так и последующем анализе найденной топологии. Для количественной оценки данной погрешности мы воспользуемся величиной

[8] Все полученные деревья, а также исходные лексикостатистические данные представлены в сопровождающих материалах, которые доступны онлайн на сайте ВЯР.

[9] При этом разница между первым (корневым) и вторым узлами древа, соответствующим отделению болгарского и началу разделения остальных групп, достигает 2% (что эквивалентно разнице в 2 слова при сравнении 100-словных списков).

[10] То есть точно заданными.

[11] К сожалению, это важное обстоятельство, указание на которое содержится в самом названии лексикостатистики, в большинстве случаев попросту игнорируется при анализе лексикостатистических расчетов, что неизбежно приводит к абсурдным результатам и в конечном итоге дискредитирует весь метод в целом.

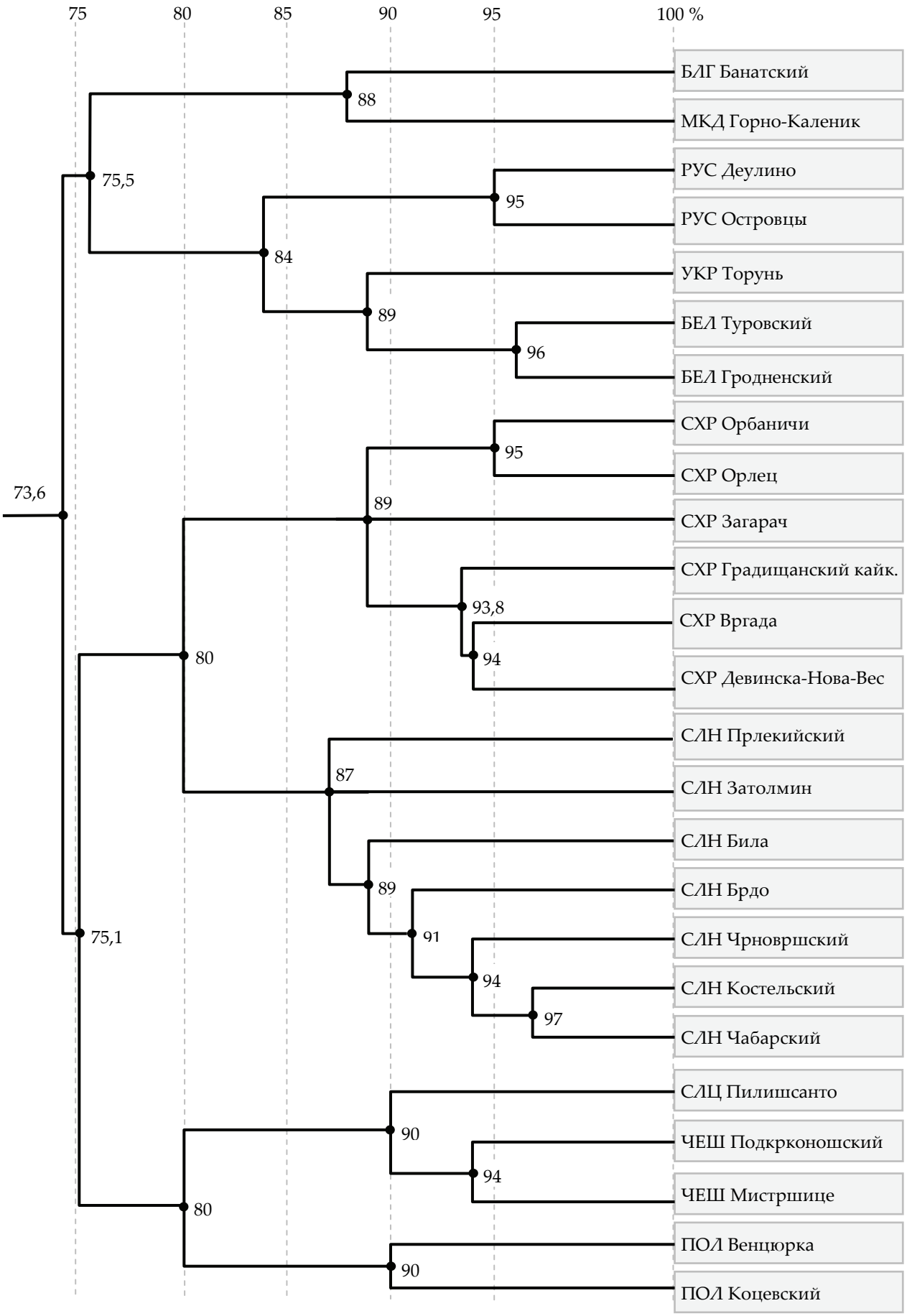


Рисунок 2. Лексикостатистическая классификация 25 славянских языков, полученная с помощью Starling (метод «Standard»). Значения на шкале и рядом с узлами древа соответствуют процентам совпадений.

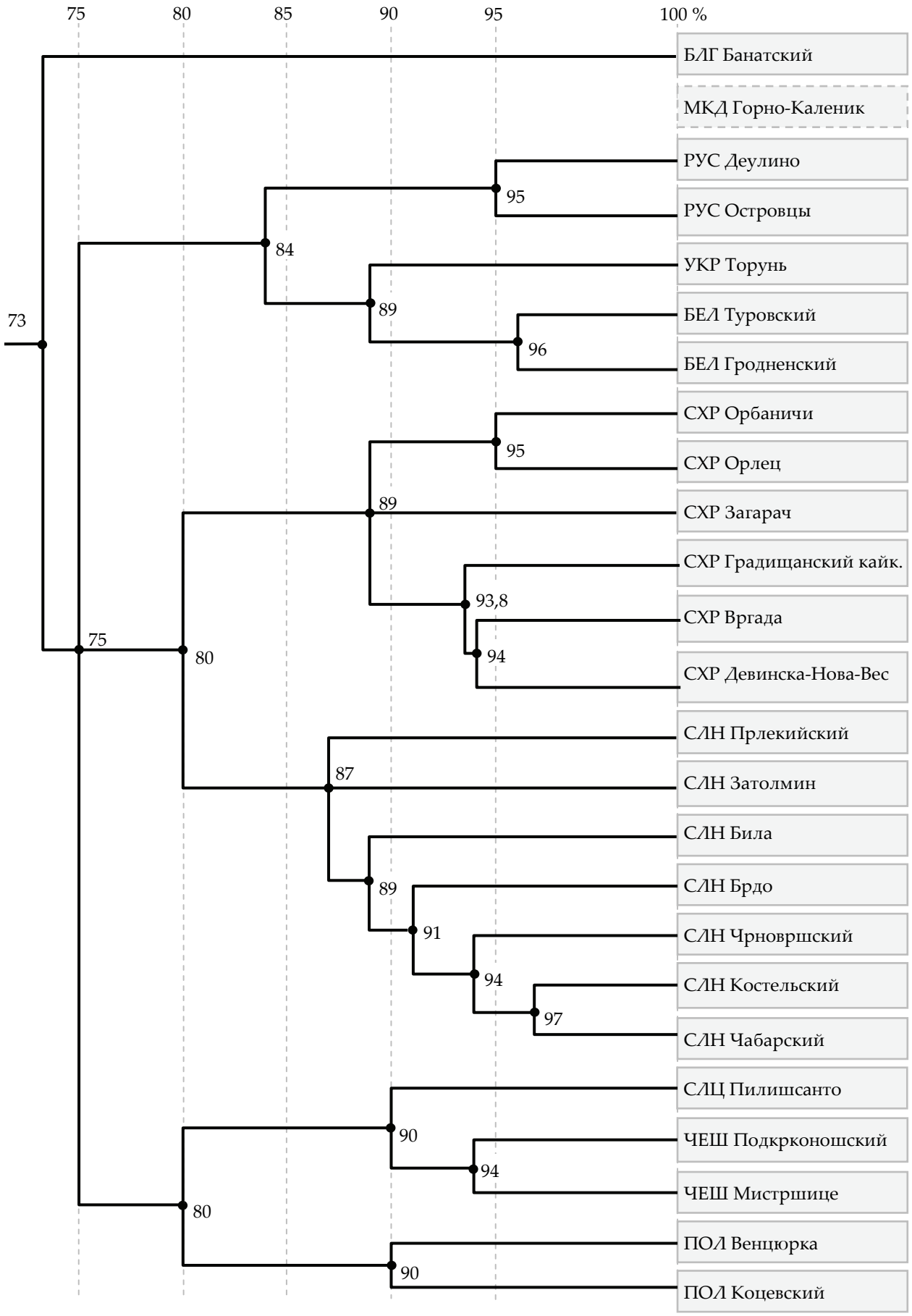


Рисунок 3. Генеалогическое древо 24 славянских идиомов, после исключения македонского. Значения на шкале и рядом с узлами древа соответствуют процентам совпадений.

среднего абсолютного отклонения[12], общий смысл которой поясним на следующем примере (см. табл. 2 и рис. 4):

Языки	A	B	C
A	-	90	84
B	90	-	86
C	84	86	-

Таблица 2. Проценты совпадений между списками языков A, B и C

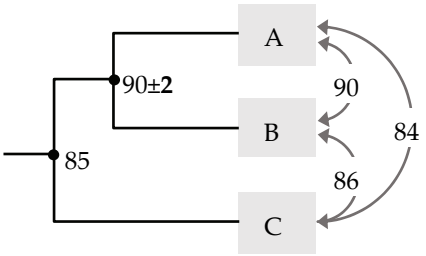


Рисунок 4. Генеалогическое древо языков A, B и C, полученное на основе таблицы

Согласно исходным данным (табл. 2), языки A и B являются ближайшими родственниками и образуют первый узел древа с процентом совпадений N_{AB}=90, после чего к ним остается присоединить оставшийся идиом C (рис. 4). При этом количество совпадений между языками A и C и между языками B и C отличается на два слова (N_{AC}=84; N_{BC}=86). Если мы исключим возможность заимствований и повторных сближений, то в рамках классической модели дивергенции такое расхождение можно объяснить только неравномерностью процесса замен в базисной лексике идиомов A и B, а именно: либо ускоренным лексическим изменением языка A, либо, наоборот, замедленным изменением языка B. Следовательно, в первом случае (с учетом двух «лишних» замен), количество общих слов в списках языков A и B составит N_{AB}=90+2=92, а во втором N_{AB}=90–2=88. Таким образом расчетная доля совпадений для узла A — B может лежать в диапазоне от 88% до 92%, что численно соответствует величине абсолютного отклонения E_{AB}, которая, для данного примера[13], рассчитывается по формуле:

$$E_{AB} = | N_{AC} - N_{BC} | = | 84 - 86 | = 2$$

Перейдем теперь к анализу полученных ранее классификаций славянских языков на основе средних абсолютных отклонений, рассчитанных для каждого узла (рис. 5 и рис. 6). Нетрудно заметить, что в строении обоих деревьев присутствует большое количество узлов, абсолютные отклонения которых накладываются друг на друга. Более того, среди них можно выделить несколько групп, в которых диапазоны отклонений включают в себя сами расчетные значения, полученные для соседних узлов. Особенно показательным примером является последовательное объединение идиомов Брда, Билы, Затолмина и Прлекии внутри словенской подгруппы, образовавших непрерывную цепочку из трех узлов с взаимным перекрытием. Причем в случае с Билой абсолютное отклонение составляет 2,6% и охватывает сразу два соседних узла.

[12] Методика расчета среднего абсолютного отклонения подробно излагается в работах Васильев 2010: 538–540; Васильев, Коган 2013: 160.

[13] В общем случае (для группы с произвольным количеством идиомов) используется более сложная формула — см. Васильев 2010: 540.

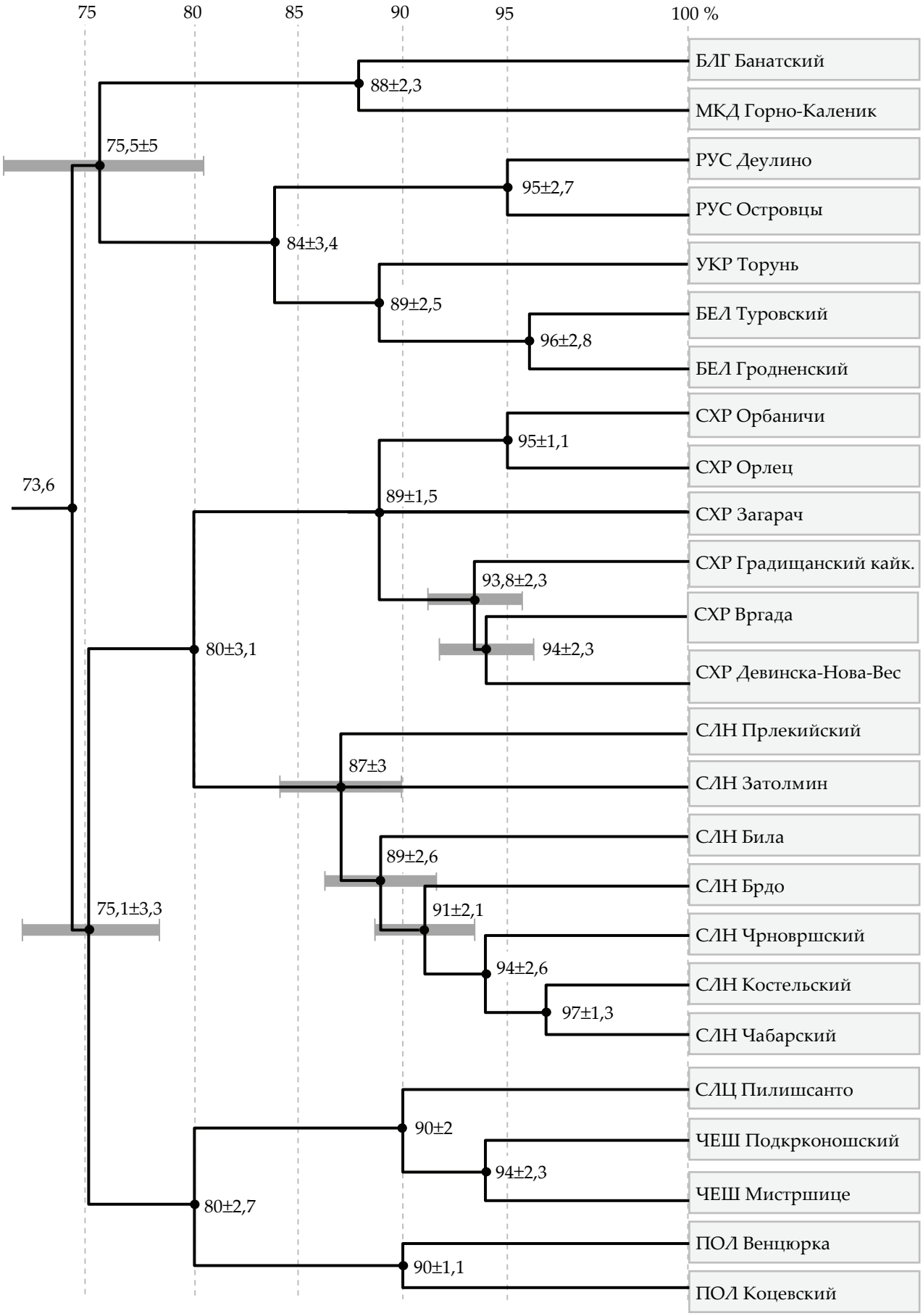


Рисунок 5. Генеалогическое древо 25 славянских идиомов с указанными процентами совпадений и значениями средних абсолютных отклонений. Диапазоны отклонений, перекрывающие соседние узлы, показаны отрезками.

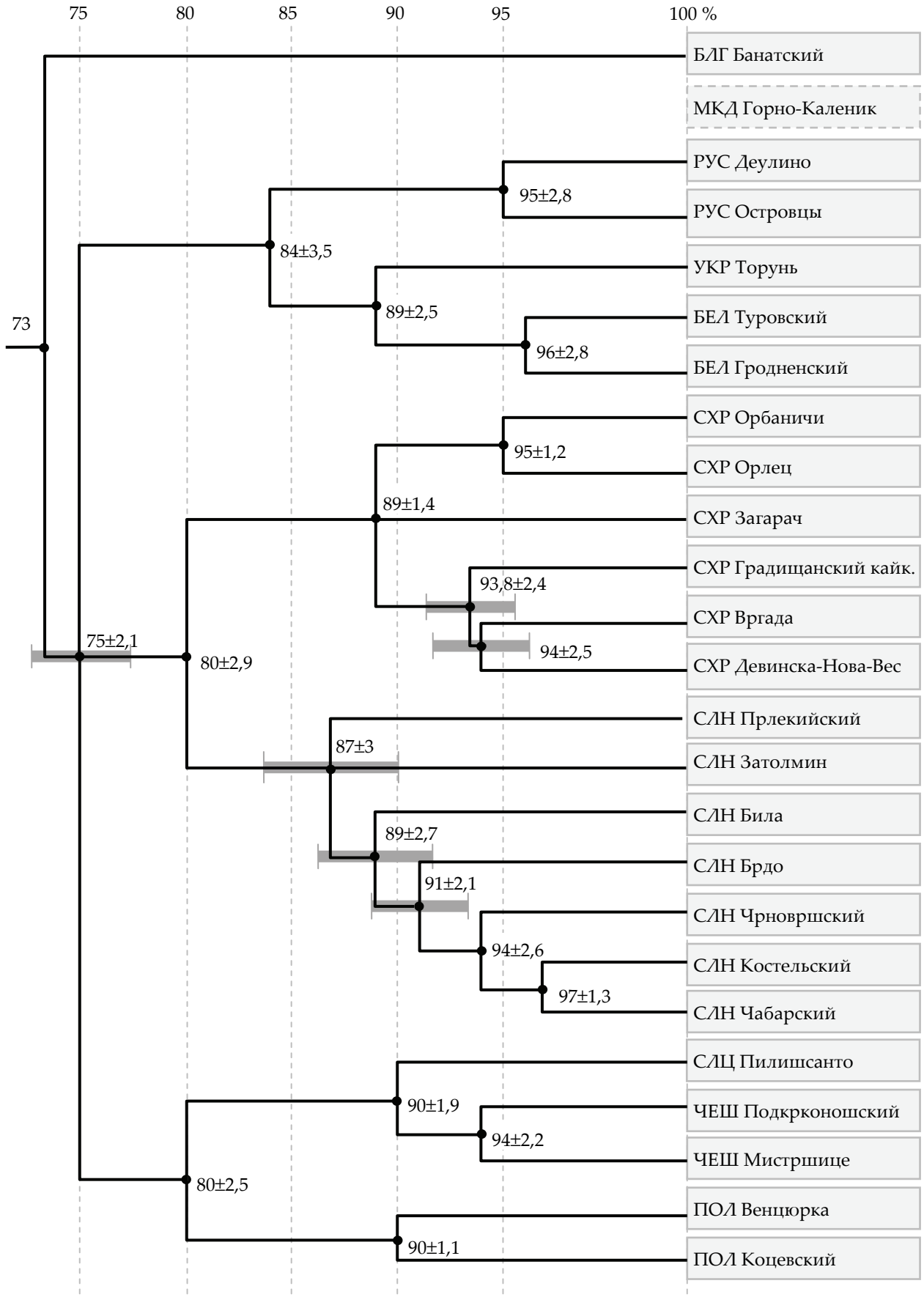


Рисунок 6. Генеалогическое древо 24 славянских языков после исключения македонского. Рядом с узлами указаны значения процентов совпадений и средних абсолютных отклонений. Диапазоны отклонений, перекрывающие соседние узлы, показаны отрезками.

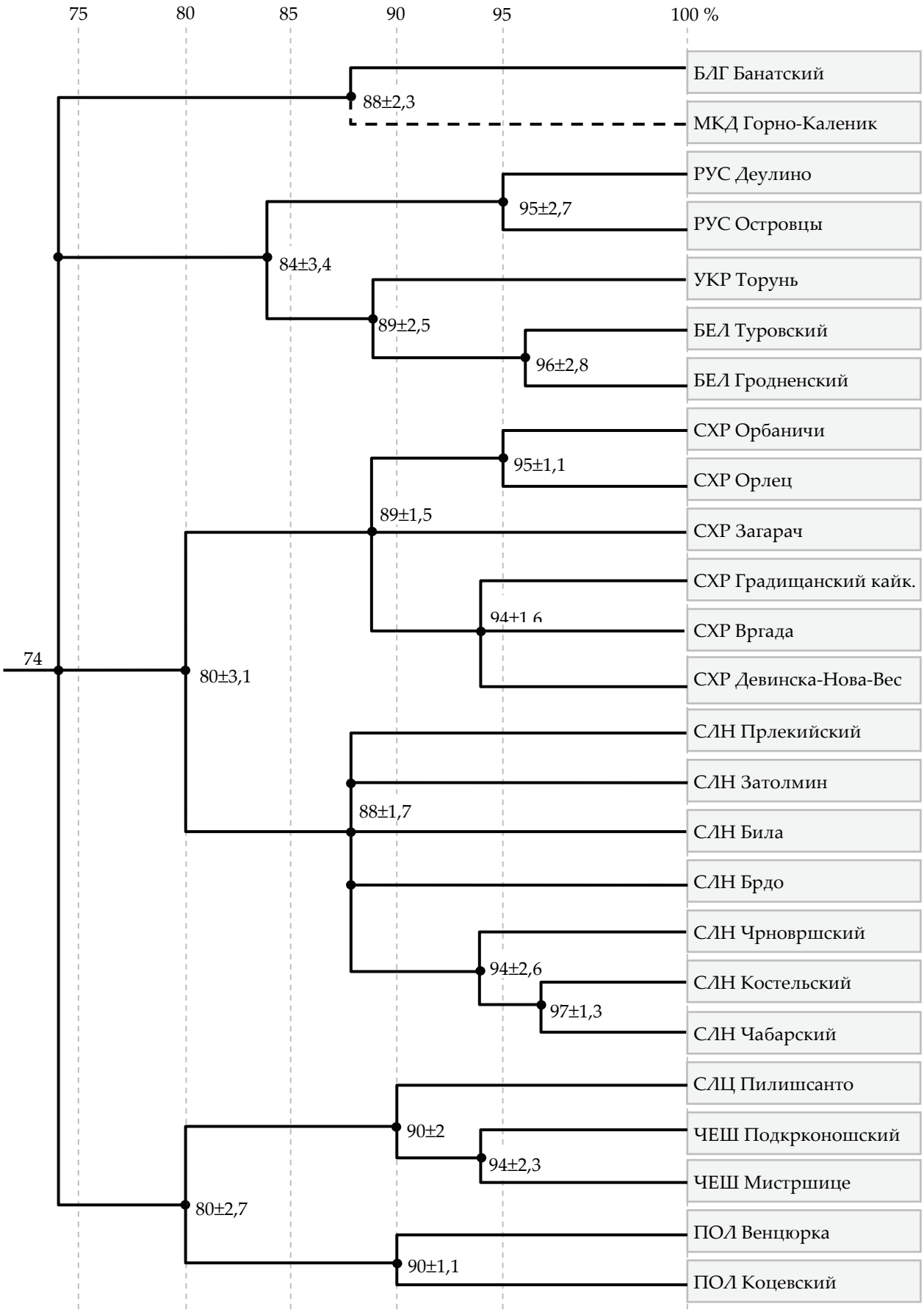


Рисунок 7. Генеалогическое древо славянских языков после объединения узлов с взаимно перекрывающимися диапазонами отклонений. Рядом с узлами приведены значения процентов совпадений и средних абсолютных отклонений.

Похожую ситуацию мы наблюдаем в основании деревьев. На рис. 5 корневой узел (со значением 73,6) перекрывается отклонениями обоих последующих узлов, в один из которых входят восточнославянский и болгаро-македонский таксоны, а в другой — западнославянский и сербохорватско-словенский таксоны. Аналогично на рис. 6 узел, соответствующий отделению болгарского, лежит в пределах отклонений узла, объединяющего все остальные ветви древа[14].

В этом и аналогичных примерах разница между процентами совпадений близкорасположенных узлов лежит в пределах погрешности, что свидетельствует об их *статистической неразличимости* на основе имеющихся данных. Следовательно, такие узлы можно[15] рассматривать как некий континуум и заменить их одним или несколькими более крупными узлами, после чего рассчитать для них новые значения долей совпадений, а также средние абсолютные отклонения. При этом, если узлы, полученные в результате объединения, также окажутся взаимно перекрывающимися, то процедуру можно повторять до тех пор, пока расстояние между любыми соседними узлами не будет превышать величину отклонений[16].

Применяя описанную методику к рассматриваемым классификациям, мы обнаружим, что после объединения узлов с перекрывающимися средними отклонениями и пересчета соответствующих значений оба древа оказались идентичными и приобрели вид, показанный на рис. 7. Очевидным образом это устраняет вариативность, вызванную изъятием македонского говора, поскольку конфигурация древа теперь остается неизменной вне зависимости от того, какую выборку (полную или сокращенную) мы используем. Кроме того, таксономическое «прореживание» позволило сократить количество фиктивных узлов[17] в группе словенских говоров, а также других ветвях древа и тем самым значительно смягчить проблему «топологического шума», обозначенную ранее.

Учитывая эффективность предложенной методики в случае с македонским, можно предположить, что покажет хорошие результаты и в остальных двух случаях, выявленных нами при исследовании топологии древа. Напомним, что они оба связаны с сербохорватскими говорами: чакавским острова Вргада и градищанским кайкавским. В отличие от предыдущего примера, исключение каждого из этих идиомов привело только к локальным изменениям внутри самой сербохорватской подгруппы и не затронуло остальные ветви древа. Тем не менее, в каждом случае это заметно повлияло на конечный вид классификации (см. рис. 9). В частности, после изъятия чакавского (рис. 9б) структура древа меняется до неузнаваемости: ранее плотная группа, состоявшая из кайкавского и двух чакавских говоров (Вргада и Девинска-Нова-Вес) распадается, причем один из них (Вргада) объединяется с другим чакавским (Орбаничи), второй образует

[14] Определить величину среднего абсолютного отклонения для корневого узла древа невозможно в силу особенностей методики — а именно, отсутствия «внешних» (по отношению к образованному узлу) языков, относительно которых можно было бы выполнить расчеты.

[15] Подчеркнем — статистическая неразличимость узлов не обязывает нас к их объединению, а только *указывает* на такую возможность. Поэтому, при наличии дополнительных (содержательных) аргументов в пользу дифференциации, близкие узлы не следует объединять, даже если диапазоны их отклонений взаимно перекрываются.

[16] Описанная методика представляет по сути своеобразный «топологический фильтр», который позволяет устранить случайные «помехи» в виде незначимых узлов, вызванных статистическими отклонениями в исходных лексических данных и «засоряющих» полезную структуру древа.

[17] Как мы уже отмечали, появление этих узлов вызвано недостатками бинарного принципа кластеризации, заложенного в методике «присоединения соседей», который не позволяет объединить более двух таксонов за один раз.

группу со штокавским (Загарач), а кайкавский становится обособленным идиомом, первым отделившимся от всей группы. В случае с градищанским (рис. 9в) мы не наблюдаем каких-либо радикальных изменений в топологии, однако обратим внимание на расположение главного узла, связывающего основные три ветви подгруппы. Его значение увеличилось сразу на 3 процента (с 89% до 92%), в результате чего произошло его сближение с узлом Вргада — Девинска-Нова-Вес (94%). Если теперь в каждом из трех фрагментов (а, б, в) мы объединим узлы с перекрывающимися отклонениями (они обведены пунктиром), то получим три несовпадающие топологии, что очевидно свидетельствует о неэффективности нашей методики в данном случае.

Как показывает дальнейший анализ, причина неудачи кроется в еще одном неучтенном факторе, а именно — особом способе подсчета процентов совпадений для узлов древа, реализованном в Starling. Согласно описанию методики расчетов в работе (Бурлак, Старостин 2005: 163–167), при объединении близкородственных языков (с долей общей лексики более 70%), следует выбирать не среднюю, а минимальную долю совпадений между ними. Авторы объясняют это тем, что «при близком родстве языков возможно вторичное их сближение, при котором трудно отличить более поздние заимствования от исконно родственной лексики»[18]. Поясним этот принцип на уже знакомом нам примере с языками A, B и C и рассчитаем долю совпадения для узла, связывающего идиом C с группой A+B.

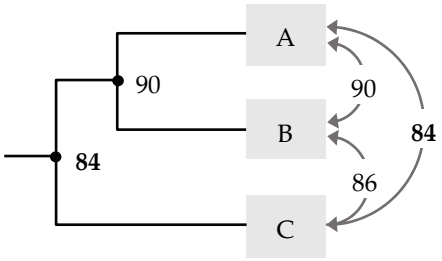
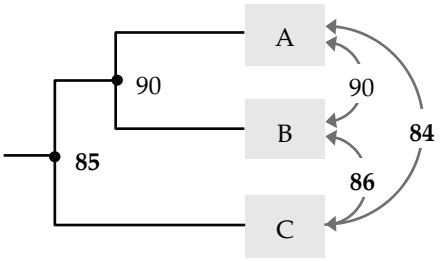


Рисунок 8. а) Расчет процентов совпадений по минимальному значению (Starling).



б) Расчет процентов совпадений по среднему значению.

Поскольку процент совпадений между списками идиомов A – C и B – C не совпадает и больше 70%, то мы, следуя данному правилу, должны выбрать наименьшее из двух значений — т. е. $N_{мин} = N_{AC} = 84$ (рис. 8а). Отметим, что выбранное значение будет отличаться от среднего процента совпадений, который для тех же языков составит $N_{ср} = (N_{AC} + N_{BC})/2 = (86+84)/2 = 85$ (рис. 8б). Причем это отличие может быть существенным, если разница между минимальным и максимальным долями совпадений окажется больше. Например, если мы примем количество общих слов в языках A и C (N_{AC}) равным 80, то среднее и минимальная доли совпадений будут отличаться уже на 3 слова: $N_{ср} = (86+80)/2 = 83$.

Несмотря на справедливость доводов, приводимых в пользу выбора минимального значения, использование такого подхода в предложенном виде трудно признать оправданным. Как уже говорилось выше, расхождения в процентах совпадений между объединяемыми языками или группами языков могут быть вызваны не только вторичным сближением между ними[19], но и — в значительно большей степени — самим случайным характером процесса лексических замен, в результате которого в лексике двух родственных языков за один и тот же выбранный промежуток времени может измениться разное

[18] Там же: 164.

[19] Которое проявляется в невыявленных поздних заимствованиях, завышающих процент совпадений при сравнении списков.

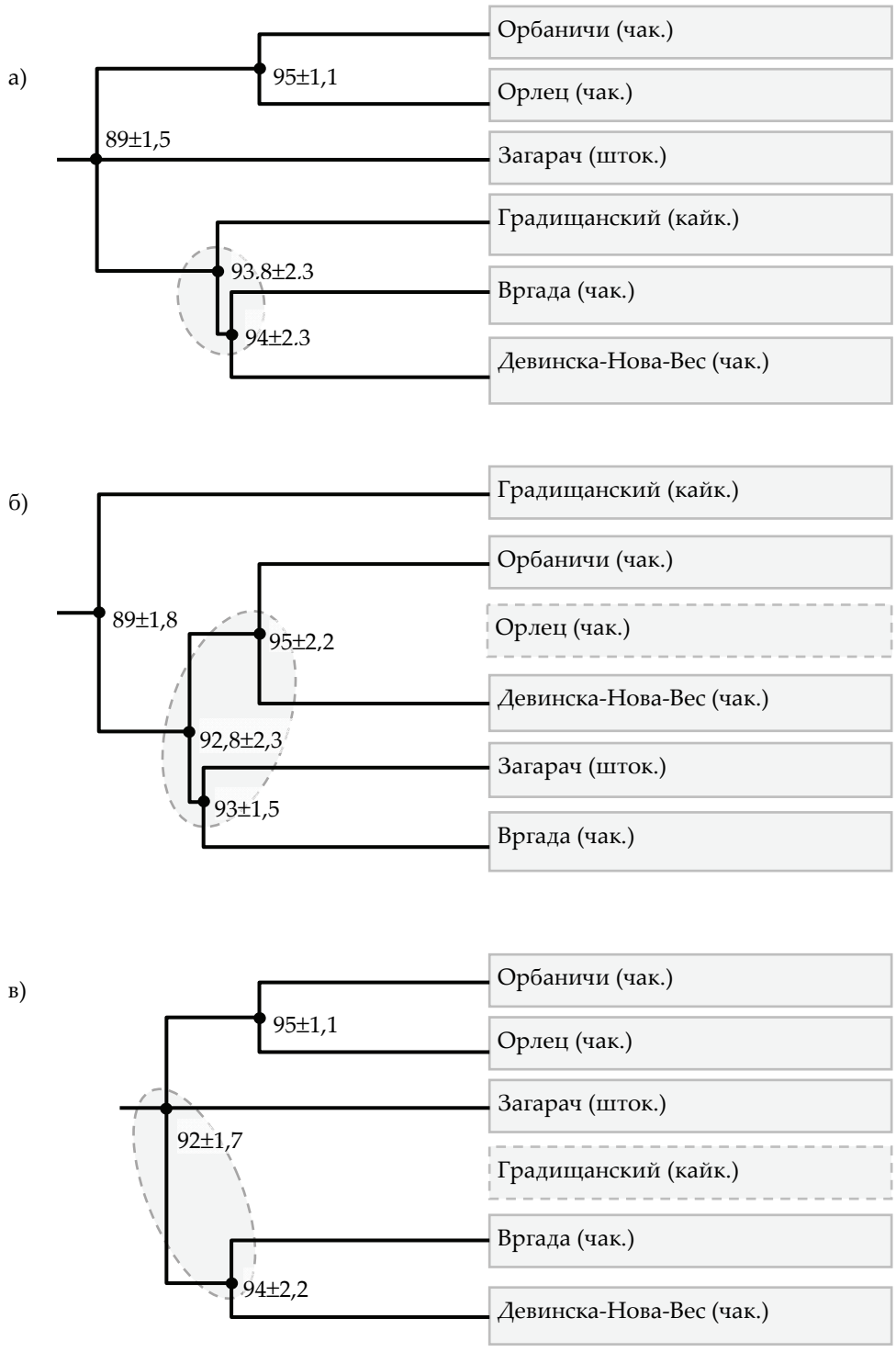


Рисунок 9. Древо, полученное для трех разных выборок сербохорватских идиомов:
а) полный список языков;
б) после исключения чакавского говора д. Орлец;
в) после исключения градищанского кайкавского.
Доли совпадений для узлов рассчитаны по наименьшим значениям.

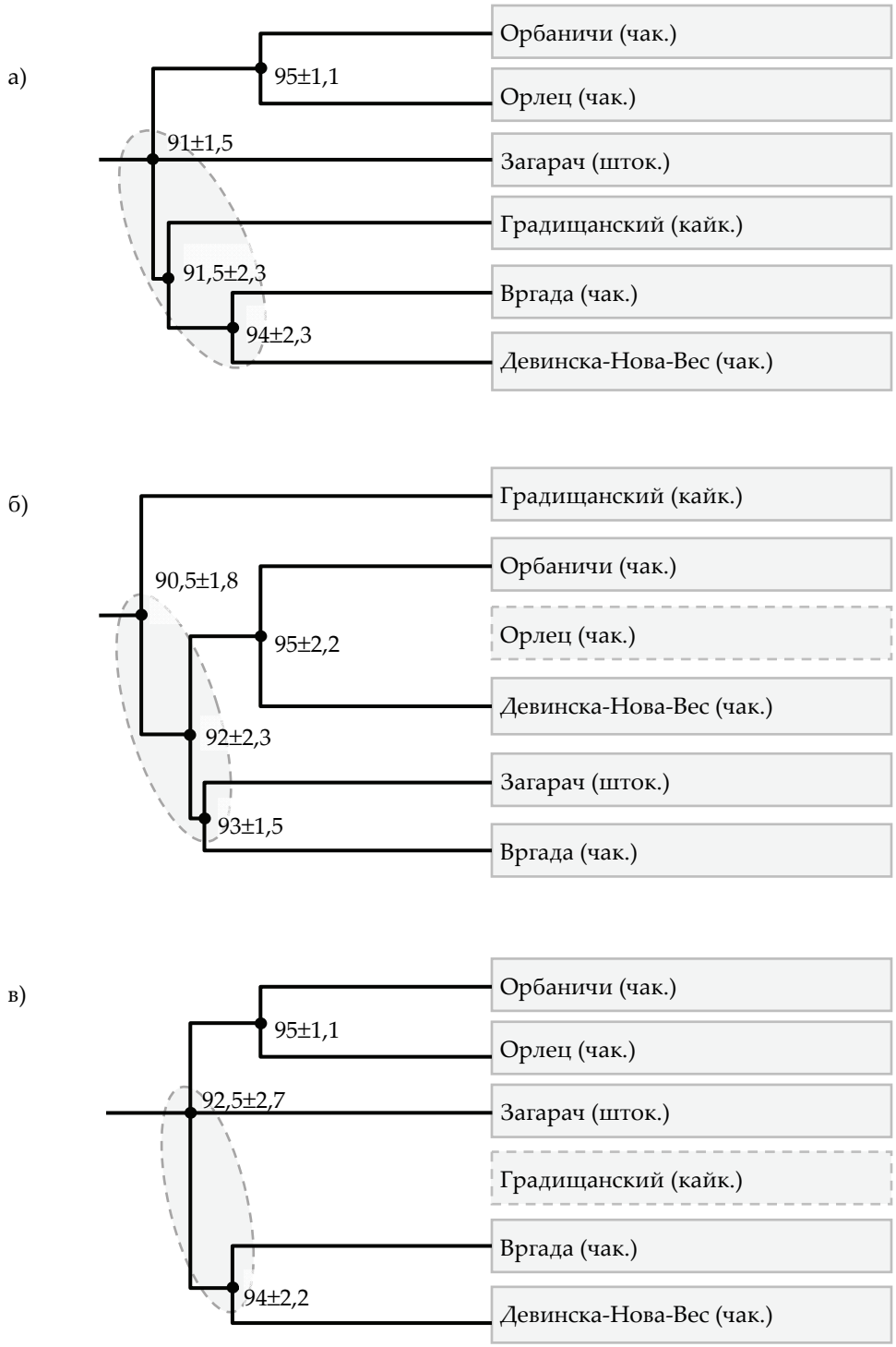


Рисунок 10. Древо, полученное, полученная для трех выборок сербохорватских идиомов после пересчета долей совпадений по средним значениям:

а) полный список языков;

б) после исключения чакавского говора д. Орлец;

в) после исключения градищанского кайкавского.

количество значений. Подчеркнем, что данная неравномерность (в отличие от вторичных сближений) не зависит от условий дивергенции языков и привносит неизбежную погрешность в *любые* лексикостатистические расчеты, как при построении деревьев, так и при расчете глоттохронологических датировок[20]. Другими словами, расхождение между долями совпадений в рамках этих погрешностей не является аномалией само по себе, а лишь отражает случайную природу лексического процесса и в большинстве случаев не требует корректировки. Поэтому попытка устранения расхождений путем отбрасывания больших значений на практике приводит к систематическому искажению исходных данных[21], в результате чего мы получаем заведомо заниженные проценты совпадений для абсолютного большинства языков, имеющих более 70% общей лексики[22].

Посмотрим, насколько существенным оказалось это искажение в случае с классификацией сербохорватской группы. Для этого пересчитаем все доли совпадений по средним значениям (рис. 10) и сравним их с рассмотренными ранее.

Прежде всего отметим, что группировка идиомов во всех трех деревьях осталась прежней. В то же время, как и ожидалось, переход к средним значениям привел к увеличению долей совпадений в основании деревьев, что отразилось в заметном сокращении расстояний между узлами. Так, в первом фрагменте с полным набором идиомов (рис. 10а), разрыв между первым и вторым узлом сократился с 5% до 0,5%, что фактически означает их полное совпадение. Во втором и третьем случаях (рис. 10б, в) это расстояние уменьшилось соответственно с 4% до 1,5% и с 2% до 1,5%, в результате чего разница между узлами оказалась в переделах статистической погрешности. Благодаря этим, на первый взгляд, несущественным изменениям, после объединения перекрывающихся узлов (обведены пунктиром) конфигурация двух деревьев (рис.10а и 10в) стала полностью идентичной, а третьего (с исключенным говором д. Орлец, рис. 10б) — очень близкой к ним[23]. Таким образом, переход к средним значениям при расчете долей совпадений, а также последующее устранение незначимых узлов древа позволили добиться топологической стабильности и прозрачности древа во всех трех случаях, выявленных нами в ходе анализа.

Вернемся теперь к исходной классификации (рис. 2) и повторим обе вышеописанные процедуры (пересчет долей совпадений по средним значениям и объединение перекрывающихся узлов) для полного генеалогического древа 25 славянских идиомов. Результаты вычисления долей совпадений по средним значениям приведены на рис. 11. Сравнение полученного древа с рис. 5 наглядно демонстрирует, насколько существенным

[20] Количественная оценка этой неравномерности для разных временных интервалов дана в статье Васильев, Саенко 2016: 272–275, а также Васильев, Саенко 2017: 128–133. Как показывают проведенные расчеты и результаты моделирования, представленные в статье, именно этот вероятностный характер процесса замен имеет определяющее значение для точности лексикостатистических расчетов.

[21] Здесь нужно добавить, что в подобной ситуации неопределенности (т. е. когда невозможно установить, какие из данных достоверны, а какие — искажены), в статистике принято использовать именно среднее значение. Применительно к нашему случаю это означает, что, если мы не можем установить факт влияния внешних факторов (будь то повторное сближение или согласованные изменения в лексике языков), то любые отклонения следует считать статистическими и, следовательно, использовать средние доли совпадений, так как замедление или ускорение процесса замен равновероятно. Предлагаемый же подход, очевидно носит не статистический, а детерминированный характер.

[22] Просматривая Таблицу 3, нетрудно убедиться, что к ним относятся почти все рассматриваемые славянские идиомы.

[23] Отделение Девинска-Нова-Вес от Вргады и присоединение к говору д. Орбаничи объясняется тем, что в отсутствие орлецкого говора, они становятся ближайшими родственниками (95% совпадений) среди оставшихся идиомов, и поэтому связываются первыми.

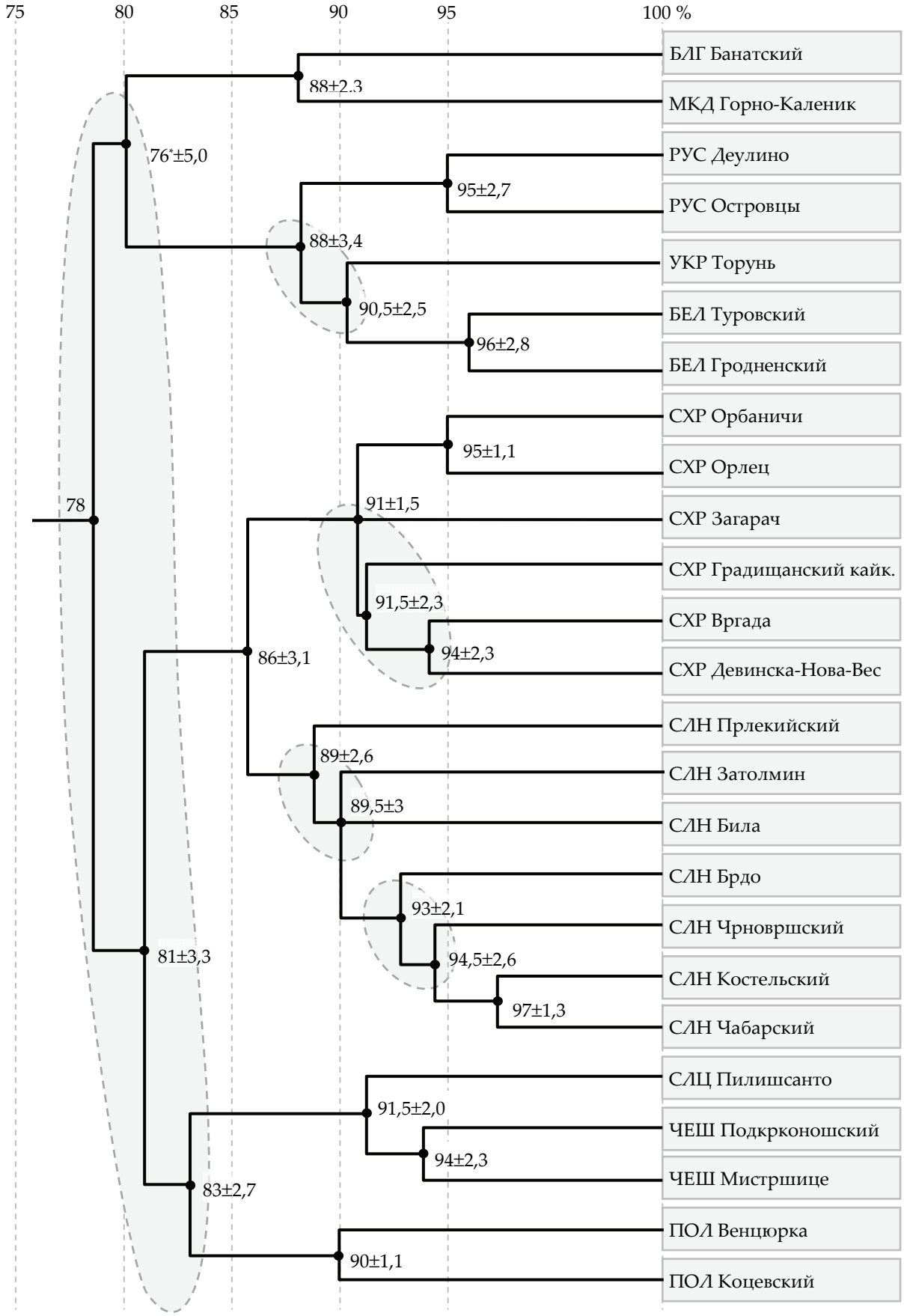


Рисунок 11. Генеалогическое древо 25 славянских идиомов с процентами совпадений, рассчитанными по средним значениям. Узлы с взаимно перекрывающимися диапазонами средних абсолютных отклонений обведены пунктиром.

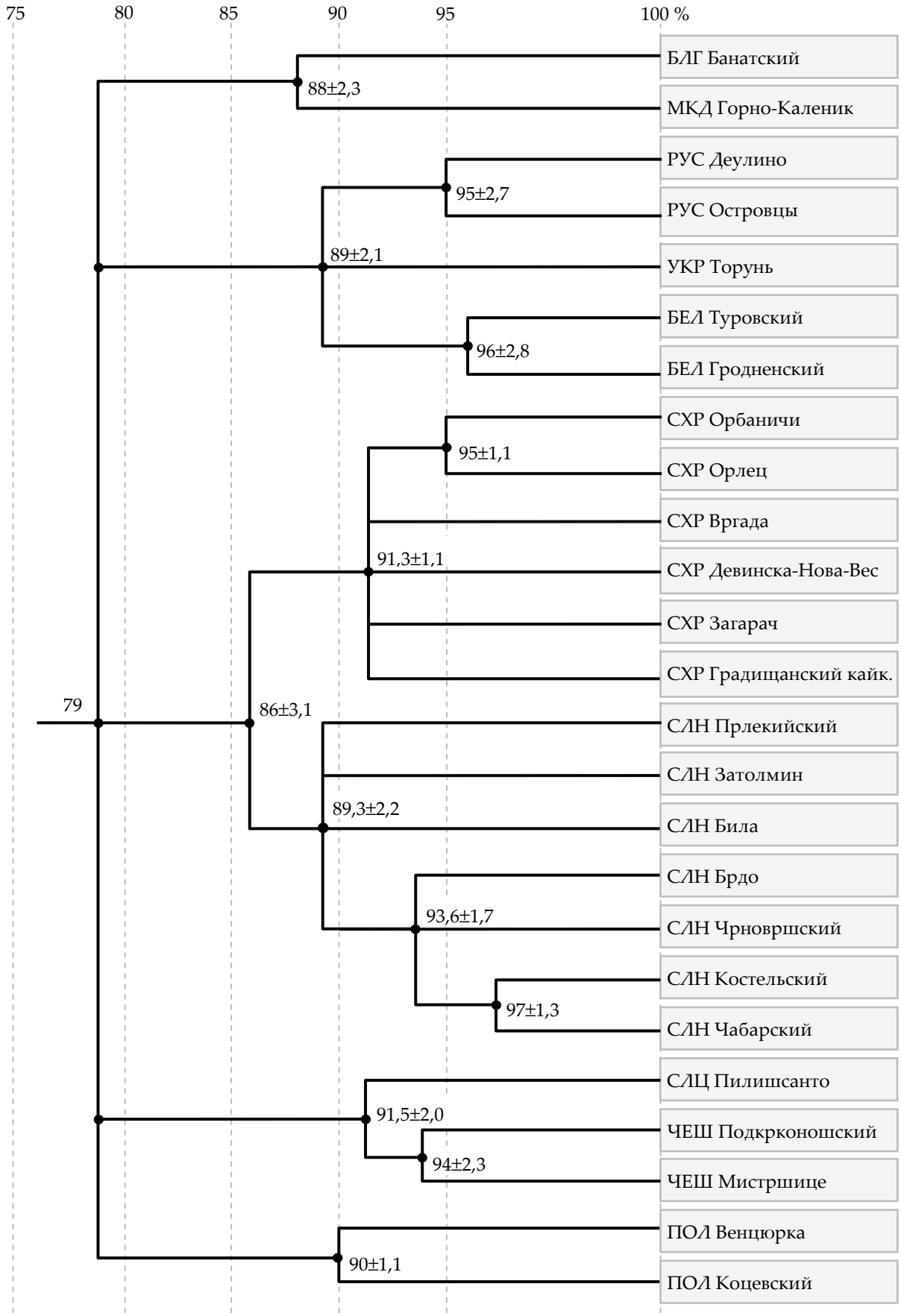


Рисунок 12. Генеалогическое древо 25 славянских идиомов с процентами совпадений, рассчитанными по средним значениям, после объединения взаимно перекрывающихся узлов.

может быть отличие между наименьшими и средними процентами совпадений. Так значение узла, объединяющего словенскую и сербохорватскую ветви, увеличилось на 6% (с 80 до 86%), а узла восточнославянских языков — на 4% (с 84 до 88%). При этом доли совпадений некоторых узлов после пересчета стали очень близкими: например, словенские прлекийский, затолминский и Била или рассмотренные выше сербохорватские идиомы. Наконец, в одном случае изменения в долях совпадений потребовали пересмотра самой структуры древа: в результате увеличения значения корневого узла с 73,6 до 78% он оказался правее узла, связывающего восточнославянскую и болгаромакедонскую ветви со средним процентом совпадений 76%, что противоречит найденной топологии[24].

Если мы перейдем к итоговому виду классификации (рис. 12), объединив узлы с перекрывающимися отклонениями в соответствии с предложенной методикой, то увидим, что различия между деревьями (ср. с рис. 7) станут еще более заметными. Вследствие уменьшения расстояний между узлами некоторые из них оказалась статистическим неразличимыми, что позволило заменить их одним общим узлом, объединяющим сразу три (например, восточнославянская подгруппа), четыре (словенская и сербохорватская подгруппы) или даже пять ветвей (а также корневой узел древа). Важно отметить, что благодаря переходу к средним значениям нам удалось не только снизить «бинарность» топологии, вызванную несовершенством процедуры объединения таксонов, но и упростить содержательную интерпретацию древа[25].

Если сравнивать полученное древо с исходным (рис. 2), то можно отметить целый ряд положительных изменений в итоговой топологии. Узлы перестали быть строго бинарными, что приблизило древо ко классификациям, построенным традиционными методами. Другим важным преимуществом стало отсутствие «лесенки» в сербохорватской и словенской ветвях: 110-словных списков явно недостаточно для построения точной классификации столь близких диалектов. Нельзя не отметить, что те говоры, которые все же объединены в бинарные узлы (Орбаничи и Орлец, костельский и чабарский), действительно очень близки друг к другу как в географическом, так и генетическом отношении.

Исчезли фантомные корневые объединения, вызванные тем, что алгоритм Starling вынуждает всё сводить к бинарному виду. На первый взгляд, некоторую проблему может представлять то, что по сравнению с первым древом западнославянская ветвь распалась на две части: польскую и чешско-словацкую. Однако имеется не так много древних фонетических инноваций, объединяющих западнославянские языки и противопоставляющих их всем остальным. В первую очередь это аффрикативизация *ť > *c и *ď > *ʒ, а также переход *x по второй и третьей палатализациям в *š, а не *s'[26]. В то же время, допустим, рефлексация сочетаний вида TorT и TolT сближает чешско-словацкую под-

[24] Данный пример свидетельствует о том, что использование минимальных значений вместо средних приводит не только к систематическому занижению долей совпадений, но влияет также на саму последовательность объединения таксонов, а следовательно — непосредственно сказывается на полученной конфигурации древа.

[25] Прежде всего это проявляется в уменьшении количества незначимых узлов, которые невозможно сопоставить каким-либо фактическим или предполагаемым событиям в истории развития языков.

[26] Часто в качестве западнославянских черт рассматривают сохранение групп *tl и *dl, а также *kw и *gw перед *ě и *i. Однако эти особенности не являются эксклюзивно западнославянскими и, что важнее, представляют из себя архаизмы, а не инновации, что серьезно снижает их ценность для генеалогической классификации.

группу с южнославянскими языками, а не лехитскими и лужицкими. Таким образом, западнославянскую подгруппу (как и южнославянскую) нельзя считать таксоном с доказанным статусом.

Фактически, именно отсутствием объединения болгаро-македонского и сербохорватско-словенского таксонов в южнославянскую подгруппу, а чешско-словацкого и лехитского — в западнославянскую, полученное нами древо и отличается от общепринятой классификации славянских языков, представленной в (Иванов 1990: 95).

Оговорим, что полученная нами схема не может служить в качестве аргумента против необходимости выделения южно- и западнославянской подгрупп, поскольку она базируется на неполном материале: отсутствуют данные лужицких, кашубского, словинского и полабского языков, а также торлакского наречия. Кроме того, как уже отмечалось ранее, даже выполнение условия о взаимном перекрытии соседних узлов не обязывает нас к их объединению, а лишь указывает на такую возможность. Поэтому при наличии дополнительных доводов в пользу сохранения западнославянской общности, она может быть выделена на итоговом древе.

В завершение ещё раз напомним, что предложенная методика не является самостоятельным методом классификации и не формирует её структуру, а применяется к *уже построенному* генеалогическому древу и позволяет оценить достоверность его топологии на основе статистических расчетов. С одной стороны, это несколько ограничивает ее возможности[27], но с другой — делает ее универсальной и дает возможность использовать ее для анализа любых лексикостатистических классификаций, вне зависимости от способа их построения.

IV. Заключение

Подведем итоги нашего исследования, сформулировав основные полученные теоретические выводы и практические результаты:

a. Накопленный опыт лексикостатистической классификации языков с помощью системы Starling свидетельствует о том, что, несмотря на общую правдоподобность и полезность получаемых результатов, в построенных генеалогических деревьях обнаруживаются внутренние несоответствия и артефакты, не поддающиеся объяснению или противоречащие известным данным.
b. Наиболее распространенными из этих несоответствий являются:

- *проблема вариативности* — неустойчивость конфигурации древа при изменении количества или состава идиомов;
- *проблема избыточной кластеризации* — древо содержит большое количество близкорасположенных узлов, интерпретация которых проблематична или невозможна.

c. Основная причина возникновения указанных проблем заключается в несовершенстве методики построения древа, реализованной в Starling, которая, с одной стороны, игнорирует вероятностный характер лексических расчетов, а с другой — привносит фиксированные поправки в исходные данные, что приводит к появлению статистических погрешностей, а также систематических ошибок в полученной классификации.

[27] Так как она, очевидным образом, не может изменить порядок объединения узлов, хотя и указывает на необходимость такого изменения (как в случае с болгаро-македонским/восточнославянским узлом на рис. 11).

d. В работе предложена методика анализа классификации, позволяющая количественно оценить возникающие погрешности на основе величины среднего абсолютного отклонения, а также минимизировать их влияние на строение древа с помощью процедуры устранения недостоверных узлов.
e. Применение методики для анализа генеалогического древа 25 славянских идиомов показало, что она позволяет эффективно решать обе выявленные проблемы, связанные с нестабильностью топологии и избыточной кластеризацией древа, а также значительно улучшить полученную классификацию древа с точки зрения ее содержательной интерпретации.
f. Универсальность разработанной методики дает возможность применять ее для анализа любых генеалогических деревьев, полученных на основе лексикостатистических расчетов, независимо от того, каким способом они были построены.
g. Все основные процедуры методики хорошо формализуются и могут быть реализованы в виде дополнительного модуля Starling, или самостоятельной утилиты, удобной для практического использования при анализе и уточнении лексикостатистических классификаций.

Литература

Бурлак, С. А., С. А. Старостин. 2005. *Сравнительно-историческое языкознание*. Москва: Академия.

Васильев, М. Е. 2010. Об использовании лексического критерия для построения генеалогической классификации. *Бюллетень Общества востоковедов РАН* 17: 530–572.

Васильев, М. Е., А. И. Коган. 2013. К вопросу о восточнодардской языковой общности. *Вопросы языкового родства* 10: 149–177.

Васильев, М. Е., М. Н. Саенко. 2016. К вопросу о точности глоттохронологии: датирование процесса лексических замен по данным романских языков. *Вопросы языкового родства* 14/3–4: 259–277.

Васильев, М. Е., М. Н. Саенко. 2017. К вопросу о точности глоттохронологии: датирование языковой дивергенции по данным романских языков. *Вопросы языкового родства* 15/1–2: 114–135.

Грунтов, И. А., О. М. Мазо. 2015. Классификация монгольских языков по лексикостатистическим данным. *Вопросы языкового родства* 13/3–4: 205–255.

Иванов Вяч. Вс. 1990. Генеалогическая классификация языков. В: *Лингвистический энциклопедический словарь*. М.: Советская энциклопедия: 93–98.

Николаев, С. Л., М. Н. Толстая. 2001. *Словарь карпатоукраинского торуньского говора*. М.: Институт славяноведения РАН.

Поздняков, К. И. 2014. О пороге родства и индексе стабильности в базисной лексике при массовом сравнении: атлантические языки. *Вопросы языкового родства* 11: 187–225.

ССРНГ 1969 = *Словарь современного русского народного говора (д. Деулино Рязанского района Рязанской области)*. Москва: Наука

Стойков, С. 1968. *Лексиката на банатския говор*. София: Издателство на Българската академия на науките.

Стойков, С. 2002. *Българска диалектология*. София: Проф. Марин Дринов.

Сцяшковіч, Т. Ф. 1972. *Матэрыялы да слоўніка Гродзенскай вобласці*. Мінск: Навука і тэхніка.

Сцяшковіч, Т. Ф. 1983. *Слоўнік Гродзенскай вобласці*. Мінск: Навука і тэхніка.

ТС = *Тураўскі слоўнік*. 1982–1987. Тамы 1–5. Мінск: Навука и тэхніка.

Ћупић Д., Ћупић Ж. 1997. *Речник говора Загарача*. Београд: Српска академија наука и уметности, Институт за српски језик САНУ.

Хонселаар, З. 2001. *Говор деревни Островцы Псковской области*. Amsterdam: Rodopi.

References

Bachmannová, Jarmila. 2016. *Slovník podkrkonošského nářečí*. Praha: Academia.

Blažek, Václav. Klasifikace slovanských jazyků. In: Petr Karlík, Marek Nekula, Jana Pleskalová (eds.). *CzechEncy — Nový encyklopedický slovník češtiny*. URL: https://www.czechency.org/slovnik/KLASIFIKACE SLOVANSKÝCH JAZYKŮ.

Burlak, Svetlana A., Sergey A. Starostin. 2005. *Sravnitel'no-istoricheskoje jazykoznanije*. Moscow: Academia.

Čujec Stres, Helena. 2011–2014. *Slovar zatolminskega govora*. Dela 1–2. Zatolmin: Stres inženiring.

Ćupić, Drago, Željko Ćupić. 1997. *Rečnik govora Zagarača*. Beograd: Srpska akademija nauka i umetnosti, Institut za srpski jezik SANU.

Gregor, Ferenc. 1975. *Der slowakische Dialekt von Pilisszántó*. Budapest: Akadémiai Kiadó.

Gregorič, Jože. 2014. *Kostelski slovar*. Ljubljana: Založba ZRC.

Gruntov, Ilya A., Olga M. Mazo. 2015. Klassifikacija mongol'skix jazykov po leksikostatisticheskim dannym. *Journal of Language Relationship* 13/3–4: 205–255.

Hill, Peter. 1991. *The Dialect of Gorno Kalenik*. Columbus: Slavica Pub.

Honselaar, Zep. 2001. *Govor derevni Ostrovtsy Pskovskoy oblasti*. Amsterdam: Rodopi.

Houtzagers, Peter. 1985. *The Čakavian dialect of Orlec and the islands of Cres*. Amsterdam: Brill.

Houtzagers, Peter. 1999. *The Kajkavian Dialect of Hidegség and Fertőhomok*. Amsterdam: Rodopi.

Ivanov, Vyacheslav V. 1990. Genealogicheskaja klassifikacija jazykov. In: *Lingvisticheskij enciklopedicheskij slovar'*: 93–98. Moscow: Sovetskaja enciklopedija.

Jurišić, Blaž. 1973. *Rječnik govora otoka Vrgade*. Zagreb: Izdavački zavod Jugoslavenske akademije; GZH — ZRINSKI.

Kalsbeek, Janneke. 1998. *The Čakavian Dialect of Orbanići near Žminj in Istria*. Amsterdam — Atlanta: Rodopi.

Kogan, Anton I. 2016. Genealogical classification of New Indo-Aryan languages and lexicostatistics. *Journal of Language Relationship* 14/3–4: 227–258.

Kucała, Marian. 1957. *Porównawczy słownik trzech wsi małopolskich*. Wrocław: Zakład im. Ossolińskich — Wydawnictwo PAN.

Malina, Ignát. 1946. *Slovník nářečí mistřického*. Praha: Česká akademie věd a umění.

Malnar, Slavko. 2008. *Rječnik govora čabarskog kraja*. Čabar: Matica hrvatska — Ogranak u Čabru.

Nikolaev, Sergey L., Marfa N. Tolstaja. 2001. *Slovar' karpatoukrainskogo torun'skogo govora*. Moscow: Institut slavianovedenija RAN.

Pozdniakov, Konstantin I. 2014. O poroge rodstva i indekse stabil'nosti v bazisnoj leksike pri massovom sravnenii: atlanticheskije jazyki. *Journal of Language Relationship* 11: 187–225.

Pronk, Tijmen. 2009. *The Slovene Dialect of Egg and Potschach in the Gailtal, Austria*. Amsterdam — New York: Rodopi.

Rajh, Bernard. 2010. *Gúčati po antùjoško. Gradivo za narečni slovar severozahodnoprleškega govora*. Maribor: Filozofska fakulteta — Mednarodna založba Oddelka za slovanske jezike in književnosti.

Saitou, Naruya, Masatoshi Nei. 1987. The Neighbor-joining Method: A New Method for Reconstructing Phylogenetic Trees. *Molecular Biology and Evolution* 4 (4): 406–425.

Sciaškovič, Tatjana F. 1972. *Materyjały da słoŭnika Hrodzienskaj vobłasci*. Minsk: Navuka i technika.

Sciaškovič, Tatjana F. 1983. *Słoŭnik Hrodzienskaj vobłasci*. Minsk: Navuka i technika.

SSRNG 1969 = Ossovetski I.A. (ed.). *Slovar' sovremennogo russkogo narodnogo govora (d. Deulino Riazanskogo rajona Riazanskoy oblasti)*. Moskva: Nauka.

Starostin, George. 2010. Preliminary lexicostatistics as a basis for language classification: A new approach. *Journal of Language Relationship* 3: 79–116.

Steenwijk, Han. 1992. *The Slovene dialect of Resia: San Giorgo*. Amsterdam — Atlanta: Rodopi.

Stojkov, Stojko. 1968. *Leksikata na banatskija govor*. Sofija: Izdatelstvo na Bălgarskata akademija na naukite.

Stojkov, Stojko. 2002. *Bălgarska dialektologija*. Sofija: Prof. Marin Drinov.

Sychta, Bernard. 1980–1985. *Słownictwo kociewskie*. Tomy 1–3. Wrocław — Warszawa — Kraków — Gdańsk: Zakład narodowy im. Ossolińskich — Wydawnictwo PAN.

Tominec, Ivan. 1964. *Črnovrški dialect*. Ljubljana: Slovenska akademija znanosti in umetnosti.

TS = Kryvitski, A. A. (ed.). 1982–1987. *Turaŭski słoŭnik*. Tamy 1–5. Minsk: Navuka i technika.

Vasilyev, Mikhail E. 2010. Ob ispol'zovanii leksicheskogo kriterija dl'a postrojenija genealogicheskoj klassifikacii. *B'ulleten' Obshchestva vostokovedov RAN* 17: 530–572.

Vasilyev, Mikhail E., Anton I. Kogan. 2013. K voprosu o vostochnodardskoj jazykovoj obshchnosti. *Journal of Language Relationship* 10: 149–177.

Vasilyev, Mikhail E., Mikhail N. Saenko. 2016. K voprosu o tochnosti glottoxronologii: datirovanije processa leksicheskix zamen po dannym romanskix jazykov. *Journal of Language Relationship* 14/3–4: 259–277.

Vasilyev, Mikhail E., Mikhail N. Saenko. 2017. K voprosu o tochnosti glottoxronologii: datirovanije jazykovoj divergencii po dannym romanskix jazykov. *Journal of Language Relationship* 15/1–2: 114–135.
Vážný, Václav. 1927. *Čakavské nářečí v slovenském Podunají*. Bratislava: Filosofická fakulta University Komenského.
Vydrin, Valentin. 2009. On the Problem of the Proto-Mande Homeland. *Journal of Language Relationship* 1: 107–142.

Mikhail Vasilyev, Mikhail Saenko. An analysis of the topology and estimation of accuracy for lexicostatistical classifications (on the data of Slavic languages)

Today, lexicostatistical methods are widely used in comparative-historical linguistics to establish linguistic kinship and build genealogical classifications. In works by Russian comparative linguists the most common technique is construction of phylogenetic trees obtained with the aid of the Starling software, developed by Sergei Starostin at the end of the 20th century. Starostin's algorithm was based on a modified method of "neighbor joining" and yielded satisfactory or plausible results in the vast majority of cases. At the same time, many researchers have pointed out a number of significant shortcomings in the obtained classifications, the most serious of which are the instability of the tree caused by even minimal changes in the number of idioms, as well as detection of a large number of fictitious taxa and nodes that are poorly explained or even contradict existing concepts. This article provides a detailed examination of these shortcomings based on the example of a new lexicostatistical classification for 25 Slavic lects. Upon detailed analysis, we propose a special procedure that allows to minimize the negative effect of identified deficiencies on the structure of the tree, making use of statistical analysis of the resulting topology and capable of identifying unreliable nodes within it. The technique is simple enough to to be practically implemented in the form of an additional Starling component or a separate application.

Keywords: lexicostatistics; neighbor-joining method; genealogical classification; mean absolute deviation.

Приложение

	BAN	GKM	ZAG	ORB	ORL	VRG	DNV	BRG	CVS	KOS	ZTL	RES	GLT	PRL	CAB	PKC	MIS	PLS	WLP	KGP	TUR	GRD	TOR	DEU	OST
BAN	1	0.88	0.8	0.76	0.77	0.77	0.77	0.76	0.73	0.77	0.68	0.72	0.75	0.73	0.76	0.74	0.72	0.71	0.67	0.68	0.72	0.73	0.76	0.73	0.75
GKM	0.88	1	0.84	0.79	0.79	0.8	0.8	0.78	0.74	0.77	0.71	0.74	0.73	0.72	0.75	0.77	0.74	0.72	0.68	0.68	0.77	0.79	0.79	0.77	0.77
ZAG	0.8	0.84	1	0.89	0.89	0.93	0.92	0.89	0.84	0.88	0.8	0.83	0.85	0.83	0.88	0.82	0.8	0.8	0.76	0.78	0.81	0.82	0.85	0.79	0.8
ORB	0.76	0.79	0.89	1	0.95	0.93	0.95	0.9	0.87	0.91	0.8	0.84	0.85	0.86	0.9	0.83	0.82	0.83	0.79	0.81	0.79	0.82	0.83	0.76	0.8
ORL	0.77	0.79	0.89	0.95	1	0.92	0.93	0.89	0.86	0.89	0.81	0.85	0.85	0.84	0.88	0.83	0.82	0.82	0.78	0.77	0.78	0.81	0.82	0.75	0.78
VRG	0.77	0.8	0.93	0.93	0.92	1	0.94	0.89	0.87	0.91	0.8	0.84	0.85	0.84	0.89	0.81	0.81	0.82	0.78	0.79	0.8	0.83	0.85	0.77	0.81
DNV	0.77	0.8	0.92	0.95	0.93	0.94	1	0.94	0.9	0.94	0.84	0.89	0.89	0.88	0.93	0.86	0.84	0.83	0.81	0.82	0.81	0.84	0.85	0.77	0.8
BRG	0.76	0.78	0.89	0.9	0.89	0.89	0.94	1	0.87	0.88	0.8	0.87	0.86	0.86	0.87	0.82	0.82	0.8	0.78	0.78	0.78	0.81	0.83	0.76	0.81
CVS	0.73	0.74	0.84	0.87	0.86	0.87	0.9	0.87	1	0.94	0.94	0.9	0.94	0.91	0.95	0.83	0.79	0.79	0.79	0.8	0.76	0.77	0.77	0.76	0.8
KOS	0.77	0.77	0.88	0.91	0.89	0.91	0.94	0.88	0.94	1	0.88	0.89	0.91	0.91	0.97	0.86	0.85	0.84	0.81	0.83	0.8	0.82	0.82	0.76	0.79
ZTL	0.68	0.71	0.8	0.8	0.81	0.8	0.84	0.8	0.94	0.88	1	0.87	0.88	0.87	0.91	0.79	0.75	0.73	0.76	0.75	0.73	0.75	0.74	0.73	0.77
RES	0.72	0.74	0.83	0.84	0.85	0.84	0.89	0.87	0.9	0.89	0.87	1	0.89	0.86	0.89	0.82	0.79	0.78	0.76	0.79	0.79	0.8	0.81	0.75	0.81
GLT	0.75	0.73	0.85	0.85	0.85	0.85	0.89	0.86	0.94	0.91	0.88	0.89	1	0.89	0.93	0.84	0.81	0.82	0.79	0.8	0.75	0.77	0.79	0.75	0.8
PRL	0.73	0.72	0.83	0.86	0.84	0.84	0.88	0.86	0.91	0.91	0.87	0.86	0.89	1	0.9	0.83	0.85	0.82	0.82	0.82	0.78	0.8	0.81	0.76	0.79
CAB	0.76	0.75	0.88	0.9	0.88	0.89	0.93	0.87	0.95	0.97	0.91	0.89	0.93	0.9	1	0.86	0.83	0.82	0.8	0.81	0.78	0.79	0.79	0.76	0.79
PKC	0.74	0.77	0.82	0.83	0.83	0.81	0.86	0.82	0.83	0.86	0.79	0.82	0.84	0.83	0.86	1	0.94	0.9	0.8	0.81	0.78	0.81	0.82	0.76	0.79
MIS	0.72	0.74	0.8	0.82	0.82	0.81	0.84	0.82	0.79	0.85	0.75	0.79	0.81	0.85	0.83	0.94	1	0.93	0.84	0.85	0.77	0.83	0.83	0.74	0.75
PLS	0.71	0.72	0.8	0.83	0.82	0.82	0.83	0.8	0.79	0.84	0.73	0.78	0.82	0.82	0.82	0.9	0.93	1	0.83	0.85	0.79	0.86	0.83	0.78	0.78
WLP	0.67	0.68	0.76	0.79	0.78	0.78	0.81	0.78	0.79	0.81	0.76	0.76	0.79	0.82	0.8	0.8	0.84	0.83	1	0.9	0.78	0.84	0.84	0.77	0.77
KGP	0.68	0.68	0.78	0.81	0.77	0.79	0.82	0.78	0.8	0.83	0.75	0.79	0.8	0.82	0.81	0.81	0.85	0.85	0.9	1	0.78	0.83	0.83	0.75	0.77
TUR	0.72	0.77	0.81	0.79	0.78	0.8	0.81	0.78	0.76	0.8	0.73	0.79	0.75	0.78	0.78	0.78	0.77	0.79	0.78	0.78	1	0.96	0.89	0.88	0.89
GRD	0.73	0.79	0.82	0.82	0.81	0.83	0.84	0.81	0.77	0.82	0.75	0.8	0.77	0.8	0.79	0.81	0.83	0.86	0.84	0.83	0.96	1	0.92	0.9	0.91
TOR	0.76	0.79	0.85	0.83	0.82	0.85	0.85	0.83	0.77	0.82	0.74	0.81	0.79	0.81	0.79	0.82	0.83	0.83	0.84	0.83	0.89	0.92	1	0.84	0.87
DEU	0.73	0.77	0.79	0.76	0.75	0.77	0.77	0.76	0.76	0.76	0.73	0.75	0.75	0.76	0.76	0.76	0.74	0.78	0.77	0.75	0.88	0.9	0.84	1	0.95
OST	0.75	0.77	0.8	0.8	0.78	0.81	0.8	0.81	0.8	0.79	0.77	0.81	0.8	0.79	0.79	0.79	0.75	0.78	0.77	0.77	0.89	0.91	0.87	0.95	1

Таблица 3. Исходная таблица долей совпадений между 110-словными списками 25 славянских идиомов.

Условные обозначения: BAN — Банатский болгарский; GKM — Македонский говор д. Горно-Каленик; ZAG — Штокавский сербохорватский племени Загарач; ORB — Чакавский сербохорватский говор д. Орбаничи; ORL — Чакавский сербохорватский говор д. Орлец и острова Црес; VRG — Чакавский сербохорватский говор острова Вргада; DNV — Чакавский сербохорватский говор д. Девинска-Нова-Вес; BRG — Градищанский кайкавский сербохорватский; CVS — Чрноврш-ский словенский; KOS — Костельский словенский; ZTL — Затолминский словенский; RES — Резьянский словенский д. Била.; GLT — Словенский д. Брдо; PRL — Прлешский словенский; CAB — Словенский говор Чабара и окрестностей; PKC — Подкрконошский чешский; MIS — Моравский чешский д. Мистршице; PLS — Словацкий говор д. Пилишсанто; WLP — Малопольский диалект д. Венцюрка; KGP — Коцевский великопольский; TUR — Белорусские говоры Турова и окрест-ностей; GRD — Белорусские говоры Гродненской области; TOR — Украинский говор д. Торунь; DEU — Русский говор д. Деулино; OST — Русский говор д. Островцы