Dialectal layers in West Iranian: A hierarchical dirichlet process approach to linguistic relationships

Cathcart, Chundra A

DOI: https://doi.org/10.1111/1467-968x.12225

Transactions of the Philological Society Volume 120:1 (2022) 1–31 doi: 10.1111/1467-968X.12225

DIALECTAL LAYERS IN WEST IRANIAN: A HIERARCHICAL DIRICHLET PROCESS APPROACH TO LINGUISTIC RELATIONSHIPS[1]

By Chundra A. Cathcart
University of Zurich

(Submitted: 31 December, 2019; Accepted: 22 September, 2021)

Abstract

This paper addresses a series of complex and unresolved issues in the historical phonology of West Iranian languages, (Persian, Kurdish, Balochi, and other languages), which display a high degree of irregular, non-Lautgesetzlich behaviour. Most of this irregularity is undoubtedly due to language contact; I argue, however, that an oversimplified view of the processes at work has prevailed in the literature on West Iranian dialectology, with specialists assuming that deviations from an expected outcome in a given non-Persian language are due to lexical borrowing from some chronological stage of Persian. It is demonstrated that this qualitative approach yields at times problematic conclusions stemming from the lack of explicit probabilistic inferences regarding the distribution of the data: Persian may not be the sole donor language; additionally, borrowing at the lexical level is not always the mechanism that introduces irregularity. In many cases, the possibility that West Iranian languages show different reflexes in different conditioning environments remains under-explored. I employ a novel Bayesian approach designed to overcome these problems and tease apart the different determinants of irregularity in patterns of West Iranian sound change. This methodology helps to provisionally resolve a number of outstanding questions in the literature on West Iranian dialectology concerning the dialectal affiliation of certain sound changes. I outline future directions for work of this sort.

Dieser Artikel befasst sich mit einigen komplexen Problemen der historischen Phonologie der westiranischen Sprachen (Persisch, Kurdisch, Belutschi usw.). Diese Sprachen weisen eine Vielzahl von nichtlautgesetzlichen Reflexen auf, die sich entgegen einer häufig in der Literatur zu findenden Annahme nicht immer als Resultat einer Übernahme von Lehnwörtern aus dem Persischen erklären lassen. Es kann nämlich nicht ausgeschlossen werden, dass auch Lautwandel selbst entlehnt wurden bzw. dass sich einzelne vermeintlich irreguläre Lautwandel bei genauerem Hinsehen als lautgesetzlich erweisen. In diesem Beitrag stelle ich daher eine probabilistische Methode vor, die es erlaubt, die Verteilung der irregulären Lautwandelreflexe in den westiranischen Sprachen zu erklären. In einem Ausblick diskutiere ich dann weiterhin, wie sich die

[1] All errors and infelicities are my own responsibility. Many of the issues treated in this paper are inspired by discussions with Martin Schwartz. I am additionally grateful for comments and suggestions provided by Tim Aufderheide, Florian Wandl, two anonymous referees, and editor James Clackson, as well as audiences at the Universities of Zurich and Tübingen. Supplementary information containing a data appendix can be found at the PhilSoc website (https://philsoc.org.uk/transactions); additional code can be found at https://github.com/chundrac/w_ir_layers.

erarbeitete Methode auf andere Probleme der historisch-vergleichenden Sprachwissenschaft anwenden lässt.

[German]

1. Introduction

Isoglosses based on sound changes differentiating the West Iranian languages, a group comprising Persian, Kurdish, Balochi, and other speech varieties, have long been of interest to linguists. The West Iranian languages are traditionally divided into Northwest (containing Kurdish, Balochi, etc.) and Southwest (containing Persian and closely related dialects) subgroups, the latter of which can be defined by a small number of phonological and morphological innovations that have taken place before the attestation of Old Persian, its oldest member. At the same time, a comparable (if not larger) number of Persian innovations have taken place after Old and Middle Persian, and similar innovations can be seen in other West Iranian languages, showing the effect of complex areal networks that have existed during the development of these languages, shown in Figure 1.

A number of reflexes can be identified as Southwest Iranian or Northwest Iranian on the basis of the languages in which they occur; however, language contact has complicated the picture significantly. In some cases, it is not clear what the 'correct' outcome of a given Proto-Iranian sound should be; for instance, in the word for 'spleen' (Proto-Iranian **spr̥jan-*), Kurdish *sipił* shows what is thought to be a typically SWIr outcome (**r/r̥j > l*), while Persian *supurz* shows a typically NWIr outcome (**r/r̥j > rz*).

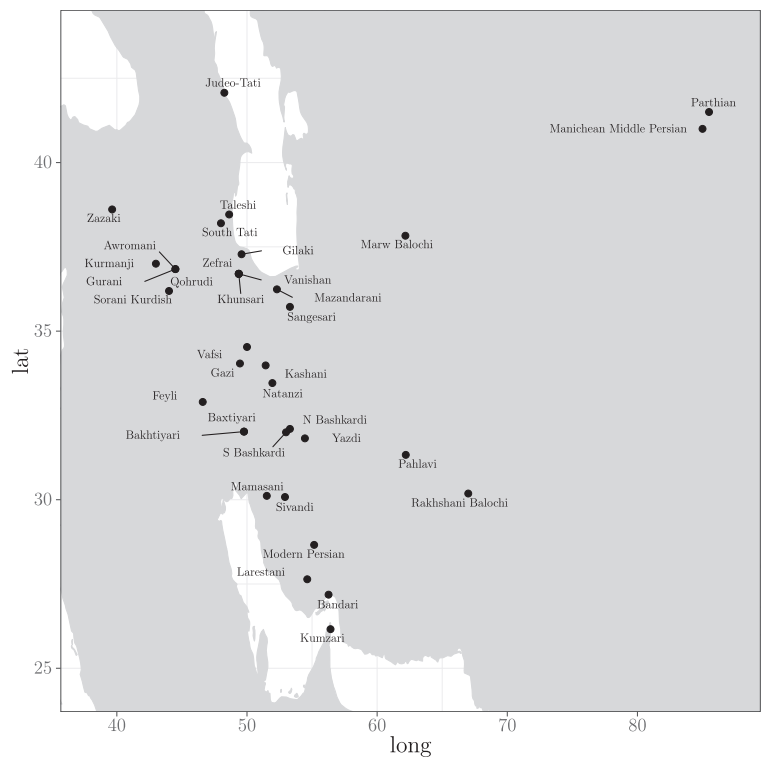

Figure 1. Approximate locations of languages in sample

Researchers in West Iranian dialectology have developed a set of diagnostics for marking individual words as loans in specific languages, but some of these diagnostics are better founded than others; in general, the picture is often so noisy, and these heuristics are so tightly intertwined, that all the facts cannot be qualitatively resolved within a traditional comparative-historical framework. I propose an alternative way of analysing West Iranian data that integrates insights from the comparative method with probabilistic modelling. While previous research has tended to make hard decisions regarding a language's regular reflexes of sound change, this study avoids this approach; instead, I employ a quantitative approach intended to let regular behaviour fall out of the data.

This paper investigates this variation in historical phonology across etymological reflexes and languages on a large scale. Specifically, I use a Bayesian probabilistic model, the hierarchical Dirichlet process (HDP), to reduce the dimensionality of the data seen within and across languages into a set of latent, unobserved components representing dialect membership which can be shared by multiple languages. The HDP is non-parametric, meaning that there is no upper bound on the number of latent features inferred. Both languages and phonological variants are associated with the presence of a latent feature. This allows us to identify potential networks of language contact across the dataset.

This methodology sheds light on a number of unresolved issues in the literature on West Iranian dialectology. I find, unsurprisingly, that West Iranian languages show admixture (to differing degrees) from two major dialect components, roughly corresponding to Northwest and Southwest Iranian dialect groups. I provisionally resolve a small number of questions regarding the dialectal provenance of certain types of sound change; while the impact of this paper's results is somewhat limited due to the relatively small size of the data used, the results are interpretable, and the methodology I use is promising. I discuss future directions for models of this sort.

This paper is structured as follows: Section 2 introduces the concept of West Iranian and the languages surveyed by this paper. Section 3 gives a condensed overview of the problems posed by West Iranian historical phonology for the traditional comparative method of historical linguistics. Section 4 provides an in-depth (but not exhaustive) catalogue of sound changes in West Iranian languages, pinpointing a few problematic examples where traditional methodologies have led researchers to potentially unjustified conclusions regarding language contact within Iranian. Sections 5–7 provide a conceptual overview of the HDP and related models, highlighting their relevance to the issues outlined in the previous section; an outline of the rationale for the model used in this paper; and a high-level description of the inference procedure employed. Results and discussion follow in sections 8–9; technical details of the model employed can be found in the Appendix.

2. The West Iranian languages

The Iranian languages are traditionally divided into East and West subgroups, but the genetic status of these labels is shaky. Historically, Bartholomae (1883: 1) divided Old Iranian into western and eastern variants, the former represented by Old Persian, and the latter by Avestan. The *Grundriss der iranischen Philologie*, particularly Wilhelm Geiger's contribution, provides a great deal of information on dialectology and subgrouping of contemporary Iranian languages. In this work, the chief distinction that cuts across Iranian is between "Persian" and "Non-Persian" dialects (Geiger 1901: 414). East and West are used as purely geographic labels: at one point, Balochi, generally classified in later work as a West Iranian language, is referred to as East Iranian (p. 414). There is not full agreement regarding which Iranian languages are western and which are eastern; the languages Ormuri and Parachi are considered to be West Iranian languages by some scholars (Grierson 1918; Oranskij 1977;

Efimov 1986), but the consensus following Morgenstierne (1929) places them in East Iranian. The problematic nature of the geographic labels was noted at an early date by Bailey (1933). Sims-Williams (1996: 651) states that East Iranian is better understood as a *Sprachbund* than a genetic grouping, as there are very few non-trivial innovations shared by all languages in this group. Wendtland (2009) finds that there are no secure shared phonological or morphological characteristics between the East Iranian languages, and argues against Northeast and Southeast subgroups (a division provisionally suggested in Morgenstierne 1926 and followed in Oranskij 1977; Kieffer 1989; and elsewhere). Cathcart (2015), Korn (2016, 2019) argue there are virtually no non-trivial innovations shared among West Iranian languages that could serve as diagnostics for subgrouping.

Regardless of the genetic status of West Iranian, the label is meaningful, not only from a typological standpoint (West Iranian languages are highly convergent in their morphosyntax) but in terms of many of the diachronic trends displayed by West Iranian languages. Contact with non-Indo-European linguistic stocks, such as Turkic and Semitic, may have aided in shaping the linguistic profiles of West Iranian languages (Stilo 2005; 2018). Even if there are no good shared genetic innovations among West Iranian languages, the study of inter-dialectal West Iranian contact has the potential to shed light on the socio-historical development of Iran and surrounding regions.

The West Iranian languages are traditionally divided into Northwest and Southwest groups. The Southwest group, comprising Old, Middle and New Persian, as well as closely related dialects such as Bashkardi, Kumzari, Judeo-Tati, and others, is generally viewed as a genetic subgroup, defined by a small number of innovations. The Northwest group has fewer subgroup-defining innovations uniting it. The finer details of this distinction are not of particular importance to this paper, as the major goal here is for dialectal groupings to fall out of the behavior displayed by languages in this paper's sample, which are listed in Table 1.

3. West Iranian historical phonological variation

West Iranian languages show a great deal of deviation from expected outcomes of historical phonology. This is clear in the oldest language; Old Persian contains a number of words which display the reflexes *s*, *z*, and *sp* and *zb* for Proto-Iranian (PIr) **ć*, **ǰ*, **ću̯*, and **ǰu̯* instead of the expected outcomes *θ*, *d*, *s* and *z*. This has led scholars to draw a distinction between 'proper Old Persian' and 'Median' forms (cf. Hoffmann 1976: 60ff.), the latter label an allusion to the confederation which preceded the Achaemenid Empire (with which Old Persian is associated). Although we can reliably identify only a single form as explicitly Median (*σπακα* 'dog', recorded by Herodotus, which shows the sound change **ću̯* > *sp*), a number of Old Iranian onomastic items are generally assumed to be Median.

Words containing irregular historical phonological reflexes are common in Middle and New Persian as well, and are generally ascribed to contact with Northwest Iranian languages (although there are several probable loans from East Iranian as well; see Sims-Williams 1989: 167). Northwest Iranian languages show the same degree of irregularity and contain a number of clear loans from various chronological stages of Persian, which is not surprising, given the sociopolitical influence of the Persian language in Iranian antiquity and onward.

It is likely that a number of mechanisms have worked together to create the complex patterns seen across West Iranian. These include (but are not limited to) language-internal factors such as the following:

- poorly understood conditioning environments: we may not fully understand the factors influencing regular sound change within languages; and
- analogical change, including paradigmatic levelling and extension, contamination, etc.

Table 1. West Iranian languages in the data set, along with alternative names and subvariants (in italics)

Language	Alternative names/ subdialects	Sources	Glottocode
*†Middle Persian (Manichean, MMP)		Durkin-Meisterernst 2004	midd1352
*†Middle Persian (Pahlavi, Phl)	*Book Pahlavi*, *Psalter Pahlavi*	MacKenzie 1971	pahl1241
*Parthian (Manichean, Pth)		Durkin-Meisterernst 2004	part1239
Awromani	*Pawai*	Benedictsen & Christensen 1921	gura1251
†Bakhtiyari		Anonby & Asadi 2014	bakh1245
Marw Balochi		Elfenbein 1963	west2368
Rakhshani Balochi		Barker 1969	west2368
†Bandari		Pelevin 2010	band1335
†N. Bashkardi		Skjærvø1988	bash1263
†S. Bashkardi		Skjærvø1988	bash1263
Zoroastrian Dari	Yazdi, Gabri (pejorative)	Ivanow 1940, Vahman & Asatrian 2002	zoro1242
†Feyli		Mann 1909	sout2640
Gazi		Eilers 1978	gazi1243
Gilaki	*Rashti*	Rastorgueva et al. 2012	gila1241
Gurani	*Kandulai*	Benedictsen & Christensen 1921, Hadank 1930	gura1251
Kashani		Žukovskij 1922	kash1282
Khunsari		Eilers 1976	khun1255
†Kumzari		Thomas 1930, van der Wal Anonby 2015	kumz1235
Kurmanji Kurdish		Soane 1913, Thackston 2006	nort2641
Sorani Kurdish		Blau 1980	cent1972
†Larestani		Kamioka and Yamada 1979	lari1253
†Mamasani		Mann 1909	mama1269
Mazandarani	Tabari	Nawata 1984	maza1291
Natanzi		Žukovskij 1922	nata1252
†New Persian		Steingass 1892	midd1352
Qohrudi	Soi	Žukovskij 1922, Mann and Hadank 1906–1932	khun1255
Sangesari		Azami & Windfuhr 1972	sang1315
Sivandi († according to Windfuhr 2009 but not Windfuhr 1991)		Lecoq 1979	siva1239
Taleshi	Talysh	Schulze 2000, Paul 2011	taly1247
†Judeo-Tati		Miller 1892, Authier 2012	jude1256
S. Tati	*Challi, Eshtehardi, Takestani*	Yar-shater 1969	take1255
Vafsi		Stilo 2004	vafs1240
Vanishani		Žukovskij 1922	khun1255
Zazaki	Dimli	Paul 1998b	diml1238
Zefrai		Žukovskij 1922	khun1255

Along with sources from which information was taken, and compatible glottocodes (Hammarström et al. 2017). Frequently used abbreviations for language names are provided. Pre-modern languages are indicated with an asterisk, Southwest Iranian languages with †. Transcriptions for pre-modern forms follow the sources cited.

Additionally, inter-language factors like the following are almost certainly involved:

- borrowing of lexical items; and
- lexical diffusion of sound changes.

As mentioned above, most explanations of irregularity appeal to lexical borrowing, often from an identifiable source such as Persian. However, it is additionally possible that more than one dialectal source of similar-looking reflexes was involved (e.g., the change *$\underset{\sim}{u}$- > *b* may not have been restricted to Persian); furthermore, it is possible that under the umbrella of widespread multilingualism, speakers imposed sound changes from one dialect onto words from other speech varieties. In certain *Mischformen* it is quite clear that diffusion of sound changes was at

work, rather than wholesale lexical borrowing. In many cases, however, it is not possible to distinguish between the two mechanisms. Additionally, it is not entirely clear whether similar-looking sound changes should be treated as unified, stemming from a single speech variant, or whether nearly identical sound changes were developed in parallel in different speech communities, possibly at different times. I identify some of these problems in the survey of sound change below, and propose a data-driven solution to at least some of these issues.

4. West Iranian historical phonology

Below, I give a synopsis of historical phonological innovations in West Iranian languages (viewed through the lens of Persian, which has the best-documented historical record) and discuss outstanding problems. These developments are given in rough chronological order (where a chronology can be securely established), starting with innovations preceding Old Persian, and so on, focusing on some particularly vexing problems.

Dialectal differentiation is visible in the earliest attested West Iranian records, which consist of Achaemenid Old Persian inscriptions, as well as fragmentary Median records. At this stage, several phonological and morphological innovations that define Southwest Iranian as a subgroup can be identified (see below).

The *locus classicus* of West Iranian dialectal differentiation is Tedesco's (1921) study of Middle Persian/Parthian isoglosses in the Manichean texts of Turfan. Lentz (1926) discusses dialectal variation found in the Šāh-nāma. In many cases this variation can be periodized with respect to when variant forms were introduced, especially in the case of Persian (cf. Paul 2005). The isoglosses identified in these works have served as the basis for a large number of dialectological investigations. Over the past century, the list of variables has been supplemented (Bailey 1933, Krahnke 1976, Stilo 1981), and scholars have debated which features in particular are the most meaningful for West Iranian dialectology (Paul 1998a, Korn 2003, Windfuhr 2009), in terms of joint versus independent innovations.

4.1. Changes to PIr *ć, *ǰ

The changes *ć > θ (> Middle Persian, New Persian *h*), *ǰ > *d*, *ću̯ > *s*, *ǰu̯ > *z*, *θr > ç,[2] are found in a stratum of Old Persian (OP) vocabulary, and thought to be the expected outcome in Southwest Iranian languages. However, OP also exhibits a number of doublets or irregular reflexes of the aforementioned Proto-Iranian sounds, usually ascribed (as mentioned above) to Median admixture, though we know little about the true nature of the Median language, given the paucity of records.

This variation is well attested in Old Persian: PIr *ć > *θ* in one layer of vocabulary, but *s* elsewhere; PIr *ǰ > *d* in (likely) the same stratum, but *z* elsewhere. This variation is described further below.

4.1.1. PIr *ć-

OP (or post-OP) initial *θ*- consistently corresponds to Middle Persian (MP; I distinguish between Phl and MMP only for forms exhibiting variation between the two dialects) *s*-:[3]

[2] The phonetic nature of this sound is unclear; see Kümmel (2007: 287ff.) for a survey of possible realizations.

[3] The symbol >, used to represent descent between forms with regular sound change, is used somewhat liberally in this paper in that it may connect OP, MP, and New Persian (NP) forms that are not actually in a relationship of direct descent (given contact); → denotes an analogical change, possibly in combination with the operation of regular sound change.

OP *a**θ**anga-* > ***θ**anga-*: MP ***s**ang* 'stone'.

The fact that OP *θ* develops to MP *h* in most environments has led many scholars to assume that forms with MP *s-* are NW Iranian loans (cf. Gershevitch 1962a: 2); however (Salemann 1901) takes initial MP *s-* to be the regular reflex of earlier *θ-*. The development of PIr **ć-* to *h-* is shown in a single form, NP ***h**adba* 'centipede' found in the Burhān-i Qāṭi', a 17th century dictionary (Morgenstierne 1932: 55).

*4.1.2. PIr *ću̯*

Reflexes of PIr **ću̯* are highly probative with respect to Iranian subgrouping; the 'proper' Southwest Iranian outcome is taken to be *s*, while Khotanese Saka and Wakhi show *š*. Kurdish and Balochi, showing 'transitional' behaviour between Northwest and Southwest Iranian, appear to participate in the change **ću̯* > *s* with Southwest Iranian, but not the changes **ć* >*θ*>*h*, **ǰ* > *d*. In most Southwest Iranian dialects the change **ću̯* > *s* must postdate the change **ć* >*θ*. Northwest Iranian languages (other than Kurdish and Balochi) and East Iranian languages (other than Khotanese Saka and Wakhi) show *sp*, or a sequence of sounds thought to descend from it, e.g., Ossetic *fs*; Khunsari, Gazi, Sangesari *asm* 'horse' < **aspa-*. Zoroastrian Dari *sv* is most likely secondary rather than an archaism preserving the glide **u̯*, as PIr **sp* surfaces as *sv* as well, e.g., ***sv**arz* 'spleen' < ****sp**ṛ́jan-* (Vahman and Asatrian 2002:21)

The change **ću̯* > *sp* cannot be reconstructed to a hypothetical ancestor of the Central Iranian languages which share it (cf. Skjærvø 2009: 50–51) without excluding Kurdish and Balochi from this group, but these languages cannot be placed in the Southwest Iranian group: in most Southwest Iranian dialects, the change **ću̯* > *s* must postdate the change **ć* (which probably had a phonetic value close to [s]) >*θ*, as **ću̯* became **θ* only in highly marginal dialects, e.g., ***t**eš* 'louse' in Judeo-Shirazi (a dialect closely related to Persian but somewhat differentiated in terms of historical phonology), if from ****θ**iša-* < ****ću̯**iša-* (Morgenstierne 1960; Borjian 2020). It is likely that changes to **ću̯* represent a sort of areal diffusion among (originally) non-peripheral Iranian languages, albeit an old one which has operated prior to early Median and Scythian onomastic items and is found in the archaic Avestan language as well. It is worth noting that similar fortition of OIA *śv* has taken place in the peripheral Indo-Aryan language Khowar as well as some Nuristani languages, though Morgenstierne (1926; 1932) cautions against connecting these developments with the Iranian one.

Persian shows reflexes containing *sp* at all chronological stages. New Persian also shows the cluster *sf*. Henning argues that this cluster cannot be secondary from earlier **sp*, and could instead be from a dialect in which PIr **ću̯* "resulted directly in *sf*" (Henning 1963: 71, fn. 13). Schwartz (2006: 223) argues against the influence of Arabic (which resulted in certain sporadic *p* > *f* changes, since Arabic lacks *p*) in certain words with *sf*. The circumstances under which NP *sf* came about remain unclear.

*4.1.3. PIr *ćr*

A small number of Old and Middle Persian words show OP *ç*, MP *s* for PIr **ćr*, e.g., **ni-**ćr**ai-* 'restore' > OP *niyaçāray-* caus. (contaminated with *dāraya-*, according to Kent 1951: 188), MP *nisāy-* 'conveying, dispatch' (Cheung 2007: 355). Kent (1942: 80) claims that OP **çunautiy* 'hear' 3sg (< **ćrunauti*) yields NP *šunūdan*, but the latter form is better connected with **xšnau-* 'hear' (Cheung 2007: 456). Elsewhere, Middle and New Persian show *s(V)r* and on occasion *š*:

PIr ****ćr**āu̯aia-* 'hear' (caus.) > MP ***sr**āy* 'sing' > NP ***sur**ū(dan)*/***sar**āy* (Cheung 2007: 357)
PIr ****ćr**auni-* > NP ***sur**ūn* 'buttocks'

PIr **hu̯aćrū-* > NP *xu**sr**ū*, *xušū* 'mother-in-law' (possibly under influence from *xusūr* 'father-in-law', as suggested by a reviewer)
PIr **aćru-(ka-)* 'teardrop' > Phl *a**rs***, MMP *a**sr***; NP *a**rs***, *ašk*

*4.1.4. PIr *ćn, *ǰn*

Initial Proto-Iranian **ćn-* appears to surface as *sn-* in Iranian, though evidence is restricted to reflexes of one etymon in two languages (YAv ***sn**aθ-*, MP ***sn**āh-* 'strike' < PIr ****ćn**aθH-*; Cheung 2007: 349); medial **-ćn-* > OP *-šn-*, e.g., **u̯ać-**n**a-* > OP *vašna-* 'will, favour' (OAv *vasnā*, YAv *vasna* 'wish' inst.sg. may show the effects of analogy rather than a regular outcome; see Hoffmann and Forssman 2004: 102 as well as Schwartz 2010 for an alternative view).

PIr **ǰn* appears to have become OP *xšn-* (> MP *šn-* > NP *š(V)n-*) word-initially (e.g., PIr ****ǰn**ā-s(ć)a-* > *OP **xšn**āsa-* 'know' > MP *(i)**šn**ās-* > NP ***šin**ās-* 'recognize') and *-šn-* word-medially (e.g., **i̯a**ǰn**a-* > Phl *ǰa**šn*** > NP *ja**šn*** 'festival', **ga**ǰn**a-* > NP *ga**šn*** 'abundance'). From what we can tell, Northwest Iranian languages appear to have medial *-zn-* (on metathesis to *-nz-* in Median onomastic items, see Gershevitch 1962a), e.g., Parthian *ga**zn*** 'treasure' vs. NP *ga**šn*** 'abundance' < **ga**ǰn**a-*; Persian shows some forms with *-zn-*, possibly loans from Northwest or East Iranian, e.g., NP *gavazn*, cf. Sogdian *γ(')wzn-*, Khotanese Saka *ggūyzna-* (Gershevitch 1954: 57). At the same time, Northwest Iranian languages appear to agree with Persian in reflecting *(x)šn-* for initial **ǰn-*, e.g., ****ǰn**ā-(s)ća-* 'know' > Parthian ***išn**ās*, Qohrudi *e**šn**ās-* etc. (Cheung 2007: 466).

*4.1.5. PIr *ǰu̯*

The sequence **ǰu̯* is found in only a small number of Proto-Iranian etyma. Old Persian contains reflexes of only two of these forms, *patiya**zb**ayam* 'proclaim' 1sg. impf. (< **pati-**ǰu̯**a*i̯*am* < **pati-a-ǰuH-a*i̯*a-m*) and *hi**z**āna-* 'tongue' (< **h**ǐǰu̯**āna-*). The latter form is believed by most scholars to be "proper" OP (Kent 1951), and the former a Median loan.[4] The Middle Persian word for tongue is *uzwān*/*izwān*, and cannot be taken as a direct reflex of the OP form (since *z* from other sources does not change to *zw*); Middle Persian however shows the development **ǰu̯* > *z* in *parzīr-* 'keep away' if from **para-*/*pari-**ǰu̯**ar-* (Cheung 2007: 475). New Persian and related dialects tend to show *zVb* or *zVw* (e.g., NP *za**b**ān*, *zuwān* 'tongue' < **h**ǐǰu̯**āna-*).

*4.1.6. PIr *ći̯*

The development of PIr **ći̯* in Persian is not entirely clear. Its fate is intertwined with that of the cluster **θi̯* (< PIIr **ti̯*), whose regular Old Persian reflex is thought to be *šiy* (Kent 1951: 32); cf. **hu̯ai-pa**θi̯**a-* > MP *xwēš* 'self'.

Old Persian attests the cluster **ći̯* only word-medially, where it surfaces as *θiy* (showing characteristic OP anaptyxis between consonants and glides), e.g., *viθiyā* 'house' loc.sg., possibly via paradigmatic leveling of the stem *viθ-* (< **u̯ić-*). Old Persian does not directly attest this cluster word-initially; Middle Persian shows varying reflexes:

PIr ****ći̯**āu̯a-(ka-)* > MP ***siy**āh* > NP ***sy**āh* 'black'
PIr ****ći̯**aina-(mr̥ga-)* > MP *sēn murw* 'a fabulous bird' > NP *sīmurγ* (MacKenzie 1971: 74).
Young Avestan *saēna-* 'eagle' may show a dissimilation **ći̯* > **ć* the presence of the off-

[4] Verbal adjectives involving this form have found their way into Aśoka's Aramaic inscriptions (Schwarzschild 1960).

glide of the diphthong *ai*, which may also account for Persian *s*, but this development was clearly not pan-Iranian, since the initial consonant of Balochi *šēnak* 'falcon, hawk' (Korn 2005: 129) cannot continue PIr **ć-*.

Word-internally, Middle and New Persian reflect variation between earlier **šiį̂* and **θiį̂*:

PIIr **matsi̯a-(ka-)* → **māćį̂a-ka-* (Hoffmann 1976: 637, fn. 25 attributes the long vowel to Vr̥ddhi) > MP *māhīg* 'fish' > NP *māhī*
PIIr **tusći̯a-(ka-)* > likely PIr **tućį̂a-ka-* > MP *tuhīg* 'barren'
PIr **kaćį̂apa-ka-* > MP *kašavag* 'tortoise' > NP *kašav*, *kašaf*; cf. Bandari *kāsapošt* 'turtle', Balochi *kasīp* ~ *kasīb* 'turtle, tortoise', Kurmanji *kīsal*, Sivandi *kalapošt* (showing the East Iranian change **š* > *l*, according to Asatrian 2012; cf. Sims-Williams 1989: 167), Bakhtiyari *kāsepošt* 'turtle', Zazaki *kese*
PIr **u̯aćį̂ah-* > MP *wēš* 'more' > NP *bēš* 'more'; cf. Sivandi *vīštar*, Balochi (Marw) *geštir* 'greater, oftener'
PIr **kaćį̂ah-* > MP *keh* 'small(er), young(er)' > NP *kih* 'small'
PIr **maćį̂ah-* > MP *meh*, *mahy* (MMP <mhy>) 'great(er), old(er)' > NP *mih* 'big'

Variation of this sort led Gershevitch (1962a: 19–22) to argue that *θ* was an optional pronunciation of *s* in Old Persian (see Hoffmann cf. 1976: 637, fn. 25 for a disagreeing view). Klingenschmitt (2000: 203) ascribes alternation between pre-OP **-iį̂a-* and **-į̂a-* to variation among suffixes, with pre-OP **θį̂* yielding *šiy* and pre-OP **θiį̂* yielding *θiy*. Cantera (2009) invokes a rhythmic law proposed by Klingenschmitt (2000: 203) to account for phonological irregularities in Middle Persian nouns. Armed with these ideas, we can account for some of the variation within Persian, if we assume the pre-forms **mā**θiį̂a**-ka-* and **tu**θiį̂a**-ka-* versus **kaćį̂ápa-ka-* and **u̯áćį̂ah-* (the stress placement assumed here follows Back 1978: 30ff.), but this still does not explain *h* in reflexes of **kaćį̂ah-* and **maćį̂ah-*, which should have undergone the same development as **u̯áćį̂ah-*, as noted by Gershevitch.

In Persian, as elsewhere in West Iranian, language contact, analogy, and prosodically conditioned change have interacted to bring about the complex variation seen in reflexes of PIr **ći̯* and related sounds. The limited knowledge we have of late Old Persian prosody can help to tease out the role of the last mechanism, but only to a certain extent. It is tempting to account for variation in West Iranian languages with no documented history in a similar manner, but this is purely speculative. An instructive example is the following thought experiment: Korn (2005: 284) contends while discussing Balochi *kāsib*/*kasīp* 'turtle, tortoise' that 'a genuine Bal. word should show *š*'. However, given that Balochi *ī* reflects PIr **-iį̆ā̆-* (cf. Korn 2005: 105), a pre-Balochi form **kaćiį̆apa-* is not inconceivable on historical phonological grounds, but perhaps overly speculative since we know virtually nothing about the phonotactics, syllabification, and stress pattern of Balochi's precursor. However, we also do not know whether *š* is the Balochi reflex of **ći̯* across the board (as assumed by Korn), or only in specific environments. Ultimately, we may benefit from relaxing some of these assumptions and employing a probabilistic model that allows us to make generalizations regarding languages' diachronic behavior on the basis of intra-language and inter-language distributions of sound changes.

*4.2. PIr *θ*

PIr **θ* changes into OP *θ* (> MP, NP *h*) in most conditioning environments, though it may develop into MP *s-* word-initially, e.g., PIr ****θ***axta-* (cf. Khwarezmian ***θ***γd*) > NP ***s**axt* 'hard'. The change **θr* > *ç* [s] is well established, as is **θį̂* > *šy* (mentioned above), though numerous exceptions to these developments exist as well.

*4.2.1. PIr *θn*

There are relatively few Proto-Iranian sources of the cluster **θn*, but these are realized as *šn* across the board in West Iranian, to the exclusion of the possible Median proper name in Akkadian *Pa-at-ni-e-ša-* = Med *Paθnīi̯ēša-* **paθnī-aiša-* 'looking for a wife' (Tavernier 2007: 273).[5]

PIr **ara**θn**i-* > OP *ara**šn**i-* 'cubit' > MP *ārešn* > NP *ăreš**(n)***
PIr **dmāna-pa**θn**i-* > MP, Parthian *bāmbi**šn*** **'queen'**
PIr **-i-**θn**a-* > MP abstract noun suffix *-i**šn*** > NP *-iš*; cf. Zazaki infinitive suffix *-iš* (Benveniste 1935: 105)

Middle Persian *ārenč* > NP *āranj* 'elbow', a doublet with *ăreš(n)* 'cubit', is most likely a loan from a source closely related to Sogdian (cf. *ārin**č*** < **ara**θn**i-ka-*). If NP *āmvasnī* 'rival wife' is to be connected with **ham-paθnī-* (Tafazzoli 1974: 119; Monchi-Zadeh 1990: 134) it perhaps shows a secondary change **š* > *s* seen in some other words.

*4.3. PIr *št, *žd*

Old, Middle and New Persian (along with other Iranian languages) show variation between *st* and *št* for PIr **št*; later language attests variation between *zd* and *žd* (e.g., MP *mizd* 'reward', NP *mizd*, *mužd*). There is disagreement as to whether OP *st* for *št* is due to analogy (Kent 1951: 34) or a sound change defining Southwest Iranian (Skjærvø 1989), and what the relationship of this behaviour is to similar-looking developments in the later language. Lipp (2009: 196ff.) states that OP *-st-* (found as a reflex of PIE **-k̂-t-*, **-ĝ-t-*) is due to analogy, while other developments are due to a phonological change predating Middle Persian:

> PIE **h₃reĝ-to-* > PIr **ra**šta-*** → OP *rā**sta-*** 'right' > MP *rā**st***> NP *rā**st***
> PIr *mu**šti-*** 'fist' > MP *mu**št***, *mu**st***> NP *mu**st***
> PIr **-i**šta-*** (superlative suffix) > MP *-i**st***; e.g., Phl *bāli**st***, MMP *bāri**st*** 'highest' < **barǰi**šta-***; Phl *xwāli**st***, MMP *xwāri**st*** 'sweetest' < **hu̯arǰi**šta-*** (cf. Iron Ossetic *xorz*, Digor Ossetic *xuarz* 'good')

*4.4. *r* + CORONAL change

4.4.1. Change to l

A number of West Iranian forms show a sound change whereby **r* + CORONAL sequences become *l*. This behaviour is common in Middle and New Persian, perhaps representing a regular sound change which operated between Old and Middle Persian:

PIr **ǰr̥**d**-* > MP *di**l*** > NP *di**l*** 'heart'
PIr **u̯(a)**rd**a-* > MP *gu**l*** > NP *gu**l*** 'flower'
PIr **ćar**d**a-* > OP *θard-* > Phl *sā**l***, MMP *sā**r***>NP *sā**l*** 'year'
PIr **p(a)**rd**anku-* > NP *pa**l**ang* 'panther'; cf. Vedic *pr̥**d**āku-* (with meaning 'leopard' in the Paippalāda recension of the Atharva Veda, Zehnder 1999: 59), Sogdian *pwr**δ***'*nk* 'panther, leopard'

[5] A possible but highly unlikely exception is the OP form *kr̥nuvaka-* 'stonemason' (found in the Susa inscription of Darius, Schmitt 2009: 133, 145), if from **kr̥t-nu-aka-* following Kent (1942: 80), though Kent (1951: 180) is less certain regarding the presence of **-t-* and other scholars (e.g., Brust 2018: 163) make no mention of etymological **-t-*; a preform **kr̥-nu-aka-* is far more likely.

[6] This form is likely a Wanderwort, but seems reconstructible to Proto-Iranian; see Lubotsky (2001).

PIr *b**ŕ́**ant- > MP bu**l**and > NP bu**l**and ‘high’
PIr *na**r**d- ‘lament, moan’ > MP nā**l**- > NP nā**l**(īdan) ‘lament’ (Cheung 2007: 282)
PIr *ba**ŕ́**ād(a)- > MP bā**l**āy > NP bā**l**ā ‘height’
PIr *ma**ŕ́**- > MP māl- ‘rub, sweep’ > mā**l**īdan ‘rub, polish’
PIr *r̥**ĵ**ifia- > Phl āluh, MMP āluf > NP ā**l**uh ‘eagle’
PIr *ĵa**r̂n**u-mani- ‘gold neck’ > NP dā**l**-man ‘black eagle’ (Schwartz 1971: 292, fn. 14)
PIr *g(a)**r̥n**a-ka- (cf. Old Indic ga á-) > NP ga**l**(**l**)a ‘flock’ (Schwartz 1971: 292, fn. 14)
PIr *p**r̥**tu- → *p**r̥θ**u- > MP pu**hl** > NP pu**̆l** ‘bridge’
PIr *par**ć**u- → *par**θ**(a)u̯-a-ka- > MP pa**hl**ūg ‘side, rib’ > NP pa**hl**ū ‘side’
PIr *čaθu̯(a)r-ćat- (cf. Emmerick 1992: 309) > PSWIr *čaθu̯**r̂θ**at- > MP čehe**l**>NP čihi**l** ‘forty’ (Emmerick 1992: 309), Judeo-Tati č**ül** (Authier 2012: 88), cf. Zazaki čew**res** (Paul 1998a: 61)

In some cases, this development has operated across an intervening vowel, likely unstressed:

PIr *ć**a**r(**a**)-**d**āra- (Klingenschmitt 2000: 194) > Phl ‘Träger der Mund; Oberster’ s**ā**l**ā**r (MMP s**ā**r**ā**r ‘leader’) > NP s**ā**l**ā**r ‘leader’ (cf. NP sa**r**-**d**ār, perhaps a later compound)
PIr *pa**r**i-**d**āna- > NP p**ā**l**ā**n ‘pack-saddle’ (cf. Sogdian py**rδ**nn ‘saddle’, cf. Sims-Williams 1989: 181)
PIr *pa**r**i-**d**aĭ̯a- > NP p**ā**l**ē**z ‘garden’ (Cheung 2007: 53)

However, this development is not exceptionless: it does not operate in forms like NP pa**d**arzah ‘a wrapper in which clothes are folded up’, if from *pa**r**i-**d**arĵ-aka- (Cheung 2007: 63, marked as a loanword perhaps due to *z* < **ĵ*), which appears to have undergone a dissimilatory development *r*…*r* > ∅…*r* that is not paralleled in **ćar(a)-dāra-*. It is unlikely that language-internal factors (i.e., different conditioning environments) can account for the entire range of variation seen within Persian.

The lateralization of **r* + CORONAL sequences, from what we can see, post-dates Old Persian. However, there are a large number of exceptions to this rule within Middle and New Persian; for example, NP bu**l**and forms a doublet with bu**r**z, thought to represent a Northwest Iranian form (Beekes 1997: 3). For some etyma, Persian lacks *l*, while a non-Persian reflex displays it, e.g., NP su**r**pu**r**z ‘spleen’ versus Kd sipiø ‘*id.*’ < PIr *sp**r̥ĵ**an-. The uncertainty surrounding this behaviour can be summed up by the following comment by MacKenzie (1961: 78) on the outcome of PIr **rd*/**rĵ* in Kurdish: ‘I do not think it is possible to be certain which is the true Kurdish development, but whether we consider the many words with *l*/ø as native or loan-words their preponderance is significant’.

Gurani contains the forms zi**l** ‘heart’ (< **ĵ**r̥d***-) and wi**l**ī ‘flower’ (< **u̯**r̥d**a-*, suffixation unclear), which cannot be Persian loans; in the first, the change to OP *d* predates lateralization of **r*/**r̥d-*. In the second form, it is most probable that the change to *g-* in MP *gul* was triggered by the following **-r̥-*, which subsequently underwent lateralization (e.g., **u̯r̥da- > **gVrd- > *gul*, see Section 4.6).[8] Whether *l* in these forms owes itself to Persian influence as opposed to some other source is unclear.

[7] Perhaps a de-instrumental *d*-stem built to PIE **b^{h}erǵ^{h}-eh$_1$-*, cf. Latin *mercēs*, *mercēdis* ‘wages’, *herēs*, *herēdis* ‘heir’ (Weiss 2009: 304–5).

[8] If Semnani *val(a)* reflects full-grade **u̯arda-*, it could in theory be a Persian loan, since changes affecting MP *wa-* post-date the **rd* > *l* change; however, there is no concrete evidence that Persian continues **u̯arda-*.

*4.4.2. *rn > rr*

The change **rn > rr* is attested in Middle Persian and onward, as seen in the following examples:[9]

PIr **hu̯**rn**ah-/*far**n**ah-* (on the reconstruction of this etymon see Skjærvø1983; Lubotsky 2002) > Phl *xwa**rr**ah*, MMP *farrah* > NP *fa**rr*** 'glory'
PIr **u̯a**rn**a-ka-* 'lamb' > MP *wa**rr**ag* > NP *ba**rr**ah*
PIr **par**n**a-* > MP *par**r*** >NP *par(**r**)* 'feather'
OP *k**rn**uvakā* <*k-r-n -u-v-k-a*> 'stonemason' > MP *ki**rr**ōg* 'artisan'
PIr **d(a)**r-n-*** > MP *da**rr***- > NP *da**rr***- 'to rend, tear up'
PIr **da**rn**a-ka-* > NP *da**rr**ah* 'valley', cf. YAv *uši.darəna-* 'having reddish cracks', name of a mountain (Horn 1893: 124–5; Humbach and Ichaporia 1998:1 80)
PIr **us-pr̥**n**a-* > MP *aspurr* 'accomplished' (Klingenschmitt 2000: 228)

It has been suggested that the changes **rn > rr* and **rn > l* (see above) are interconnected, and that *l(l) ~ r(r)* variation in reflexes of **rn* represents dialectal variation within West Iranian (Schwartz 1971: 292, fn. 14).

Middle and New West Iranian languages as a whole show an overwhelming tendency toward the change **rn > r(r)*. West Iranian words for 'lamb', if reflexes of **u̯arna-(ka-)* (< PIE **u̯r̥h$_1$n-*; Mayrhofer 1992: 225–6), show this behavior across the board:

MP *wa**rr**ag* > NP *ba**rr**a*; Parthian *wa**rr**ag*; Awromani *værā*; Balochi *gwārag*; S Bashkardi *va**r**k*; Gurani *varāla*, *valala* (< **u̯arna-la-*?); Zazaki *vo**r**ek* 'lamb'

However, Balochi and Parthian forms show the change **rn > n(n)*; Zazaki shows *rn* only via analogical maintenance or restoration, but otherwise *r ~ r* (Korn 2005: 133–4).

4.5. r ~ l variation

Proto-Indo-European **l* surfaces as *r* in the vast majority of Iranian languages. PIE **l > *r* is often given as a Proto-Iranian sound change in most handbooks, yet there are a number of exceptions to this development (Schwartz 2008), indicating that PIE **l* has been conserved in some peripheral dialects. Northwestern dialects also contain morphological variants with *l* lost by Persian with congeners in Indic, e.g., Kashani *engulī*, Mazandarani *engel* (cf. Old Indic *aṅgúli-*) against NP *angušt* (cf. OInd *aṅgúṣṭha-*) 'finger' (Horn 1893; Krahnke 1976: 226–8).[10]

However, some cases of West Iranian *l* may be secondary rather than archaic (Hübschmann 1895: 262ff.). It is not clear, for example, where forms like S. Tati (Ebrahim-abadi) *nālbanda* ~ (Sagz-abadi) *nārbanda* 'elm' (Yar-shater 1969: 71) = NP *nārvan* belong.[11] Similarly, one finds S. Tati *kelma* 'worm' = NP *kirm*; S. Tati *anjila* (Yar-shater 1969: 71), Vidari *injil* (Baghbidi 2005: 36) = NP *anjīr* 'fig' (forms elsewhere in Iranian point to **r*, e.g., Sogdian *ančēr*, *anjēr*; Gharib 1995: 37). For 'worm', the evidence clearly points to an Indo-Iranian etymon **kr(i)mi-* containing *r*, and any instances of *l* in Iranian languages should be secondary (e.g., Ossetic *kælm* shows expected **r > l* change in anticipation of **i* or **i̯*). Likely innovations are also found in Kurdish *valg*, Judeo-Tati *velg* (Miller 1892), etc. = NP *barg* < **u̯arka-* (Horn 1893:

[9] The *rr* sequence found in NP *xurram* 'joyful, lucky' and some other forms may be secondary (cf. Horn 1893: 106; Hübschmann 1895: 55).

[10] If the term for 'shepherd' in languages spoken in the Caspian region, *gāleš*, comes, as suggested by Asatrian (2002), from **gau̯a-raxša-* 'cow protector' (cf. Old Indic *go-rak a-*), then the presence of *l* is in agreement with PIE **h$_2$leks-*, pointing to another possible archaism.

[11] The elm appears to be the frequent target of folk etymology in Iranian languages (Henning 1963: 70); it is possible that Tati speakers have conflated the tree's name with *nāl-band* 'smith, farrier' (Arabic *na'l* 'horseshoe'), on the basis of some perceived but non-obvious connection to horseshoes.

47); this variant surfaces in the Dari dialect of New Persian as *balg* (Korn 2005: 160). Non-archaic *l* can also be found in NP *šikār* 'hunt' vs. Bandari, Bakhtiyari *eškāl*, S. Bashkardi *šekāl* 'mountain sheep', if from a verbal root **skar-* with no good Indo-European cognates (Cheung 2007: 346). Ultimately, *r*∼*l* variation across West Iranian is due not only to preservation of original PIE **l*, but also a secondary change to *l* from original **r*, especially evident in loans originally from non-Iranian languages, e.g., Judeo-Isfahani *kelews* 'celery' = NP *karafs* (Stilo 2007) ← Arabic. We can be sure of the directionality in cases where there is secure evidence from outside of Indo-Iranian, but in the absence of such information, it can be difficult to tease apart primary and secondary *l*; it is equally unclear whether all variant pronunciations stem from the same dialectal source.

4.6. *Changes to PIr *u̯-*

Reflexes of PIr **u̯-* are characterized by a high degree of irregularity across West Iranian.[12] Developments within Persian serve to demonstrate the complexity of these developments. Proto-Iranian **u̯-* surfaces as Middle Persian *g-* before **r̥* and **i̯*, but is otherwise unchanged in Middle Persian (with a few stray exceptions; see below):

PIr **u̯r̥tka-* > MP ***g**urdag* > NP ***g**urdah* 'kidney'
PIr **u̯r̥ka-* > MP ***g**urg* > NP ***g**urg* 'wolf'
PIr **u̯r̥pa-ka-* > MP ***g**urbag* > NP ***g**urbah* 'cat' (cf. YAv *urupi-* 'dog, fox (?)' < **u̯r̥pi-*)
PIr **u̯(a)r̥da-*[13] > MP ***g**ul* > NP ***g**ul* 'flower, rose'
PIr **u̯r̥šna-ka-*[14] 'hungry' > MP ***g**ušnag*, ***g**ursag*; NP ***g**ušnah*, ***g**urusnah*; Bakhtiyari ***g**osne*; Balochi (Marw) ***g**ušnag*; Gaz ***v**ašše*; Larestani ***g**ošna*; Mazandarani ***v**ašnā*; Sivandi ***f**eše* 'qui a faim' (showing the Central Iranian development **u̯-* > *f-*, c.f., Asatrian 2012); Taleshi ***v**eši*; S Tati ***g**ošna*; Zaz ***v**eyšān*
PIr **u̯i̯āna-* > OP *****v**iyāna* (?) > MP ***g**yān* > NP ***j**ān* 'life, soul'
PIr **u̯i̯āka-* > OP *****v**iyāka* (?) > MP ***g**yāg* > NP ***j**āh* 'place'

Generally speaking, PIr **u̯ĭ-* > MP *wi-* > NP *gu-*:

PIr **u̯i-nāća-* > MP ***w**ināh* > NP ***g**unāh* 'sin'
PIr **u̯i-čāra-* > MP ***w**izār* > NP ***g**uzār(dan)*
PIr **u̯i-dāna-* > MP ***w**iyān* > NP ***g**iyān* 'tent' (cf. OInd *vi-dhā-* 'furnish, spread, diffuse'?)
PIr **u̯a/inǰēča-ka-* (?) > MP ***w**inǰišk* > ***b**injišk* ∼ ***g**unjišk* 'sparrow', cf. Baxtiyārī ***b**ingišt* (Schapka 1972: 236); cf. Challi ***v**eškenj* (Yar-shater 1969: 69)

However, some exceptions exist:

MP ***w**iškar* > NP ***b**išgar(d)* 'hunting ground'
MP ***w**iyābān* 'desert' > NP ***b**iyābān*

In the following forms, PIr **u̯V(C)r-* > MP *wVr-* > NP *gVr-*:

PIr **u̯arāǰa-* > MP ***w**arāz* > NP ***g**urāz* 'boar'
PIr **u̯art-* 'turn' > OP ***v**-r-t-* > MP ***w**ardišn* 'turning' > NP ***g**ardiš* (MacKenzie 1971: 87; Cheung 2007: 423–5)

[12] See Schwartz (1982) on certain conditioned reflexes of this sound.

[13] YAv *varəδa-* 'rose' points to (and Semn *val(a)* 'flower' seems to point to — perhaps also Pth *w'r*/wa:r/) **u̯arda-*, while the Persian forms, along with Gor *wilī*, may point to **u̯r̥da-* (MacKenzie 1961: 77 gives the former etymon for all these forms).

[14] See Klingenschmitt 2000: 208 regarding the reconstruction of this form (departing from the earlier reconstruction of Hübschmann 1895: 92), as well as the double reflex *š* ∼ *r(V)s*.

PIr **u̯ájra-* > MP ***warz*** > NP ***gurz*** ‘mace’ (cf. Bal ***burz*** ‘club’, Elfenbein 1963: 25)

Change to *g-* does not operate in the following words beginning with PIr **u̯V(C)r-*; most, but not all, have a grave (i.e., labial, labiodental, or velar) consonant later in the word:

PIr **u̯arna-ka-* > MP ***warrag*** > NP ***barra*** ‘lamb’
PIr **u̯ájr(a)-ka-* > OP ***v-z-r-k***/vazṛka-/[15] > Phl, MMP ***wuzurg*** (cf. Pazand ***guzurg***, Bailey 1933: 56) > NP ***buzurg***
PIr **u̯arka-* > MP ***warg*** ~ ***walg*** > NP ***barg*** ‘leaf’
PIr **u̯afra-* > MP ***wafr*** ‘snow’ > **βafr-* (cf. Judeo-Persian <bp¯r> [bafr], Paul 2013: 50) > NP ***barf***
PIr **u̯ara-* > MP ***war*** ‘breast’ > NP ***bar***
PIr **u̯ar-ma-* (? cf. Horn 1893: 298) > MP ***warm*** ‘pond’ > NP ***barm***
MP ***warm*** ‘memory’ > NP ***barm***
MP ***wardag*** ‘captive, prisoner’ > NP ***barda***
PIr **u̯arźa-* > MP ***warz*** ‘work, agriculture’ > NP ***barz*** ‘a sown field, agriculture’ (cf. NP *varzīdan* ‘sow a field’, with *v-*)

Elsewhere, PIr **u̯-* > MP *w-* > NP *b-*:

PIr **u̯ahāna-ka-* (Gershevitch 1952) > MP ***wihānag*** (Phl <**w**h’n(k)>, MMP <**w**h’n(g)>) > NP ***bahāna*** ‘reason, pretext’
PIr **u̯ahāra-* > MP ***wahār*** > NP ***bahār*** ‘spring’
PIr **u̯ata-* > NP ***wad*** > NP ***bad*** ‘bad’
PIr **u̯at-čaka-* > MP ***waččag*** > NP ***baččah*** ‘child’
PIr **u̯ana-* > MP ***wan*** > NP ***bun*** ‘tree’
PIr **u̯āta-* > MP ***wād*** > NP ***bād*** ‘wind’
PIr **u̯īćati-* > MP *wīst* > NP ***bīst*** ‘twenty’ (note also *s* for PIr **ć*, while other decads show *h*)
PIr **u̯ahi̯a-* > Phl ***weh***, MMP ***weh***, ***wahy*** > NP ***bih*** ‘better, good’
PIr **u̯aći̯a-* > MP ***wēš*** > NP ***bēš*** ‘more’
PIr **u̯ahišta-* ‘best’ > MP ***wahišt*** > NP ***bihišt***; cf. the name of a 4th cent. CE Christian martyr, *Gu(hi)štāzād* (Peeters 1910), the first member of which **u̯ahišta-*
PIr **u̯a/iṅǰēča-ka-* (?) > NP ***binjišk*** ~***gunjišk*** ‘sparrow’ (see above)
PIr **u̯rīñǰi-* > MP ***briňǰ*** > NP ***birinj*** ~***gurinj*** ‘rice’ (cf. AV+ *vrīhí-*)

As is apparent, none of the sound laws sketched above is exceptionless. It is almost certain that contact between closely related dialects is responsible for some of the doublets seen above. But it is also clear that succinct generalizations regarding the behavior of PIr **u̯-* in different conditioning environments are hard to come by. This issue has not received a systematic treatment in the literature. Lentz (1926: 280–1) seems to consider **u̯-* > *b-* the regular Southwest Iranian outcome. MacKenzie (1971: 76) takes the change **u̯-* > *b-* as a feature shared by Persian and Northern and Central Kurdish dialects, whereas ‘[i]n most other W.Ir dialects *w-* is little modified in this position, while in Bal. it has developed into *g(w)-*’.

Attempts to establish the regular behaviour of PIr **u̯-* for non-Persian West Iranian languages have proved as difficult as for Persian. Early Judeo-Persian records, thought to typify a link between Middle and Modern Persian, present an equally challenging picture (Paul 2013: 35ff.). An errant strain of Middle Persian shows *g-* for expected *b-*, e.g., Pazand *guzurg*: NP *buzurg* (Bailey 1933: 56). A large number of West Iranian languages leave **u̯-* more or less unmodified (surfacing as *v*, *w* or *f* but more importantly not merging with PIr **g-*,

[15] Schmitt (1989: 69) and others give this reading, departing from *vazraka-* (found in Kent 1951), on the basis of the later forms.

*b-), but forms with *g-* and *b-* still preponderate. For instance, while Zazaki usually shows *v-* (e.g., *vā* 'wind'), the word for 'blood' is *gūnī* < **u̯ahuni-*[16] (Paul 1998b) South Tati *varga* 'leaf' sits alongside *behār-* 'spring' (Yar-shater 1969: 95, 103, 110). The Kurmanji dialect of Kurdish shows a preference for *b-* where other languages do not, e.g., *burāz*, *vurāz* 'boar': NP *gurāz*; *birsī*, *birčī* 'hungry': NP *gurusnah* (Soane 1913; Thackston 2006; Chyet 2003), but elsewhere agrees with Persian, e.g., *gurg*, *gūr* 'wolf'.

If a regular outcome can be established for a given non-Persian language, there is a tendency to assume that any words containing deviations from it are loans from Persian (though this approach is in general avoided by Korn 2005). For instance, Marw Balochi *burz* 'mace' (< **u̯ájra-*; note the metathesis identical to Persian) does not show expected *g(w)-*, hence, Elfenbein (1963: 25) marks it as a 'Persic' loan. However, there is no reason to expect NP *b-* in a reflex of a Middle Persian word with an initial syllable of the shape **u̯ar(C)-*, unless a grave consonant is found later in the word (and if the sound law sketched above is accurate). The Northern Kurdish dialect Kurmanji does, as mentioned above; this behaviour can be found sporadically in other non-Persian languages as well (e.g., Mamasani *burâz* 'wild pig', Mann 1909: 184). Given this evidence, these languages may be more viable donors for Balochi *burz* than Persian (the metathesis found in both of the forms is another question entirely).

4.7. Metathesis

Over the course of Persian history, more than one development of metathesis has taken place (Hübschmann 1895: 266–7), involving the re-sequencing of word-final and some word-internal clusters ending in *r* (and on occasion *l*). By the advent of Middle Persian, we see ***narm*** 'soft' < **na**mr**a-* and ***warz*** 'club, mace' < **u̯a**jr**a-*. Fricative + *r*/*l* clusters (as well as some fricative + fricative clusters) have undergone metathesis after Middle Persian attestations:

PIr **u̯a**fr**a-* 'snow' > MP ***wafr*** > NP ***barf***
PIr **ta**xr**a-* 'bitter' > Phl ***taxl***, MMP ***tahr***>NP ***talx***; Phl ***taxlīh*** 'bitterness' > NP ***talxī*** (the latter change could be analogical)
PIr **ča**xr**a-* 'wheel' > MP ***čaxr***>NP ***čarx***
PIr **a**ć**ru-* 'tear' > Phl ***ars***, MMP ***ars***, ***asr*** > NP ***ars*** (alongside *ašk* < **a**ć**ru-ka*)

Other West Iranian languages vary as to whether they show metathesis in the same words; this variation is often language internal:

PIr **u̯a**fr**a-* > Bal ***barp***; Gaz ***vaf***, ***-varf*** (in compounds); Gur ***varwa***; Khun ***varf***; Lar ***vafr***, ***barf***; Maz ***varf***; Siv ***varf***; Tal ***var***; S Tati ***vara***; Zaz ***vewr***; Judeo-Tati ***vähr*** 'snow' can be found in the materials of Miller (1892: 59), but Authier (2012: 323) gives ***verf***.
PIr **ta**xr**a-*>Bal (Rakhshani) *ta(**h**)**l***

Language contact must have played a role in bringing about intense variation, but the exact mechanisms are unclear. Metathesis is generally associated with Persian, since it can be documented in Persian's history. However, it is not clear whether the presence of metathesis in a non-Persian language is due to wholesale lexical borrowing or lexical diffusion (i.e., the adoption of the pronunciation *rC* for earlier *Cr*). Lexical borrowing from Persian tends to be assumed in the literature. For languages with *varf*: NP *barf*, it is assumed that the loan is from Middle Persian, or some period predating the change of MP *w-* to NP *b-*; for instance,

[16] The word for blood shows irregular historical phonology across west Iranian. NP *xūn* (MP *xōn*) has either undergone a metathesis between **u̯* and **h*, or was subject to the same irregular *x*-prothesis as MP *xāyag*, NP *xāya* 'egg', *xirs* 'bear'. Parthian has *guxn*, with unexpected *g-*; Sivandi has *fūn*.

Eilers (1978: 749) derives Gazi *värf* 'snow' from MP *varf* [*sic*]. However, this is unlikely to be the case. If we take Judeo-Persian to be representative of the link between Middle and New Persian (cf. MacKenzie 2003), then Judeo-Persian forms like <*bpr*> [bafr] (Paul 2013: 50) make it clear that metathesis postdates the merger of MP *w-* with *b-*, and that an intermediate stage **warf* was unlikely. Additionally, *w-*, *v-*, etc. cannot be secondary from earlier **b-* in the forms given above, since most of the languages mentioned show *b-* for original PIr **b-*.[17]

This detail aside, there are other reasons to question the account of lexical borrowing from Persian: first, this metathesis may not be a solely Persian development. Since most West Iranian languages (with exceptions, e.g., Yarshater 1962) lost final syllable nuclei, it is likely that many languages had words ending in *-xr*, *-fr*, etc., clusters which posed articulatory and perceptual problems, and were resolved in a variety of ways, including metathesis. Second, many of the above forms can be analysed only as *Mischformen*, vitiating a lexical borrowing account. Instead, it is possible that speakers in a situation of heavy multilingualism imposed pronunciations from forms in one language upon their cognates in another, a well-documented phenomenon in situations of multidialectalism, generally affecting less frequently uttered words (Phillips 1984; Stollenwerk 1986; Wieling et al. 2011).

4.8. *Changes affecting *dr*

Gershevitch (1962b: 78–9) discusses reflexes of the word for 'spade', demonstrating that some modern West Iranian languages reflect a form **barda-* (metathesised from **badra-*, which is internally derived from **badar-*). The source of metathesis in **barda-* is unclear. (Schwartz 1971: 297–8) shows that Iranian languages continue a doublet in the word for 'grape', **angudra-* (> MP, NP *angūr*) ~ **angurda-(ka-)* (> NP *angurda*), the latter being secondary and a likely East Iranian loan into Persian and other languages. It is not clear whether the metathesis in **barda-* is a related phenomenon.[18]

4.9. *Prothetic x-, h-*

Two separate protheses have operated during the history of Persian. The first involves sporadic insertion of *x-* before an initial vowel, and predates Middle Persian; the second involves sporadic insertion of *h-* before an initial vowel, and predates New Persian.

These developments can be seen elsewhere in West Iranian, e.g., *xotkā* 'duck' (language unmarked by Asatrian 2012: 113) < **āti-ka-*; Kumzari, Bandari, Larestani *xars* 'tear' < **aćru-* (cf. Bakhtiyari *hars*, Zazaki *hesri*). Korn (2005: 155–9) provides a detailed treatment of this issue, and makes a strong case that some items showing initial *h-* in both Balochi and Kurdish are due to contact, though elsewhere, the sporadic presence of *h-* may be a sort of hypercorrection, as in many English dialects (Wells 1982: 252–6), and not necessarily due to wholesale lexical borrowing (further bolstered by the fact that many Iranian languages lose initial *h-* under varying circumstances, e.g., **hĭǰu̯āna-* 'tongue' > MP *izwān*, *uzwān*).

[17] Lenition often affects earlier intervocalic labial consonants, e.g., Qohrudi *vīxōvā* = NP *bē-x^{v}āb* 'sleeplessness' (Žukovskij 1922: 79), NP *bē-* < MP *abē* < **apa-ika-* (Durkin-Meisterernst 2014).

[18] Bailey's (1973) derivation of the ethnonym Baloch from **baδlaut-čī* < **u̯adra-u̯at(č)ī* '[land] having water [channels]' (cf. the Greek toponym *Gedrosia*) is criticized by Korn (2005: 47) on the grounds that there is no parallel for **dr* > **δl* > *l*. However, this form may speak to a near-identical metathesis to **badra-*, **angudra-* etc., though the change **dr* > **rd* > *l* is a not a common Balochi development.

4.10. č ~ š

Some quasi-systematic variation between *š* and *č* is found in forms across West Iranian. In some cases, original *č* becomes *š* due to the interference of Arabic, which lacks a phoneme *č* (in the relevant dialects), as in *šatranj* ~ *šatrang* 'chess' < MP *čatrang* (← Old Indic *catur-aṅga-*).

In other forms, as noted by Horn (1901: 71), *č* is secondary, e.g., Zor Yazdi *čūm* 'supper' = NP *šām* 'evening' (1st member < **xšap-*, cf. YAv *xšāfniia-*, Bartholomae 1904: 553); Kashani *čiltúk* 'unhulled rice' = NP *šaltok*; Kashani *cepūn* 'herdsman', Kurdish *čuwān* (*čōpān* 'butcher') = NP *čubān*, *šubān* < **fšu-pāna-* (Horn 1901). Martin Schwartz (p.c.) points out that reflexes of the latter etymon may have undergone influence from NP *čūb* 'staff, crook'.

*4.11. *t > r*

The change **t > r* in North Tati dialects was noted by Henning (1954: 173). This change is seen in other languages, e.g., Judeo-Yazdi *čer-* 'go' (< **čiu̯ta-*), Judeo-Isfahani *čer-* 'know' (< **čait-*), Kumzari *spīr*, North Bashkardi *espīr* 'white' < **ću̯aita-*. Some Central Dialects show variation between *šīr* ~ *šīt* for 'milk', though this may be due to the continuation of separate etyma **xšīra-* and **xšu̯ifta-*.

4.12. Other developments

Above, a number of developments thought to be of interest to West Iranian dialectology were discussed. In this study, it is not possible to consider all possible meaningful changes, including vowel fronting (Krahnke 1976), *p* ~ *f* variation (e.g., S Tati *fercel* 'dirty': Bakhtiyari *parčal*), and other isoglosses. A hope is that as digitization efforts grow, fully data-driven approaches will allow us to take into account a wider range of innovations (see Section 9 for details).

4.13. Key issues

The foregoing sections served to illustrate the difficulties posed for the traditional comparative method by West Iranian sound change. Along the way, some problematic analytical decisions made by scholars have been highlighted, which are restated here:

- Elfenbein (1963) assumes that Marw Balochi *burz* 'mace' is a Persian loan, given unexpected *b-*, but it could easily be from another language (Section 4.6)
- Eilers (1978) assumes that Gazi *värf* is a loan from Middle Persian **warf*, but no such form existed, given the relative chronology between the developments **u̯-* > *b-* and **-fr* > *rf*; if the metathesis shown by the Gazi form is due to Persian influence, lexical diffusion rather than lexical borrowing was likely involved (Section 4.7)
- Korn (2005) assumes that PIr **ći̯* > Balochi *š* in all conditioning environments, and hence, that Balochi *kāsib*/*kasīp* 'turtle, tortoise', is a loan, but we cannot be sure this is the case (Section 4.1.6)

It is hoped that the qualitative points made or revived here – namely that some of the segmental and prosodic contextual factors involved in West Iranian sound laws are indeterminate, that not all donor languages are necessarily Persian, and that pure lexical borrowing is not the sole mechanism of contact – are convincing on their own merits. Still, it remains difficult to resolve many of the questions raised above within the constraints of the

traditional comparative method. In general, it is difficult to maintain a bird's-eye view of the many innovations and archaisms that cut across the West Iranian lexicon; while discussing one type of variation, another type is ignored (the above discussion is no exception). The remainder of this paper develops a probabilistic methodology designed to relieve historical linguists of the need to make hard decisions regarding phonological outcomes in a dialectal group, and instead let regularities fall out of the data.

5. Mixed membership models

As described above, West Iranian languages show admixture from an unknown number of latent (i.e., unobserved or unknown) dialectal components, each with its own individual sound laws and analogical changes. The key aim of this work is to learn which underlying components have contributed various features to the noisy pattern observed. A number of statistical techniques exist for the purpose of reducing the dimensionality of multivariate categorical data; mixed-membership models of this sort learn clusters that capture co-occurrence patterns of features in a data set in a way that the human eye cannot easily manage to do. These include certain classes of so-called generative models, which attempt to tell a story specifying one or more latent parameters which are thought to have generated the observed data. The latent parameters specified in a generative model can be estimated, usually within a Bayesian framework, which infers their posterior distributions. Bayesian modelling allows prior distributions to be imposed over these parameters, which serves as a sometimes-necessary means of ensuring that the model embodies realistic behaviour.

I draw upon probabilistic models of document classification in order to motivate the model I use in this paper. Topic modelling, which seeks to identify the topics present in a set of documents by associating the words found in them with one or more topics, is a well-known application for Bayesian mixed-membership models. Latent Dirichlet Allocation (LDA) is one such model (Blei et al. 2003); it assumes a fixed number of topics. It assumes that there is an overall distribution over possible topics, that each document has a specific distribution over topics, and that each word in each document is distributed according to a particular topic. The posterior global distribution over topics, document-specific topic distributions, and word-specific topic associations can then be inferred; it should be noted that if the procedure is entirely unsupervised, topics will receive meaningless labels such as 'Topic 1' rather than 'History', and that these labels require further interpretation. LDA is highly similar to the Structure algorithm of population genetics (Pritchard et al. 2000), which has been used in some linguistic applications (Reesink et al. 2009; Bowern 2012; Longobardi et al. 2013; Syrjänen et al. 2016). Figure 2 provides a hypothetical representation of how inferred topic

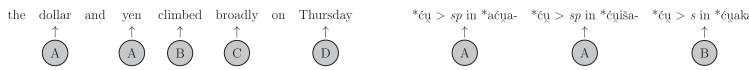

Figure 2. A schematic comparison of topic modelling and the approach used in this paper. *Notes*: In types of topic modelling such as Latent Dirichlet Allocation (LDA), it is assumed that each content word instance in each document in a corpus is generated by a given 'topic', as represented by the shaded circles. These assignments are unknown *a priori* and must be inferred from the data. The example on the left provides hypothetical posterior topic assignments for a sentence fragment, with individual topics generating word instances from similar semantic fields or spheres of reference. The example on the right provides a hypothetical assignment of dialect components to sound changes operating in words found in a single Iranian speech variety.

assignments might appear when LDA is applied to a document classification as well as the data investigated in this paper.

LDA requires practitioners to provide a specification a priori of the total number of topics assumed. It is often unreasonable to assume that an exhaustive list of possible topics has been drawn up. LDA has a non-parametric extension, the HIERARCHICAL DIRICHLET PROCESS (HDP, Teh et al. 2005; 2006), which allows for a potentially infinite number of topics. Over the course of the inference procedure, the model will return the number of topics which best explain the data.

I wish to extend the HDP model to the problem of admixture in the vocabularies of Iranian languages. By aggregating the patterns of variation in reflexes of a number of Proto-Iranian etyma, we may be able to identify components in the lexicon of each language which conceivably can be explained via historical language contact. I assume that there exists a set of areal components which underlie the variation reflected synchronically in West Iranian languages, and that we can recover their associations with variants and representation within languages.

An advantage of Bayesian models of this sort over classical methods for categorical data analysis is that they are generally robust to uneven or missing data – this is critical, given the patchy coverage for some Iranian languages. At the same time, mixed-membership models can potentially be sensitive to skews in data coverage. If a large number of features bearing on a particular isogloss are well attested in the data, but others are not, the algorithm used to infer component distributions may learn a distribution based on the former, even when the latter are highly relevant (but under-attested).[19] For this reason, I have taken pains to cast a wide net in the selection of features whilst maintaining parity in terms of the number of data points pertaining to each feature.

6. FEATURE SELECTION AND REPRESENTATION

For the upcoming analysis, words exhibiting the relevant Proto-Iranian sounds and sound sequences were collected from grammars and dictionaries by searching for the relevant semantic field, yielding a dataset of 1229 words. It is acknowledged that this means of data collection is highly limited, as some languages are better etymologized than others, and it would be preferable to take a top-down approach to data collection using a digitised etymological dictionary or etymological database, when such resources are developed.

As mentioned above, the goal here is to tease apart effects of areal contact and conditioning environments within West Iranian. As a concrete example, the presence of *b* in Sorani Kurdish *baran* (< *u̯ār-*) versus *g* in Sorani Kurdish *gurg* (< *u̯r̥ka-*) is due either to contact (e.g., the language has taken the words over from different donor languages) or different conditioning environments in the two words triggering the changes *u̯ > *b-* and *u̯ > *g-*. Information regarding conditioning environments is key to the feature representation which serves as model input. However, explicitly stipulating conditioning environments requires too many assumptions. I use the etymon itself as a proxy for conditioning environments; stating that *u̯ > *b-* in the etymon *u̯r̥ka- *'wolf'* is akin to stating that the change is triggered by the following *-r̥- and/or the following *-k-*. This would be a highly uneconomical analysis for a traditional historical grammar; however, any redundancy that this representation entails will be picked up by the model as part of the dimensionality reduction that it carries out.

A potential concern is that morphological variants of the same etymon are reflected in the catalogue of features; as mentioned above, different languages may continue different variants of a historical doublet *u̯r̥da-/*u̯arda- 'flower'. A similar concern is that of homophony

[19] A general practice is to remove uninformative and redundant features as well.

between reconstructed etyma, namely formally identical items that cannot be straightforwardly unified semantically (e.g., **u̯arma-* > NP *barm* 'pond, reservoir' and **u̯arma-* > NP *barm* 'memory'[20]). I leave the first problem untreated, with the hope that if a number of morphological variants of a single etymon are reflected in the data, this variation will be detectable in the model's output, namely via uncertainty in component level sound change distributions concerning this etymon.[21] I address the second problem by merging formally identical but semantically disparate reconstructions with one another, rather than treating them as instantiating different conditioning environments.

For the purposes of the model, each unobserved dialect component has a collection of SOUND CHANGE PARAMETERS associated with it. I envision this to be a CATEGORICAL probability distribution over the POSSIBLE OBSERVED OUTCOMES for each PIr sound of interest in each etymon (our proxy for the CONDITIONING ENVIRONMENT). These parameters can be visualized as shown in Table 2, for a given dialect component (probabilities are hypothetical).

Under the Neogrammarian hypothesis, sound change is exceptionless (Osthoff & Brugmann 1879; Bloomfield 1933; Hoenigswald 1965; Davies 1978). The probability of a sound change operating in a given speech variety is strictly categorical: one outcome will occur with 100% probability, all others with 0% probability. This paper's model RELAXES the Neogrammarian hypothesis, allowing sound change probabilities to be non-categorical. The first purpose is practical: rigid categorical-valued variables which assign zero, rather than infinitesimal probability mass to an outcome, will cause problems for the inference procedure, and enumerating all possible combinations of categorical feature states is computationally unfeasible. The second pertains to the real world, namely, to account for irregularity within a component that cannot be explained (due to analogy, so-called 'sporadic' change, or some other mechanism). However, it is still ideal to constrain these probability distributions such that they are SPARSE, with the majority of mass concentrated on one outcome, rather than SMOOTH (i.e., with mass distributed quasi-uniformly across outcomes). Ultimately, while we cannot constrain the model to enforce REGULAR sound change, we can employ priors that REGULARIZE sound change, encouraging probabilities to be very close to either 0 or 1.

For the purposes of this study, I make no attempt to model intermediate stages in sound change. For instance, it is not entirely clear whether the *f-* in Sivandi *fin* 'blood' < **u̯ahuni-* comes from an intermediate **x(u̯)*, or directly from **u̯* (though the latter scenario is more likely, as such changes are better attested in Sivandi). Techniques have been proposed for

Table 2. Hypothetical sound change probabilities for a latent dialect component

etymon	**u̯* > *b*	**u̯* > *g*	**u̯* > *w*/*v*	**ŕ́j* > *Vl*	**ŕ́j* > *Vrz*
**u̯ar-*	0.99	0.005	0.005		
**u̯ŕka-*	0.005	0.005	0.99		
**spŕ́jan-*				0.99	0.01
**ŕ́jifia-*				0.97	0.03

Note that probabilities of outcomes for the relevant PIr sound(s) sum to one, and that distributions are SPARSE (with the majority of mass concentrated on one outcome).

[20] These forms may be comprised by a single polysemous lexeme 'pond, reservoir, memory', but are given separate headwords in MacKenzie (1971); furthermore, colexification between 'pond' and 'memory' is not well attested cross-linguistically (Rzymski et al. 2020).

[21] A pervasive issue in the historical morphology of Indo-Iranian languages is the widespread use of the **-aka-* suffix. The *k* of this suffix has been elided in most modern West Iranian languages, including New Persian (Pisowicz 1985), making it difficult to determine whether certain forms in fact reflect **-aka-*. In general, I do not make a distinction between suffixed and unsuffixed forms, unless there is clear widespread evidence for a suffix, as in the case of *ašk* 'tear', found in New Persian, Gilaki, and other dialects.

reconstructing forms at intermediate nodes on fixed phylogenies (Bouchard-Côté et al. 2007; 2013), but not for situations like ours, where a form in a given language is generated by one of an unknown number of dialect components, rather than a single fixed ancestor.[22] The relatively abstract model of feature representation employed at least partly ensures that the sound changes dealt with by the model are meaningful. This paper's data set comprises 1160 sound change instances instantiating 190 unique sound change types in 32 West Iranian languages.

7. Inference

The generative process underlying the HDP and the technical details of inference can be found in the Appendix. A non-technical description of the HDP follows. Each data point (i.e., the reflex of a Proto-Iranian sound in a particular etymon in a given language, e.g., PIr *u̯- > NP *b*- in *u̯arma-*) is associated with a latent dialect component. The probability that a data point is associated with a given latent dialect component is dependent on a language-level probability distribution over dialect components $\boldsymbol{\theta}$, as well as a component-level distribution over sound changes $\boldsymbol{\phi}$. We do not know the values of these parameters, and must infer parameter values of high posterior probability (i.e., of high likelihood as well as high prior probability) from the data. Additionally, we do not know the true number of dialect components; this unknown must be learned by the model as well.

The HDP involves three hyperparameters: α is the concentration parameter of the symmetric Dirichlet prior over each dialect component's sound change distribution; the parameter γ controls the dispersion of data points across dialect components within a given language; δ controls the number of components inferred (at the risk of oversimplifying). These hyperparameters can be fixed, or (as in the case of the parameters described in the previous paragraph) given a fully Bayesian treatment by estimating them from the data.

Parameter and hyperparameter values can be estimated in several ways, including Markov chain Monte Carlo (MCMC) approaches such as Gibbs sampling (Geman & Geman 1984) or variational Bayesian methods (Bishop 2006). In the former procedure, values for each parameter are sampled stochastically on the basis of current values of all other parameters; after many iterations, the Gibbs sampler is guaranteed to draw samples from the posterior distribution of each parameter. Variational methods can be either deterministic or stochastic, and unlike MCMC methods, they assume a parametric form of the posterior distribution of each variable known as the variational posterior distribution, the parameters of which are iteratively updated. I use automatic differentiation variation inference (ADVI, Kucukelbir et al. 2017), as implemented in PyMC3 (Salvatier et al. 2016) to infer the posterior distributions of $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ (as described in the Appendix).

8. Results

As stated in the previous section, the inference procedure finds posterior probability distributions for two key parameters: $\boldsymbol{\theta}$, which gives each language's posterior distribution over dialect components; $\boldsymbol{\phi}$, which gives each dialect component's distribution over sound changes.

[22] It is also worth noting the existence of the mStruct model (Shringarpure and Xing 2009), a generalization of the Structure model which allows for mutations between historical and present-day populations. This model has the potential to account for intermediate stages, but situations of the sort described above are too rare in this paper's data set to justify the implementation of this considerably more complex methodology.

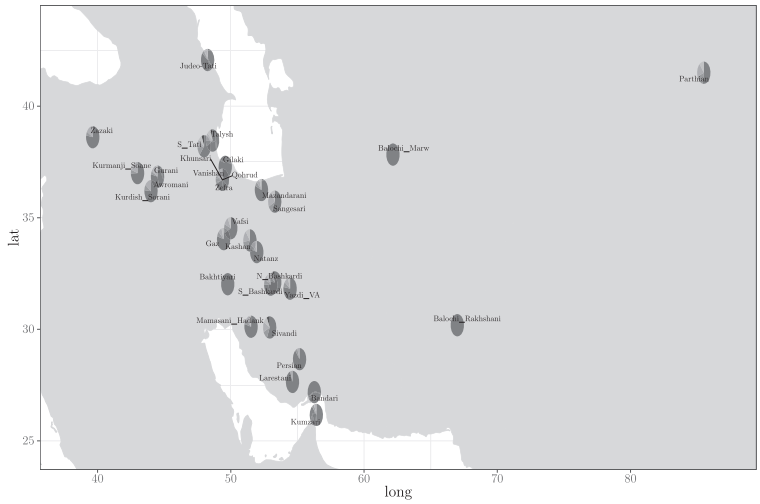

Figure 3. Language-level posterior distributions over latent dialect components

8.1. Language-level component distributions

As is clear from Figure 3, most languages in the sample show a relatively uniform profile in terms of their component makeup, favoring a small number of identical components. This pattern dovetails with received wisdom regarding the widespread dominance of Persian over other West Iranian languages in the period following the Safavid empire roughly 500 years before the present day (Borjian 2009); this homogenisation appears to have resulted in a more or less uniform profile for New West Iranian languages in terms of the sound changes reflected in their vocabularies (albeit with some degree of differentiation).

Virtually all languages in the sample show some degree of admixture from component k=1, along with differing degrees of components $k \in \{2,3,4\}$. Interestingly, k=1 appears to be strongly associated with developments that are thought to be typical of NW Iranian, such as the retention of initial *$\underset{\sim}{u}$ and the non-operation of the change *$\theta r > s$. It is not surprising that Modern Persian attests this component to a strong degree, given the well-known NW Iranian component in its vocabulary; at the same time, this proportion is higher than expected, as certain instances of SW Iranian behaviour are strongly associated with this component (e.g., *$\check{s}t > st$ in *$mu\check{s}ti$- fist, with *$\check{s}t > \check{s}t$ receiving higher posterior probability in components $k \in \{2,3\}$.

While visualising $\boldsymbol{\theta}$ gives us an overview of the predominant components present in a language's vocabulary, the picture presented is difficult to interpret in that it does not allow us to pinpoint exactly why these components are present, in terms of the reflexes to which they are linked. To gain a closer understanding of this issue, it is instructive to inspect the posterior probabilities of component membership for individual sound change instances. I describe these issues in detail in the upcoming section.

8.2. Posterior distributions over components for sound change instances

I use the MAP values of $\boldsymbol{\theta}$, $\boldsymbol{\phi}$ to reconstruct the posterior probability distribution over component membership for each individual token with index i, i.e., each sound change instance in each language in the data set, $P(z_i|\boldsymbol{\theta},\boldsymbol{\phi})$. These probability distributions are given in the Appendix, as well as a table summarising these values by averaging them across instances for each sound change type. These values allow us to address hypotheses about the provenance of certain sound changes (such as those discussed in Section 4.13). Many of these distributions

exhibit high uncertainty or entropy, with probability mass spread out across more than one dialect group rather than concentrated on a single group; this is perhaps a consequence of the relatively small size of the data set used in this study. At first glance, this uncertainty may seem to make the results difficult to interpret, but on the contrary these results are quite interpretable in that this uncertainty is relatively informative. Consider the following posterior distributions, concerning reflexes of Proto-Iranian **brz̧jant-* 'high' and **ćṷaka-* 'dog', which show the posterior probability of a sound change type given a dialect component (I exclude components where none of the probabilities exceed 0.05 for visual clarity):

etymon	sound	reflex	$p(k=1)$	$p(k=2)$	$p(k=3)$	$p(k=4)$	$p(k=5)$	$p(k=6)$
*brz̧jant-	rj	l	0.58	0.11	0.15	0.08	0.04	0.02
*brz̧jant-	rj	rj	1.00	0.00	0.00	0.00	0.00	0.00
*ćṷaka-	ćṷ	s	0.01	0.23	0.34	0.16	0.09	0.10
*ćṷaka-	ćṷ	sp	1.00	0.00	0.00	0.00	0.00	0.00

Tokens exhibiting the change **rj > rj* (our shorthand for forms such as *burz*, which do not undergo change to *l*) are associated strongly with a single latent dialect component, $k=1$, as are tokens exhibiting the change **ćṷ > sp*. Tokens exhibiting the changes **rj > l* and **ćṷ > s* do not show a particularly strong affinity with any latent dialect component. What is critical here is that changes of the former type, usually associated with Northwest Iranian languages, show behaviour that patterns much differently from changes usually associated with Southwest Iranian. This allows us to potentially classify individual change types according to whether the posterior distributions they exhibit are more in line with prototypical Northwest Iranian or Southwest Iranian sound changes.

On the basis of these distributions, I propose provisional solutions for the problems identified in Section 4.13. We find that Elfenbein's (1963) identification of Marw Balochi *burz* as a Southwest Iranian loan is indeed highly probable. Table 3 shows the component distributions of changes affecting PIr **ṷ-* in **ṷaȷ́ra-* 'mace, club'. We see that change to *b-* shows a distribution similar to those of the prototypically Southwest Iranian sound changes discussed above, while change to *g-* and *γ-* shows Northwest Iranian behaviour. Similarly, we find that change to *s* in **kaćia̯pa-* 'turtle/tortoise' patterns with canonically Northwest Iranian changes; hence, there is no strong reason to consider Balochi *kāsib/kasīp* a loan, as assumed by Korn (2005), since it patterns with many other typically Balochi features. Finally, changes concerning the etymon **ṷafra-* 'snow' suggest a Northwest Iranian origin for the presence of *w-* and a Southwest Iranian origin for metathesis in the form; hence, Gazi *värf* is probably a genuine *Mischform*, *pace* Eilers (1978), stemming perhaps from a scenario where speakers in contact with a neighbouring dialect exhibiting metathesis imposed this sound change on their inherited reflex of **ṷafra-*.

Table 3. Posterior component distributions for selected sound changes

etymon	sound	reflex	$p(k=1)$	$p(k=2)$	$p(k=3)$	$p(k=4)$	$p(k=5)$	$p(k=6)$
*ṷaȷ́ra-	ṷ	b	0.08	0.17	0.39	0.17	0.07	0.02
*ṷaȷ́ra-	ṷ	g	1.00	0.00	0.00	0.00	0.00	0.00
*ṷaȷ́ra-	ṷ	γ	1.00	0.00	0.00	0.00	0.00	0.00
*kaćia̯pa	ći	š	0.03	0.24	0.28	0.15	0.06	0.18
*kaćia̯pa	ći	s	1.00	0.00	0.00	0.00	0.00	0.00
*ṷafra	ṷ	b	0.70	0.06	0.12	0.04	0.03	0.03
*ṷafra	ṷ	w	1.00	0.00	0.00	0.00	0.00	0.00
*ṷafra	meta	meta	0.01	0.25	0.34	0.18	0.08	0.07
*ṷafra	meta	no meta	1.00	0.00	0.00	0.00	0.00	0.00
*ṷafra	meta	unclear	1.00	0.00	0.00	0.00	0.00	0.00

I exclude components with probability mass under 0.05 for visual clarity.

The results from the model are by no means the final word on these issues, and it is to be stressed that the conclusions drawn above are only tentative. It is likely that in many cases of idiosyncratic or unusual behaviour, the paucity of data employed is the culprit. I have demonstrated however that this sort of methodology serves as a promising technique for teasing apart questions concerning dialectal admixture in Iranian and other dialect groups. I am confident that this method will produce increasingly realistic and reliable results as digital resources for Iranian languages grow, facilitating big data approaches to questions such as those addressed in this paper.

9. Discussion and future directions

In this paper, I outlined a series of unresolved problems in Iranian dialectology and developed a probabilistic methodology designed to address these problems. In doing this, for the most part, I sought proof of concept as to whether Bayesian applications to Iranian dialectology might yield results which shed light on outstanding problems in the field as well as those that jibe with received wisdom. To some extent, this exercise was a success: I have shown that this model has great potential for resolving questions of the sort asked in this paper, but will benefit from further refinement. Below, I identify future directions that will improve this line of research:

9.1. Data

This paper made use of a relatively small data set compiled by hand from existing grammars. Sound changes were manually coded according to the behaviour they displayed. Additionally, only sound changes thought to be of interest to West Iranian dialectology were included in the feature catalogue. While I do not feel that this method of feature selection introduced any sort of pernicious bias that negatively affected results – after all, this paper focused on patterns displayed by sound changes thought to be probative for the purposes of Iranian dialect grouping across the vocabularies of West Iranian languages – it may be desirable to employ a more hands-off approach to feature selection and extraction, which will necessitate larger digitized etymological data sets. Additionally, this paper excluded East Iranian languages (including the languages Ormuri and Parachi), and shared patterns across both East and West Iranian should not be neglected; again, fulfilling this desideratum requires bigger data. At least two tacks can be taken for the purpose of data expansion: the first would involve digitizing of existing etymological dictionaries (Rastorgueva & Ėdel'man 2003; Cheung 2007) and converting them into a computationally tractable data format; however, no complete Iranian etymological dictionary currently exists for all parts of the lexicon, though current efforts such as the *Atlas of the languages of Iran* (Anonby et al. 2019), in its pilot phase at the time of writing, work towards filling this gap. The second approach involves applying semi-supervised cognate detection methods (List 2012; Rama 2016) to digitised Iranian word lists, which can potentially be coupled with semi-supervised methodologies for linguistic reconstruction (Meloni et al. 2019). While these methods still face many challenges, they can potentially save specialists a great deal of time and work in compiling large etymological resources. Whatever the approach employed, I believe that methods of the sort introduced in this paper will greatly benefit from the use of a larger data set. It is possible that the use of different data may yield different results from those reported in this paper.

9.2. Models

While this paper employed the HDP, several alternative types of non-parametric mixed-membership model exist. The HDP has certain properties that are undesirable for certain

uses, possibly including the dialectological application explored in this paper: specifically, the proportion of a component across all data points is correlated with its proportion within languages. It may be the case that a certain component is very rare overall, but well represented within one or a small number of languages. Certain alternatives to the HDP deal explicitly with this issue (Williamson et al. 2010).

9.3. Representation of sound change

In designing this paper's methodology, I made the radical decision to make no prior assumptions about the nature of the conditioning environments involved in the sound changes under study, instead treating entire etyma as conditioning environments. At first blush, this may seem like an implementation of the dictum that every word has its own history, attributed to dialect geographers such as Jules Gilliéron and Hugo Schuchardt. This is not the case: by linking the diachronic behavior of Proto-Iranian sounds in individual etyma to a finite number of dialect components exhibiting regularized sound change, we have inferred information regarding patterns of sound change within components as well as patterns of admixture within languages; the model ultimately embodies the interpretation of the above problem posed by dialect geographers that was provided by Bloomfield (1933: 360).

At the same time, it may be wrong to ignore the effect of phonetic similarity between conditioning environments on sound change. It may be the case that in a particular dialect component, **u̯-* undergoes a particular type of change in similar-looking etyma like **u̯ahi̯a-* and **u̯aći̯a-*, but a different change in a more dissimilar etymon such as **u̯arka-*. I have ignored this possibility; my goal was to let this systematicity fall out of the data in a bottom-up fashion. If desired, it is possible to employ a prior over sound change that can express covariance, such as the logistic normal distribution, which will encourage Proto-Iranian sounds to behave similarly in phonetically similar environments (which can potentially be operationalised via a smooth kernel function of the edit distance between the etyma containing these environments).

10. Conclusion

This paper introduced a new way of looking at Iranian dialectal relationships. The focus was on sound change in West Iranian, but this method can potentially be extended to linguistic groups of similar geographic spread and time depth. My chief goal was to provide a means for relaxing assumptions regarding the operation of individual sound changes in individual languages, and allow regular patterns to fall out of the data. Much work remains to be done in order to understand the complex history of the Iranian languages. Larger data resources are needed, and cooperation among linguists is needed in order to design and refine the probabilistic models we use; as data analysts, we need to work together to characterize the stochastic processes that we believe to have generated the data we observe, formalized in probabilistic terms. There needs to be a willingness to simplify models (if particular models are intractable), and an effort to keep models flexible, so that they can be expanded. It is likely that many of these goals are well within reach.

11. Acknowledgements

Open Access Funding provided by Universitat Zurich. [Correction added on 21st May 2022, after first online publication: CSAL funding statement has been added.]

Appendix

Model specification and inference

The generative process for the HDP involving the truncated stick-breaking construction (Ishwaran & James 2001) is given below. I set the truncation cutoff T, representing the maximum number of components, at 10.[23] L denotes the number of languages, S the number of environments in which sound changes occur, and N the number of data points in the data set. At a high level, this parameterisation allows for the prior over components to be highly skewed such that certain components are favored and certain components have prior probabilities close to zero, as justified by the data.

Draw hyperparameters α, δ, γ, which control the sparsity of $\boldsymbol{\phi}$, dispersion of data points across components within languages, and the number of components inferred:

$\boldsymbol{\beta} \sim \text{GEM}(\gamma)$ [draw atoms governing number of components learned]
$\boldsymbol{\theta}_{\ell,\cdot} \sim \text{Dirichlet}(\delta\boldsymbol{\beta}) : \ell \in \{1, \ldots, L\}$ [draw language-level distributions over component membership]
$\boldsymbol{\phi}_{t,s,\cdot} \sim \text{Dirichlet}(\alpha) : t \in \{1, \ldots, T\}, s \in \{1, \ldots, S\}$ [draw sound change parameters for every environment s and every component t]
For $i \in \{1,\ldots,N\}$ [for each data point (i.e., sound change instance)]

$z_i \sim \text{Categorical}(\boldsymbol{\theta}_{\ell_i,\cdot})$ [draw a component label]
$y_i | z_i = t \sim \text{Categorical}(\boldsymbol{\phi}_{t,x_i,\cdot})$ [sample the observed reflex on the basis of parameters associated with the label drawn in the previous step]

GEM(γ) denotes the Griffiths-Engen-McCloskey distribution, which has the following function when parameterized by γ:

$$\beta'_t \sim \text{Beta}(1, \gamma) : t \in \{1, \ldots, T\}$$

$$\beta_t = \begin{cases} \beta'_t \prod_{i=1}^{t-1}(1 - \beta'_i) & \text{if } t \in 1, \ldots, T-1 \\ \prod_{i=1}^{t-1}(1 - \beta'_i) & \text{if } t = T \end{cases}.$$

Under this process, each data point has the following likelihood:

$$P(y_i, x_i, z_i = t | \boldsymbol{\theta}, \boldsymbol{\phi}) = P(z_i = t | \boldsymbol{\theta}_{\ell_i}) P(y_i | \boldsymbol{\phi}_{t,x_i}).$$

Marginalising out the discrete variable $\boldsymbol{z}$ yields the following likelihood:

$$P(y_i, x_i | \boldsymbol{\theta}, \boldsymbol{\phi}) = \sum_{t=1}^{T} P(z_i = t | \boldsymbol{\theta}_{\ell_i}) P(y_i | \boldsymbol{\phi}_{t,x_i}).$$

The posterior distributions of $\boldsymbol{\theta}, \boldsymbol{\phi}$ can be used to reconstruct the probability that a given data point is associated with a given dialect component:

$$P(z_i = t | \boldsymbol{\theta}, \boldsymbol{\phi}, x_i, y_i) \propto \theta_{\ell_i,t} \phi_{t,x_i,y_i}.$$

[23] The choice of cutoff is arbitrary, and I have chosen a value I believe to be justified, given that the traditional literature hypothesises the existence of a much smaller number of groups (around two or three), and allows for a degree of computational tractability (larger values lead to longer runtimes for inference).

I place uninformative Gamma(1,1) priors over δ and γ, since we do not know *a priori* the degree to which data points within a given language should be dispersed across components, or how many components we should expect to find. I fix α, the concentration parameter of the symmetric Dirichlet prior over each dialect component's SOUND CHANGE distributions, at .0001 to encourage sparse sound change distributions.

I carry out inference using ADVI (Kucukelbir et al. 2017) in PyMC3 (Salvatier et al. 2016). ADVI allows users to define flexible and complex differentiable Bayesian models using a wide range of prior distributions over parameters. Certain probability distributions have constrained support: e.g., all samples from the Dirichlet distribution must be simplices summing to one; all samples from the Gamma distribution must be greater than zero. Parameters are mapped to unconstrained space and approximated with Gaussian variational posterior distributions, the parameters of which can be straightforwardly optimised using stochastic gradient descent. In mean-field ADVI, variational posteriors consist of independent Gaussian distributions, whereas in full-rank ADVI, variational posteriors make up a multivariate Gaussian distribution with non-diagonal covariance; I employ mean-field ADVI for simplicity. I optimize the model's variational parameters over four separate initializations of 100,000 iterations each, monitoring the evidence lower bound (ELBO) for convergence. The learning rate and β_1 parameter of the Adam optimizer (Kingma & Ba 2015) are set to 0.01 and 0.8, respectively. Posterior samples for each parameter are generated by drawing 500 samples from the fitted variational posterior.

Mixture models suffer from the so-called label switching problem, in which indices of identical components differ across initialisations/chains. To address this problem, I relabel the components inferred across initialisations 2–4 by permuting component labels and selecting the permutation which minimises the Kullback-Leibler divergence from the parameters for initialization 1 to the permuted parameters for the initialisation under consideration. This allows us to average parameters across initialisations, providing an approximation to the MAP configuration over component assignments for each item in the data set. Aggregating over these assignments produces MAP language-level distributions over component makeup.

Correspondence
Chundra A. Cathcart
University of Zurich
Email: chundra.a.cathcart@gmail.com

REFERENCES

ANONBY, ERIK & ASHRAF ASADI. 2014. *Bakhtiari studies: phonology, text, lexicon*. Uppsala: Acta Universitatis Upsaliensis.

ANONBY, ERIK, MORTAZA TAHERI-ARDALI & AMOS HAYES. 2019. 'The atlas of the Languages of Iran (ALI): A research overview', *Iranian Studies* 52(1-2). 199–230.

ASATRIAN, GARNIK 2002. 'The lord of cattle in Gilan', *Iran and the Caucasus* 6(1/2). 75–85.

ASATRIAN, GARNIK 2012. 'Marginal remarks on the history of some Persian words', *Iran and the Caucasus* 16(1). 105–116.

AUTHIER, GILLES 2012. *Grammaire juhuri, ou judeo-tat, langue iranienne des Juifs du Caucase de l'est. Beitrage Zur Iranistik*. Wiesbaden: Dr. Ludwig Reichert Verlag.

AZAMI, CHERAGH ALI & GERNOT WINDFUHR. 1972. *A dictionary of Sangesari with a grammatical outline*. Tehran: Franklin Book Programs.

BACK, MICHAEL 1978. *Die sassanidischen Staatsinschriften*. Leiden: Brill.

BAGHBIDI, HASAN REZAI 2005. 'Guyeš-i vidari', *Guyeš Šenāsi* 2(19). 18–26.

BAILEY, HAROLD W. 1933. 'Western Iranian dialects', *Transactions of the Philological Society* 32(1). 46–64.

BAILEY, HAROLD W. 1973. 'Mleccha, Balōč, and Gadrōsia', *Bulletin of the School of Oriental and African Studies* 36(3), 584–587.

BARKER, MUHAMMAD ABD-AL-RAHMAN 1969. *A Course in Baluchi*. Montreal: Institute of Islamic Studies, McGill University.

BARTHOLOMAE, CHRISTIAN 1883. *Handbuch der altiranischen Dialekte (Kurzgefasste vergleichende Grammatik, Lesestücke und Glossar*. Leipzig: Breitkopf & Härtel.
BARTHOLOMAE, CHRISTIAN 1904. *Altiranisches Wörterbuch*. Strassburg [Strasbourg]: Karl J. Trübner.
BEEKES, ROBERT S. P. 1997. 'Historical phonology of Iranian', *Journal of Indo-European Studies* 25(1–2). 1–26.
BENEDICTSEN, ÅGE MEYER & ARTHUR CHRISTENSEN. 1921. *Les dialectes d'Awromān et de Pāwä, Volume 6 of Historisk-filosofiske Meddelelser*. Copenhagen: Det Kgl. Danske Videnskabernes Selskab.
BENVENISTE, EMILE 1935. *Les infinitifs avestiques*. Paris: Adrien Maisonneuve.
BISHOP, CHRISTOPHER M. 2006. *Pattern recognition and machine learning*. Berlin: Springer.
BLAU, JOYCE 1980. *Manuel de Kurde (dialecte Sorani)*. Paris: Klincksieck.
BLEI, DAVID M., ANDREW Y. NG, & MICHAEL I. JORDAN. 2003. 'Latent Dirichlet allocation', *Journal of Machine Learning Research* 3(1). 993–1022.
BLOOMFIELD, LEONARD 1933. *Language*. New York: Holt, Rinehart and Winston.
BORJIAN, HABIB 2009. 'Median succumbs to Persian after three millennia of coexistence: Language shift in the Central Iranian Plateau', *Journal of Persianate Studies* 2(1). 62–87.
BORJIAN, HABIB 2020. 'The Perside language of Shiraz Jewry: A historical-comparative phonology', *Iranian Studies* 53 (3–4). 403–415.
BOUCHARD-CÔTÉ, ALEXANDRE, DAVID HALL, THOMAS L. GRIFFITHS, & DANIEL KLEIN. 2013. 'Automated reconstruction of ancient languages using probabilistic models of sound change', *Proceedings of the National Academy of Sciences* 110(11). 4224–4229.
BOUCHARD-CÔTÉ, ALEXANDRE, PERCY LIANG, THOMAS L. GRIFFITHS & DANIEL KLEIN 2007. 'A probabilistic approach to diachronic phonology', in Jason Eisner (Ed.), *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLPCoNLL)*. Prague: Association for Computational Linguistics. 887–896.
BOWERN, CLAIRE 2012. 'The riddle of Tasmanian languages', *Proceedings of the Royal Society B: Biological Sciences* 279(1747). 4590–4595.
BRUST, MANFRED 2018. *Historische Laut- und Formenlehre des Altpersischen: mit einem etymologischen Glossar*. Innsbruck: Institut für Sprachwissenschaft der Universität Innsbruck.
CANTERA, ALBERTO 2009. 'On the history of the Middle Persian nominal inflection', in Werner Sundermann, Almuth Hintze, & François de Blois (Eds.), *Exegisti monumenta: Festschrift in honour of Nicholas Sims-Williams, Volume 17 of Iranica*. Wiesbaden: Harrassowitz. 17–30.
CATHCART, CHUNDRA 2015. 'Iranian dialectology and dialectometry', Ph. D. thesis, University of California, Berkeley.
CHEUNG, JOHNNY 2007. *Etymological dictionary of the Iranian verb*. Leiden: Brill.
CHYET, MICHAEL L. 2003. *Kurdish-English dictionary/Ferhenga Kurmancî-Inglîzî; with selected etymologies by Martin Schwartz*. New Haven, CT: Yale University Press.
DAVIES, ANNA MORPURGO 1978. 'Analogy, segmentation and the early Neogrammarians', *Transactions of the Philological Society* 76(1). 36–60.
DURKIN-MEISTERERNST, DESMOND 2004. *Dictionary of Manichaean texts III*. Turnhout: Brepols.
DURKIN-MEISTERERNST, DESMOND 2014. *Grammatik des Westmitteliranischen (Parthisch und Mittelpersisch)*. Wien: Verlag der Österreichischen Akademie der Wissenschaften.
EFIMOV, VALENTIN ALEKSANDROVIČ 1986. *Jazyk ormuri v sinxronnom i istoričeskom osveščenii*. Moscow: Nauka.
EILERS, WILHELM with assistance from Ulrich Schapka. 1976. *Westiranische Mundarten aus der Sammlung Wilhelm Eilers. Vol. 1: Die Mundart von Chunsar*. Wiesbaden: Steiner.
EILERS, WILHELM with assistance from Ulrich Schapka. 1978. *Westiranische Mundarten aus der Sammlung Wilhelm Eilers. Vol. 2: Die Mundart von Gäz*. Wiesbaden: Steiner.
ELFENBEIN, JOSEF 1963. *A vocabulary of Marw Baluchi*. Naples: Istituto Universitario Orientale di Napoli.
EMMERICK, RONALD E. 1992. 'Iranian', in J. Gvozdanovic (Ed.), *Indo-European numerals*. Berlin: Mouton de Gruyter. 289–346.
GEIGER, WILHELM 1901. 'Kleinere Dialekte und Dialektgruppen', in Wilhelm Geiger & Ernst Kuhn (Eds.), *Grundriss der iranischen Philologie*, Volume 1. Strassburg [Strasbourg]: Karl J. Trübner. 287–423.
GEMAN, STUART & DONALD GEMAN. 1984. 'Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images', *IEEE Transactions on Pattern Analysis and Machine Intelligence* 6(6). 721–741.
GERSHEVITCH, ILYA 1952. 'Ancient survivals in Ossetic', *Bulletin of the School of Oriental and African Studies* 14(3). 483–495.
GERSHEVITCH, ILYA 1954. *A grammar of Manichaean Sogdian*. Oxford: Blackwell.
GERSHEVITCH, ILYA 1962a. 'Dialect variation in early Persian', *Transactions of the Philological Society* 63(1). 1–29.
GERSHEVITCH, ILYA 1962b. 'Outdoor terms in Iranian', in W. B. Henning & E. Yarshater (Eds.), *A locust's leg: Studies in honor of S.H. Taqizadeh*. London: Percy Lund, Humphries & Co. 76–84.
GHARIB, BADRESSAMAN 1995. *Sogdian dictionary: Sogdian-Persian-English*. Tehran: Farhangan Publications.
GRIERSON, GEORGE A. 1918. "The Ōrmurī or Bargistā language, an account of a little-known Eranian dialect', *Memoirs of the Asiatic Society of Bengal* 7(1). 1–101.
HADANK, KARL 1930. *Kurdisch-persische Forschungen, Abt. 3 (Nordwestiranisch) Bd. 2, Mundarten der Gûrân besonders das Kändûläi, Auramânî und Bâdschälânî*. Berlin: Walter de Gruyter.
HAMMARSTRÖM, HARALD, ROBERT FORKEL & MARTIN HASPELMATH 2017. 'Glottolog 3.3', Max Planck Institute for the Science of Human History. https://glottolog.org/
HENNING, WALTER B. 1954. 'The ancient language of Azerbaijan', *Transactions of the Philological Society* 54(1). 157–177.

Henning, Walter B. 1963. 'The Kurdish elm', *Asia Major* 10(1). 68–72.
Hoenigswald, Henry M. 1965. *Language change and linguistic reconstruction*. Chicago, IL: University of Chicago Press.
Hoffmann, Karl 1976. 'Zur altpersischen Schrift', in Johanna Narten (Ed.), *Karl Hoffmann: Aufsätze zur Indoiranistik, Volume 2*. Wiesbaden: Ludwig Reichert Verlag. 620–645.
Hoffmann, Karl & Bernhard Forssman. 2004. *Avestische Laut- und Flexionslehre* (2nd ed.), Volume 84 of Innsbrucker Beiträge zur Sprachwissenschaft. Innsbruck: Institut für Sprachwissenschaft der Universität Innsbruck.
Horn, Paul 1893. *Grundriss der neupersischen Etymologie*. Strassburg [Strasbourg]: Karl J. Trübner.
Horn, Paul 1901. 'Neupersische Schriftsprache', in Wilhelm Geiger & Ernst Kuhn (Eds.), *Grundriss der iranischen philologie*, Volume 1. Strassburg [Strasbourg]: Karl J. Trübner. 1–200.
Hübschmann, Heinrich 1895. *Persische Studien*. Strassburg [Strasbourg]: Karl J. Trübner.
Humbach, Helmut & Pallan Ichaporia. 1998. *Zamyād Yasht: Yasht 19 of the Younger Avesta: text, translation, commentary*. Wiesbaden: Harrassowitz.
Ishwaran, Hemant & Lancelot F. James. 2001. 'Gibbs sampling methods for stick-breaking priors', *Journal of the American Statistical Association* 96(453). 161–173.
Ivanow, Wladimir 1940. *The Gabri dialect spoken by the Zoroastrians of Persia, Volume 16 of Rivista degli Studi Orientali*. Rome: Scuola Orientale nella R. Università di Roma.
Kamioka, Koji & Minoru Yamada. 1979. *Lārestani studies*, Volume 1. Tokyo: Institute for the Study of Cultures of Asia and Africa.
Kent, Roland 1942. 'Vocalic r in Old Persian before n', *Language* 18(2). 79–82.
Kent, Roland 1951. *Old Persian*, Volume 33 of American Oriental Series. New Haven, CT: American Oriental Society.
Kieffer, Charles 1989. 'Le parāčī, l'ōrmuẓī', in Rüdiger Schmitt (Ed.), *Compendium Linguarum Iranicarum*. Wiesbaden: Ludwig Reichert Verlag. 445–455.
Kingma, Derek P. & Jimmy Ba 2015. 'Adam: A method for stochastic optimization'. Poster presented at International Conference on Learning Representations (ICLR). Available online at https://arxiv.org/abs/1412.6980
Klingenschmitt, Gert 2000. 'Mittelpersisch', in Berhnard Forssman & Robert Plath (Eds.), *Indoarisch, Iranisch und die Indogermanistik: Arbeitstagung der Indogermanischen Gesellschaft vom 2. bis 5. Oktober 1997 in Erlangen*. Wiesbaden: Ludwig Reichert Verlag. 191–229.
Korn, Agnes 2003. 'Balochi and the concept of North-Western Iranian', in Carina Jahani & Agnes Korn (Eds.), *The Baloch and their neighbours: Ethnic and linguistic contacts in Balochistan in historical and modern times*. Wiesbaden: Dr. Ludwig Reichert Verlag. 49–60.
Korn, Agnes 2005. *Towards a historical grammar of Balochi*. Wiesbaden: Ludwig ReichertVerlag.
Korn, Agnes 2016. 'A partial tree of Central Iranian', *Indogermanische Forschungen* 121(1). 401–434.
Korn, Agnes 2019. 'Isoglosses and subdivisions of Iranian', *Journal of Historical Linguistics* 9(2). 239–281.
Krahnke, Karl 1976. 'Linguistic Relationships in Central Iran', Ph. D. thesis, University of Michigan.
Kucukelbir, Alp, Dustin Tran, Rajesh Ranganath, Andrew Gelman, & David M. Blei. 2017. 'Automatic differentiation variational inference', *The Journal of Machine Learning Research* 18(1). 430–474.
Kümmel, Martin 2007. *Konsonantenwandel*. Wiesbaden: Dr. Ludwig Reichert Verlag.
Lecoq, Pierre 1979. *Le dialecte de Sivand*, Volume 10 of Beiträge zur Iranistik. Wiesbaden: Dr. Ludwig Reichert Verlag.
Lentz, Wolfgang 1926. 'Die nordiranischen Elemente in der neupersischen Literatursprache bei Firdosi', *Zeitschrift für Indologie und Iranistik* 4(1). 251–316.
Lipp, Reiner 2009. *Die indogermanischen und einzelsprachlichen Palatale im Indoiranischen*. Heidelberg: Carl Winter. 2 vols.
List, Johann-Mattis 2012. 'Lexstat: Automatic detection of cognates in multilingual wordlists', in Miriam Butt, Sheelagh Carpendale, Gerald Penn, Jelena Prokić, & Michael Cysouw (Eds.), *Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH*. Avignon: Association for Computational Linguistics. 117–125.
Longobardi, Giuseppe, Cristina Guardiano, Giuseppina Silvestri, Alessio Boattini, & Andrea Ceolin. 2013. 'Toward a syntactic phylogeny of modern Indo-European languages', *Journal of Historical Linguistics* 3(1). 122–152.
Lubotsky, Alexander 2001. 'The Indo-Iranian substratum', in C. Carpelan, A. Parpola, & P. Koskikallio (Eds.), Early contacts between Uralic and Indo-European: Linguistic and archaeological considerations. Papers presented at an international symposium held at the Tvärminne Research Station of the University of Helsinki 8-10 January 1999, Helsinki. 301–317.
Lubotsky, Alexander 2002. 'Scythian elements in Old Iranian', in Nicholas Sims-Williams (Ed.), *Indo-Iranian languages and peoples*, Volume 116 of Proceedings of the British Academy, Oxford. Oxford University Press. 189–202.
MacKenzie, David Neil 1961. 'The origins of Kurdish', *Transactions of the Philological Society* 60(1). 68–86.
MacKenzie, David Neil 1971. *A concise Pahlavi dictionary*. London: Oxford University Press.
MacKenzie, David Neil 2003. 'The missing link', in Ludwig Paul (Ed.), *Persian origins: Early Judaeo-Persian and the emergence of New Persian*. Wiesbaden: Harrassowitz. 103–110.
Mann, Oskar 1909. *Kurdisch-Persische Forschungen*, Abt. 1: Die Tâjîk-Mundarten der Provinz Fârs. Berlin: Reimer.
Mann, Oskar & Karl Hadank. 1906–1932. *Kurdisch-persische Forschungen*. Berlin: Walter de Gruyter.
Mayrhofer, Manfred 1992. *Etymologisches Wörterbuch des Altindoarischen*, Volume 1. Heidelberg: Winter.

MELONI, CARLO, SHAULI RAVFOGEL & YOAV GOLDBERG 2019. 'Ab antiquo: Proto-language reconstruction with RNNs', arXiv preprint arXiv:1908.02477.
MILLER, VSEVOLOD F. 1892. *Materialy dlja izučenija evrejsko-tatskogo jazyka*. St. Petersburg: Akademij Nauk.
MONCHI-ZADEH, DAVOUD 1990. *Wörter aus Xurāsān und ihre Herkunft*, Volume 15 of Acta Iranica. Leiden: Brill.
MORGENSTIERNE, GEORG 1926. *Report on a linguistic mission to Afghanistan*. Oslo: H. Aschehoug & Co.
MORGENSTIERNE, GEORG 1929. *Parachi and Ormuri*, Volume 1 of Indo-Iranian Frontier Languages. Oslo: Instituttet for Sammenlignende Kulturforskning, H. Aschehoug & Co. (W. Nygaard).
MORGENSTIERNE, GEORG 1932. 'Persian etymologies', *Norsk Tidsskrift for Sprogvidenskap* 5(1). 54–56.
MORGENSTIERNE, GEORG 1960. 'Stray notes on Persian dialects ii', *Norsk Tidsskrift for Sprogvidenskab* 19(1). 121–129.
NAWATA, TETSUO 1984. *Mazandarani*, Volume 17 of Asian and African Grammatical Manual. Tokyo: Institute for the Study of Languages and Cultures of Asia and Africa, Tokyo University of Foreign Studies.
ORANSKIJ, IOSIF M. 1963 [1977]. *Les langues iraniennes*. Translated by Joyce Blau. Paris: Klincksieck.
OSTHOFF, HERMANN & KARL BRUGMANN. 1879. *Morphologische Untersuchungen auf dem Gebiet der indogermanischen Sprachen*, Volume 2. Leipzig: Hirzel.
PAUL, DANIEL 2011. 'A comparative dialectal description of Iranian Taleshi', Ph.D. thesis, University of Manchester.
PAUL, LUDWIG 1998a. 'The position of Zazaki among West Iranian languages', in N. Sims-Williams (Ed.), Proceedings of the Third European Conference of Iranian Studies. Cambridge, 11–15 September 1995. Part I: Old and Middle Iranian Studies. European Conference of Iranian Studies. Wiesbaden: Dr. Ludwig Reichert Verlag. 163–177.
PAUL, LUDWIG 1998b. *Zazaki: Grammatik und Versuch einer Dialektologie*, Volume 18 of Beiträge zur Iranistik. Wiesbaden: Dr. Ludwig Reichert Verlag.
PAUL, LUDWIG 2005. 'The language of the Šāhnāme in historical and dialectal perspective', in Dieter Weber (Ed.), *Languages of Iran: Past and present: Iranian studies in memoriam David Neil MacKenzie*. Wiesbaden: Dr. Ludwig Reichert Verlag. 163–177.
PAUL, LUDWIG 2013. *A Grammar of Early Judaeo-Persian*. Wiesbaden: Ludwig Reichert Verlag.
PEETERS, PAUL 1910. 'S. Eleutherios-Guhištazad', *Analecta Bollandiana* 29(1). 151–156.
PELEVIN, MIKHAIL 2010. 'Materials on the Bandari dialect', *Iran and the Caucasus* 14(1). 57–78.
PHILLIPS, BETTY S. 1984. 'Word frequency and the actuation of sound change', *Language* 60(2). 320–342.
PISOWICZ, ANDRZEJ 1985. *Origins of the new and middle Persian phonological systems*. Kraków: Uniwersytet Jagiellonski.
PRITCHARD, JONATHAN K., MATTHEW STEPHENS, & PETER DONNELLY. 2000. 'Inference of population structure using multilocus genotype data', *Genetics* 155(2). 945–959.
RAMA, TARAKA 2016. 'Siamese convolutional networks for cognate identification', in Yuji Matsumoto, & Rashmi Prasad (Eds.), Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, Osaka, Japan. Osaka: The COLING 2016 Organizing Committee. 1018–1027.
RASTORGUEVA, VERA S. & DŽOJ I. EDELMAN. 2000- 2003. *Ètimologiceskij slovar' iranskix jazykov*. Moscow: Vostocnaja Literatura.
RASTORGUEVA, V. S., A. A. KERIMOVA, A. K. MAMEDZADE, L. A. PIREIKO, & J. I. EDELMAN. 2012. *The Gilaki Language*. Uppsala: Uppsala Universitet.
REESINK, GER, RUTH SINGER, & MICHAEL DUNN. 2009. 'Explaining the linguistic diversity of Sahul using population models', *PLoS Biology* 7, e1000241.
RZYMSKI, CHRISTOPH, TIAGO TRESOLDI, SIMON J. GREENHILL, MEI-SHIN WU, NATHANAEL E. SCHWEIKHARD, MARIA KOPTJEVSKAJA-TAMM, VOLKER GAST, TIMOTHEUS A. BODT, ABBIE HANTGAN, GEREON A. KAIPING, SOPHIE CHANG, YUNFAN LAI, NATALIA MOROZOVA, HEINI ARJAVA, NATALIIA HÜBLER, EZEQUIEL KOILE, STEVE PEPPER, MARIANN PROOS, BRIANA VAN EPPS, INGRID BLANCO, CAROLIN HUNDT, SERGEI MONAKHOV, KRISTINA PIANYKH, SALLONA RAMESH, RUSSEL D. GRAY, ROBERT FORKEL & LIST, JOHANN-MATTIS 2020. 'The database of cross-linguistic colexifications, reproducible analysis of cross-linguistic polysemies', *Scientific Data* 7(1). 1–12.
SALEMANN, CARL 1901. 'Mittelpersisch', in Wilhelm Geiger and Ernst Kuhn (Eds.), *Grundriss der iranischen Philologie*, Volume 1. Strassburg [Strasbourg]: Karl J. Trübner. 249–332.
SALVATIER, JOHN, THOMAS V. WIECKI, & CHRISTOPHER FONNESBECK. 2016. 'Probabilistic programming in Python using PyMC3', *PeerJ Computer Science* 2(1). 1–24.
SCHAPKA, ULRICH 1972. 'Die persischen Vogelnamen'. Ph.D. thesis, University of Würzburg.
SCHMITT, RÜDIGER 1989. 'Altpersisch', in Rüdiger Schmitt (Ed.), *Compendium Linguarum Iranicarum*. Wiesbaden: Ludwig Reichert Verlag. 56–85.
SCHMITT, RÜDIGER 2009. *Die altpersischen Inschriften der Achaimeniden*. Wiesbaden: Ludwig Reichert Verlag.
SCHULZE, WOLFGANG 2000. *Northern Talysh*. Munich: Lincom.
SCHWARTZ, MARTIN 1970 [1971]. 'On the Khwarezmian version of the muqaddimat al-adab as edited by Johannes Benzing'. *Zeitschrift der Deutschen Morgenländischen Gesellschaft* 120(1). 288–304.
SCHWARTZ, MARTIN 1982. '"Blood" in Sogdian and Old Iranian'. *Acta Iranica* 22(1). 189–196.
SCHWARTZ, MARTIN 2006. 'On Haoma, and its liturgy in the Gathas', in A. Panaino & A. Piras (Eds.), Proceedings of the 5th Conference of the Societas Iranologica Europæa, Volume 1, Milan: Mimesis. 215–224.
SCHWARTZ, MARTIN 2008. 'Iranian *l, and some Persian and Zaza etymologies', *Iran and the Caucasus* 12(1). 281–287.
SCHWARTZ, MARTIN 2010. 'On Rashnu's scales and the Chinvant's bridge, with etymological appendices', *Studia Asiatica* 11(1–2). 99–104.
SCHWARZSCHILD, LOUISE A. 1960. 'Review of Un Editto Bilingue Greco-Aramaico di Asoka by G. Pugliese Carratelli and G. Levi Della Vida', *Journal of the American Oriental Society* 80(2). 155–157.

Shringarpure, Suyash & Eric P. Xing. 2009. 'mStruct: Inference of population structure in light of both genetic admixing and allele mutations', *Genetics* 182(2). 575–593.
Sims-Williams, Nicholas 1989. 'Eastern Middle Iranian', in Rüdiger Schmitt (Ed.), *Compendium Linguarum Iranicarum*. Wiesbaden: Dr. Ludwig Reichert Verlag. 165–172.
Sims-Williams, Nicholas 1996. 'Eastern Iranian languages', *Encyclopædia Iranica* 7(6). 649–652.
Skjærvø, Prods Oktor 1983. 'Farnah-: mot mède en vieux-perse?', *Bulletin de la Société de Linguistique de Paris* 78 (1). 241–259.
Skjærvø, Prods Oktor 1988. 'Baškardi', *Encyclopædia Iranica* 3(8). 846–850.
Skjærvø, Prods Oktor 1989. 'Pashto', in Rüdiger Schmitt (Ed.), *Compendium Linguarum Iranicarum*. Wiesbaden: Ludwig Reichert Verlag. 384–410.
Skjærvø, Prods Oktor 2009. 'Old Iranian', in G. Windfuhr (Ed.), *The Iranian languages*. London: Routledge. 43–195.
Soane, E. B. 1913. *Grammar of the Kurmanji or Kurdish language*. London: Luzac.
Steingass, Francis Joseph 1892. *A comprehensive Persian-English dictionary, including the Arabic words and phrases to be met with in Persian literature*. London: Routledge & K. Paul.
Stilo, Donald 1981. 'The Tati language group in the sociolinguistic context of northwestern Iran and Transcaucasia', *Iranian Studies* 14(3–4). 137–187.
Stilo, Donald 2004. *Vafsi folk tales*. Wiesbaden: Dr. Ludwig Reichert Verlag.
Stilo, Donald 2005. 'Iranian as a buffer zone between the universal typologies of Turkic and Semitic', in Eva Csató, Bo Isaksson, & Carina Jahani (Eds.), *Linguistic convergence and areal diffusion. Case Studies from Iranian, Semitic and Turkic*. London: Routledge. 35–63.
Stilo, Donald 2007. 'Isfahan xix. Jewish dialect', *Encyclopædia Iranica* 14(1). 77–84.
Stilo, Donald 2018. 'Numeral classifier systems in the Araxes-Iran linguistic area', in William B. McGregor and Søren Wichmann (Eds.), *The diachrony of classification systems*. Amsterdam: Benjamins. 135–164.
Stollenwerk, Debra A. 1986. 'Word frequency and dialect borrowing', *Ohio State University Working Papers in Linguistics* 34(1). 133–141.
Syrjänen, Kaj, Terhi Honkola, Jyri Lehtinen, Antti Leino, & Outi Vesakoski. 2016. 'Applying population genetic approaches within languages: Finnish dialects as linguistic populations', *Language Dynamics and Change* 6(1). 235–283.
Tafazzoli, Ahmad 1974. 'Pahlavica II', *Acta Orientalia* 36(1). 113–123.
Tavernier, Jan 2007. *Iranica in the Achaemenid period (ca. 550–330 B.C.): Lexicon of old Iranian proper names and loanwords*. Leuven: Peeters.
Tedesco, Paul 1921. 'Dialektologie der westiranischen Turfantexte', *Le Monde Oriental* 15(1). 184–258.
Teh, Yee Whye, Michael I. Jordan, Matthew J. Beal, & David M. Blei. 2005. 'Sharing clusters among related groups: Hierarchical Dirichlet processes', in Yair Weiss, Bernhard Schölkopf, & John Platt (Eds.), *Advances in neural information processing systems*. Cambridge, MA: MIT Press, pp. 1385–1392.
Teh, Yee Whye, Michael I. Jordan, Matthew J. Beal, & David M. Blei. 2006. 'Hierarchical Dirichlet processes', *Journal of the American Statistical Association* 101(476). 1566–1581.
Thackston, Wheeler M. 2006. Kurmanji Kurdish. a reference grammar with selected readings. Manuscript, Harvard University.
Thomas, Bertram 1930. *The Kumzari dialect of the Shihuh tribe, Arabia and a vocabulary*. London: The Royal Asiatic society.
Vahman, Fereydun & Garnik Asatrian. 2002. *Notes on the Language and Ethnography of the Zoroastrians of Yazd*, Volume 85 of Historisk-filosofiske Meddelelser. Copenhagen: The Royal Danish Academy of Sciences and Letters.
van der Wal Anonby, Christina 2015. 'A grammar of Kumzari: A mixed Perso-Arabian language of Oman', Ph.D. thesis, Rijksuniversiteit te Leiden.
Weiss, Michael 2009. *Ann Arbor*, MI: Beech Stave Press.
Wells, John 1982. *Accents of English 1: An introduction*. Cambridge: Cambridge University Press.
Wendtland, Antje 2009. 'The position of the Pamir languages within East Iranian', *Orientalia Suecana* 58(1). 172–188.
Wieling, Martijn, John Nerbonne, & R. Harald Baayen. 2011. 'Quantitative social dialectology: Explaining linguistic variation geographically and socially', *PloS one* 6(9). e23613.
Williamson, Sinead, Chong Wang, Katherine A. Heller & David M. Blei 2010. 'The IBP compound Dirichlet process and its application to focused topic modeling', in Johannes Fürnkranz, & Thorsten Joachims (Eds.), Proceedings of the 27th International Conference on Machine Learning. Haifa, Israel. New York: Association for Computing Machinery.
Windfuhr, Gernot 1991. 'Central dialects', *Encyclopædia Iranica* 5(3). 242–252.
Windfuhr, Gernot 2009. 'Dialectology and topics', in Gernot Windfuhr (Ed.), *The Iranian languages*. London: Routledge. 5–42.
Yar-shater, Ehsan 1969. *A grammar of Southern Tati dialects*. Number 1 in Median Dialect Studies. The Hague: Mouton.
Yarshater, Ehsan 1962. 'The Tati dialects of Ramand', in Walter B. Henning and Ehsan Yarshater (Eds.), *A locust's leg: studies in honor of S.H. Taqizadeh*. London: Percy Lund, Humphries & Co. 240–245.
Zehnder, Thomas 1999. *Atharvaveda-Paippalāda*, Buch 2, Text, Übersetzung, Kommentar: eine Sammlung altindischer Zaubersprüche vom Beginn des 1. Jahrtausends v. Chr. Idstein: Schulz-Kirchner.
Žukovskij, Valentin Aleksejevič 1888–1922. 'Materialy dlja izučenija persidskix narečij. St. Petersburg: Akademija Nauk. 3 vols.