ESTABLISHING THE SUBSTRATE OF THE SINDHI LANGUAGE: THE UNATTESTED LANGUAGE OF THE SOUTHERN INDUS RIVER VALLEY CIVILIZATION

Vikhyat Kethamukkala
Dr. Michael Witzel[1]

Abstract

The enumerated number of languages that have influenced the South Asian language of Sindhi is attestable through its unique situation at the epicenter of many linguistically-diverse regions, ranging from Dravidian (North) and Iranian, to Indo-Aryan dialect regions and centralized languages. While we know much about the language now, its past paradigms are unattested and remain enigmatic, due to the language's rather multilingual situation and tedious morphological receivings from a plethora of languages, from Perso-Arab to Germanic (primarily English). This paper aims to bring some of our unique findings to light as we unravelled the etymological intricacies of the language through a delicate mix of computer programming, online registers, as well as elements of natural language processing and ample amounts of linguistic insight. Hence, the purpose of this study: to understand the languages that influenced the Sindhi language, understanding those components and their importance in shaping the language, and finally reaching the most substantive language that may be related to the shrouded language of the Southern Indus.

INTRODUCTION

The Sindhi language itself is a linguistic amalgam of multiple vernacular dialects, as well as major languages due to its situation in between multiple Indian and Middle Eastern areas. Dr. Lachman M. Khubchandani details the Sindhi linguistic-geographic minutia as "Balochi (an Iranian language) and Brahui (a North Dravidian language) in the west, Pashto (another Iranian language) in the

[1] ⋆ I'm very grateful to Dr. Michael Witzel of Harvard University with whom I've been **collaborating** for ample amounts of patience during the research and writing processes, as well as the revelatory insight and practical advice.

north, Multani and Bahawalpuri (Lahnda dialects, northwest Indo-Aryan) in the north and northeast, Marwari (a Rajasthani dialect, central Indo-Aryan) in the east, and Gujarati (central Indo-Aryan) in the south" (Cardona and Jain 2003, p. 684). Upon further inspection, we learn that the Sindhi language is also characterized by six major dialects, namely:

1) Saraiki, spoken in Upper Sindh (Siro)
2) Vicholi, in Vicholo (Central Sindh)
3) Lari, in Laru (Lower Sindh)
4) Lasi, in Lasa B'elo (spoken in a part of Balochistan called Kohistan in the west of Sindh)
5) Thari/Thareli, in Tharu (the desert region on the southeast border of Sindh and a part of the Jaisalmer district in Rajasthan)
6) Kachhi, in the Kutch region and in a part of Kathiawar in Gujarat (on the southerneastern side of Sindh)

Vicholi is considered the main, standard dialect by Sindhi speakers, and is also used for national administration, education, and more. Politically, the Kachholi language is mutually-intelligible to the standard Vicholi dialect, but that is a stone rather left unturned, as those two regions of Sindh have grown separately as units. Linguistically, Vicholi is also a prestige dialect due to its higher nature of social acceptance and national importance, which is also another reason for the Kutch-Vicholi divide. This [divide] also parallels the political and geographic

demarcation of India and Pakistan; the Kachholi speakers and people yearn for a separate cultural identity, not so much unlike Pakistan in the 1940's.

More can be said about the India-Pakistan divide in 1947, and how it affected the Sindhi population. The migration of the Hindu population greatly strained the Sindhi vernaculars, as well as their cultural ties and demographic stability. While efforts for better external communications have been planned, there is no doubt that these events [of the India-Pakistan dissolution] changed the dynamic social structure of the Sindh region. However, with recent strides in technology, as well as continued linguistic mixing, Sindh has transformed into a multilingual melting pot of various regional dialects and languages, quite the callback to the historical circumstances of the Pre-Sindhi language(s).

We can clearly see the turbulent history of Sindh, but its importance in South Asia cannot be underestimated. If we see the Sindh region geographically, its placement can be seen as a *gateway* of sorts (Cardona and Jain 2003, p. 684), used by multiple invading nations to enter into the Indian subcontinent (*See Figure 1*).

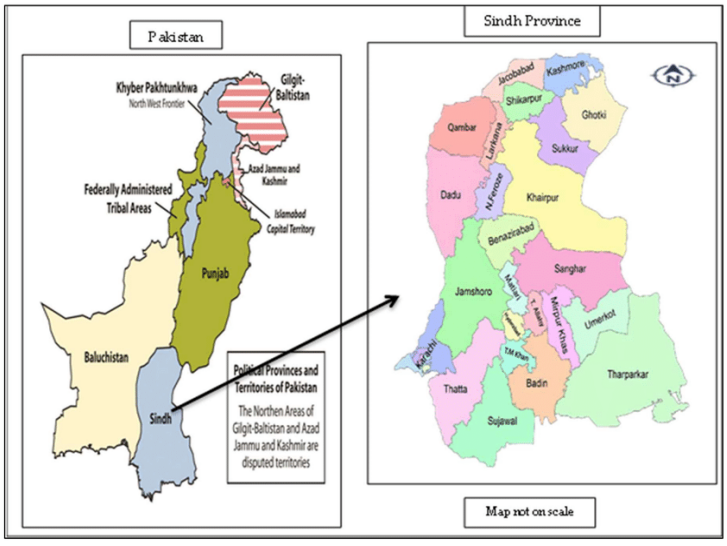

Figure 1: Geographic Situation of Sindh Province (relative to Pakistan)
Due Credit To: "*Development Framework for Agro-Based Industries in Secondary Cities of Sindh Province, Pakistan: SWOT Analysis of Ten-Year Perspective and Medium-Term Development Framework Plans*" (Kalwar, Dali, Hassan 2018)

This in turn has also made the Sindhi language (as a collective, single language) highly susceptible to loanwords and morphological borrowing. These invading nations included the Scythians, Greeks, Arabs, Indo-Aryans, and many more. After the invasion of Sindh by the Arabs (c. 711 by Muhammad ibn Qasim al-Thaqafi), the syntax and vocabulary of the Sindhi were changed over time, slowly absorbing new Persian and Arab morphosyntactic paradigms into the existing linguistic systems at that time. The evolution/transformation of the Sindhi language did not stop there however. Effects of British imperialism and industrial advancements (as well as the Sikh genesis/Bhakthi movement gaining traction; this meant that Sanskrit and Hindi-Urdu words would slowly assimilate into the language) did not stop the language trade; rather, the influence of the Sanskrit, Hindi, Perso-Arab, and Germanic (English) languages only increased, thus catapulting the Sindhi language and people into a new period of cosmpolitical associations between India, Pakistan, and many other nations. Following the divergence of Pakistan and India into the late 1990's, The Sindhi language leans in multiple directions, primarily towards it's Perso-Arab roots (written with Arabic Naskh script), while also retaining many of its constituents (Indian migrants) favor

Sanskriticized Sindhi (written in Sindhi-Devanagiri in some cases), a phenomenon known as *tatsama*.[2]

SINDHI IN PREHISTORY

While there is much that we know about the Sindhi language currently and its current state of affairs, Pre-Sindhi and its origins remain clouded in uncertainty. *Proto-Sindhian* (or Pre-Sindhi) is a language associated with prehistoric Sindh, which was discovered to have been a major nation/tribe (third millennium BC) in the Southern Indus, as revealed by the Mohen-jo-daro excavations. While the epigraphy of the seals and clay tablets found at the site are yet to be deciphered and translated, many hypotheses have still been made by philologists and linguists alike surrounding the role of Sindh, as well as its relationships with the other Indo-Aryan languages, Prakrit, Sanskrit, and more. As Dr. Gregory L. Possehl describes it, "It is one of the unsolved puzzles regarding the peoples of the ancient cities of the Indus" (Possehl 1996).

Many language families have been thrown into the mix regarding the relative closeness to the Indus inscriptions, from Dravidian and Sumerian, to claims as far as Malayo-Polynesian. Despite the sheer level of discrepancies and disparity as to where the language would have derived from, one fact remains clear among most scientists: the language/symbols of the alleged Southern Indus are arranged in *right-to-left* fashion (Farmer, Sproat,Witzel 2014).

---

[2] Skt *tasya* + *sama* = *tatsama* (the same [words] as Sanskrit and Prākrit); A process by which many modern Indo-Aryan languages adapt Sanskrit vocabulary and loanwords

While much has been considered on the grounds of (historical) linguistics and epigraphy, much more can be said about the possible cultural connections that Mohen-jo-daro holds with other ancient civilizations, like Mesopotamia. Ergo, we cannot leave Sumerian out of the equation, as suspected by many of the earlier archaeologists on the project (cf. findings of Gad 1931, Petrie 1932, Hunter 1934, cited in Possehl 1996). From that stance, some other points were made about possible Dravidian, Proto-Sanskrit, and Indo-Iranian links between the style of the Harappans.

Linguistically, it is known that Sindhi is a member of the Indo-Aryan family of languages (Grierson 1919). It shares many resemblances with some antiquated Prākrits as well with the more modern Indo-Aryan languages, like Hindi. These relationships are expressed by Dr. Ernest Trumpp: "Sindhi has remained steady in the first stage of decomposition after the old Prakrit, whereas all other cognate dialects have sunk some degrees deeper. The rules which the Prakrit grammarian Kramdishvara has laid down in reference to the Apabhramsha (Vracada)[3] are still recognizable in present Sindhi, which by no means can be stated of the other dialects. Sindhi has thus become an independent language, which, though sharing a common origin with its sister tongues is materially very different from them." (Trumpp 2005, Sindhishaan Vol. IV Issue IV).

[3] Apabhramsha is referred to as a later form of Prākrit, and is lower in prestige than Sanskrit. Vracada is a dialect of Apabhramsha; little is known about Vracada itself, except that it has some peculiarities relative to other dialects of Apabhramsha, as noted by Markandeya.

From Trumpp's perspective, Sindhi is a highly Sanskritic language, resemblant to the Prākrits of old, and maintaining their grammatical forms and systems to a fairly large degree, but what do others have to say about this?

Dr. Richard Pischel, in accordance with Trumpp, believes that Sindhi, at least historically, was directly derived from the Vracada Apabhramsha, and is referred as the 'kingdom of Sindhu', or *Sindhu-deśa*. Thus, we come across a very distinct phonological study of the Sindhi language which finds many aspects of pronunciation and word formation that are similar to those of the Dardic languages, such as Kashmiri and the somewhat reminiscent Gāndhārī Prākrit.

While many other studies related the Pre-Sindhi to language families like Semitic (to prove an idiosyncratic ancestry to the language that is different from other Indo-Aryan languages) as well as the evolution of the language through the Indus Civilization and reaching the form of *Old Sindhi*.

Beyond all of this however, we have strong linguistic evidence that Sindhi is an Indo-Aryan language, due to the mass of works that include the language under Indo-Aryanism, and the correspondences that this language has with its sister languages and close counterpart, *Lahnda* (a notable Northwestern Prākrit/Sanskrit-descendant language that is often bracketed with Sindhi).

PROCEDURE

The process itself involves multiple steps and considerations, almost as though we are peeling back the layers of a vegetable. Starting from the least effective, being English and other possible Germanic influence. These languages

have affected Sindhi more contemporarily, and were probably not present historically to account for much of the morphology and vocabulary of the language. The second layer goes to Perso-Arab loans from the 8th century, which should account for a decent number of the sample size (of *2013* words). The third layer is attributed to Indo-Aryan influence, especially from Sanskrit *tatsama* and Hindi-Urdu. This layer should account for the majority of the entries. Extra layers may include the 'Primary Prākrit' or the unattested Old Sindhi Prākrit, and/or words of Dravidian influence (as Sindh was occupied by a faction of Dravidian language speakers in 4000 BC, which split off into the Brahui tribe[4] in Balochistan). The language that remains after these etymological dissections should be the original Sindhi language, and the theoretical language of the Indus.

To begin, we have to determine the level of influence that each of those suggested language groups had on Sindhi, and make that into useful statistical data that we can compile and put to use at a later time.

We started with the Parmanand Mewaram's *Sindhi-English Dictionary* (1910)[5], which has been our main text from a lexicographic standpoint. The digital version which we received has 2013 entries and was kindly converted into a human-friendly format, but its readability needed some modification.

For this, a Python program that would not only sort through the XML file, but also transcribe its Roman equivalent with the original values, was necessary.

---

[4] A theory exists that the Brahui speakers migrated into Sindh from Central India around 1000 BC (Elenbein 1981)

[5] We received a XML file from Dr. Jim Nye, the creator of the *Dictionary of South Asian Languages (DSAL)*, a major archive and online dictionary. However, the data in the file (formatting-wise) would not match exactly with DSAL, and thus creating some translation issues in our data.

For this, we had to use an UTF-8 (Unicode Transformation Format) variable that would recognize both the Arabic characters and Roman characters in the file. Due to a corruption in the file however, a catch, or a sequence by which exceptions are handled, was needed to split the file at the line of the glitch. Having finally finished coding that segment, the file was converted into an Excel (.xls) file for easier manipulation and viewability (*See Figure 2*).

Term	Part of Speech	Inflections	Senses	Origin
ubbaararnu اُبارَنُ	verb	ubbaaru(imperative), ubbaariyo(Past Participle)	to boil, heat, rise	Indo-Aryan
baati باتِ	noun		speech, language; subject, matter	Perso-Arab Indo-Aryan
iishvaru ایِشوَرُ	noun		the manifested God (among Hindus), God	Indo-Aryan
pardesu پَردیسُ	noun		Foreign country	Indo-Aryan

Figure 2: Select few entries from Excel-format program,
Table/Program credits to Author

For the second part of the etymological/substrate identification task, we would have to go about some way in which we can compare the values in our Excel-*Mewaram* data with information from some major databases that are up-to-date with linguistic data, definitions, and provide a suitable format for our needs. For us, the choices were clear: for each language family that we want to test, we should choose a singular resource that encompasses the aforementioned needs, or find a large repository that fits all of our needs.

METHODOLOGIES

The modes in which one can define the Sindhian substrate is varied, but not impossible. Through meticulous tests and observations of peculiar linguistic cases amongst our data set of around 2000, a few conclusions can be made, but not without a multi-faceted approach to combat the nuances and technicalities that lie ahead. The paradigms of our work have been centered around three theoretical forms of translation and cross-etymological analysis: *Turner's 'A Comparative Dictionary of the Indo-Aryan Languages' and other associated works/database repositories*, *SARVA Etymological Dictionary,* and *Natural Language Processing through Google API*.

The first method by which we have gone about substrate identification is through cross-translation with Turner's 'A Comparative Dictionary of the Indo-Aryan Languages' (hereafter CDIAL). This procedure involves comparison of the words/values in our Excel readable format with the webpage format of Turner's CDIAL, Burrow and Emeneau's 'A Dravidian Etymological Dictionary', and Platt's 'A Dictionary of Urdu, Classical Hindi, and English'. This prospect, while accurate and versatile, is manually-intensive and meticulous in its current format. It would require programming and coding beyond my skill, but would also be wasteful, rather than translating the words manually. A worthwhile expense, however, is that Turner's CDIAL is particularly well-built for parsing exceptions and specific cases of words, as well as for veritability reasons. While web-scraping is not an efficient

method to receive our necessary data, it may be an option that we revisit in the future.

The second method by which we plan to set the grounds to establish the Sindhi substrate is through SARVA, an etymological/substrate dictionary co-created by Dr. Michael Witzel himself and his associate at the Tokyo University of Foreign Languages (colloquially referred to as TUFL). As stated on its homepage, "the goal of the Project is to assemble all words showing early language contact among the (known and unknown) languages of the subcontinent, in order to provide data for the reconstruction of the history of language contact in the region ... Our ultimate concern is the discovery of patterns of linguistic interaction that will lead to reconstructions of the times, places, and cultural circumstances under which prehistoric language contact took place in the subcontinent." (http://www.aa.tufs.ac.jp/sarva/materials_frame.html)

Through this software, along with its reader friendly-format and simple construction, a way in which we can take the data from the site that is marked with *S.* for Sindhi (ex. aṃśa: S. hañjhī f. ' shoulder -- blade). Through this process of demarcation, we can go through the individual cases of Sindhi substrates, removing and filing through cases of English/Germanic, Indo-Aryan, Persian, Arab, and possible Dravidian influence. Through this process of elimination, we may be able to definitively encounter the substance of the language that was closely associated with the Southern Indus. Whilst the process is just as tedious as the

aforementioned CDIAL parsing method, the results of this study could be just as rewarding, or even more from a linguistic and meritorious vantage.

The seemingly last, more innovative, yet equally tedious process is one through *natural language processing*. While the software medium through which the process can be carried out is faulty at times, this process could prove much more useful and time-saving than the other two options. Through *Google Translate's* API (Application Programming Interface), we can emulate certain characteristics of these pervasive etymological dictionaries. While the programming can be frustrating at times, especially with the logical incompatibility with the Arabic writing system and morphosyntax, these can be alleviated through third-party extensions and applications, as well as some in-built mechanics. The main obstacle that we face with this way of translation and comparison is the approach strategy. Two different criteria have presented themselves over the course of programming and experimenting with such a scheme:

1) 'Per language' is the first proposal for examining the data set and determining the language [family's] of origin. This theory can be actualized through ISO 639-x language codes, which are integral into building natural language systems and processes. While there are a plethora of codes that can be in use, of which include many of the major Indo-Aryan, Germanic/English, Perso-Arab, and Dravidian languages, the basis of Google Translate's inner mechanisms is faulty, and can sometimes experience glitches that render the output

inutilisable. One example of such a case, which I have experienced through numerous trials of debugging, is when the computer recognizes the word (in the Arabic script[6]), recognizes certain characteristics of the word as related to a certain family, and gives back the family, only to read 'Germanic/English' in many cases due to the nature of the program itself. The ineptitude of the parameters of the software make it so that the Arabic script, while programmable and can be manipulated, as well as the systems at play cannot undertake such a task logically.

2) 'Per word' is the second methodology. This concept is more agreeable systematically, and works in conjunction with the systems of Python and translation API well, yet this brings us back to the requisite cornerstones of our investigation, and the precedents and considerations that we had assessed in the aforementioned tactics of substratum perusal and morphological and historical inquiry. The tedium of this task is comparable to the first two modes, and thus, due to the familiarity and linguistic cohesiveness of the SARVA tactic, that is ultimately the mode by which we will progress our methodical research of the enigmatic Pre-Sindhi language.

[6] This can be occasionally problematic as the schema of Google Translate and its input systems do not accept romanisations, diacritics, or anything of that sort of formatting, but only under certain linguistic circumstances, they can (more often with more popular languages like Mandarin Chinese). For a language like Sindhi, which is an obscure and little-studied language, these criteria are not met.

DISCUSSION/**C**OMMENTARY

While the degrees of certitude of each of these methods are concrete and effective in their own rights, the nuanced nature of these paradigms pose both issues in respect to time and commitment to the Sindhian substrate/etymology project, as well as the accountability of each of these methods. Surely, the natural language processing route of etymological and morphological translation is faulty, but the easiest to reconcile with and work around, whilst the tried and tested method of manual etymological analysis of the data set is more meritorious in nature, as well as a more interactive and hands-on approach to the project as a collective effort, yet quite intensive (although, with more hands on deck, the best option)

Another point to consider is that these methods have all been experimented with, at least at a superficial level, especially in the cases of manual translation through Turner's CDIAL and SARVA. Through CDIAL and other affiliated dictionaries/repositories, it's a manual process, although the auto-completion feature on the site speeds up the process. Upon the two aforementioned methodologies, there are two more theoretical methods which I have thought of that could work in our favor as we progress through this project:

1) If temporary access were given to us by the maintenance team of the CDIAL online resource, the source code and hard-coded entries could be used in our favor. Think of this as a simple process of comparison. All one has to do in this situation is take the source entries and

compare with the data in the Excel file, find the matching words in both data sets, and analyze those for the 'language of origin', which is usually in the second position of the description. Once that is done, one can manipulate the format, and then do the same to crosscheck with the SARVA etymological dictionary. In its current form, this method isn't feasible due to protections and web construction of the site, but with further adjustments, massive progress can be made in this regard.

2) The second method is more of a back-up plan for the former, being that if access were granted to the source entries of the dictionary, we could build our own API (Application Programming Interface) for each of the online repositories (CDIAL, UHI [Urdu, Classical Hindi, English], SARVA, etc.), and the program itself could derive exact meanings from the foundational dictionaries themselves. A preparatory step to this however, is to convert all the entries into Unicode diacritics (primarily *macrons*), as the dictionaries recognize those best, probably due to the phonological features of each of those languages when romanized. The prospect of this method is quite optimistic as these API's can be modularized and saved for use in later translational/morphological recognition projects, and surely many other variations of the aforementioned. While the construction of the API itself can be

challenging, with the help of outside developers or professionals, the future of the project, and the results we could receive seem bright!

CONCLUSION **A**ND **C**ONSIDERATIONS

The historical altercations of the Sindhi language lend themselves to a complicated repertoire between the language and its counterparts. Despite these seemingly insurmountable nuances in the languages, and the intellectual gaps that we have yet to fill, my experience with Dr. Witzel has been ever so insightful. While work remains to be done in the areas of human and computer intelligence, translation, and historical research, the hypothesis that Indo-Aryan consists of most of the Proto-Sindhian's lexicology is not farfetched in the least, and we will continue to work towards the goals of both substrate identification and linguistic knowledge, but also towards cultural cognizance and careful considerations of the events that have transpired in the land of Sindh.