Romani Roots: A Lexicostatistical Analysis Of Romani, Hindustani, and Czech

Ryan Shosted

Follow this and additional works at: https://scholarsarchive.byu.edu/dlls
LEXICOSTATISTICS: AN INTRODUCTION

In the past few decades, the possibility of applying mathematical and statistical techniques to language study has captivated the imagination of many linguists, especially those interested in transforming their field into a more “exact” science (Embleton 1986). In the 1950s and 1960s, quantitative linguistic theory had a remarkable incarnation in the work of Morris Swadesh (Crowley 1997). Faced with the tricky and often controversial task of grouping languages into families, Swadesh hypothesized that statistical data could be used to posit accurate subgroups among related languages. Swadesh counted the cognate lexical forms shared by specific languages to determine to what extent the languages were related. He called his subgrouping technique “lexicostatistics.”

Swadesh’s technique is based on several critical assumptions. At the most basic level, the approach presupposes that languages change over time in systematic ways and that similarities among languages are the results of common “genetic” relationships among languages. This genetic analogy makes it possible to speak of parent, daughter, and sister languages (e.g., Romani and Hindustani share a common parent language, Sanskrit, and are thus considered “sister languages”). By applying lexicostatistical techniques, linguists attempt to quantify the rate and extent of lexical change in languages over time and then determine the genetic relatedness of the idioms. In short, lexicostatistics build on the rudimentary family tree model of language typology. While the hierarchical, genetic structure of the family tree model is maintained, lexicostatistical methods allow the linguist to organize language subgroups around percentage values of relatedness (e.g., languages with a determined relatedness of 81–100% are “dialects of a language,” languages with a determined relatedness of 36–81% are “languages of a family,” and so forth—for a full listing, see table 1). The higher the value, the more closely the languages are related. The method for determining this value of relatedness will be discussed shortly.

Lexicostatistics hinge on the theoretical proposition that languages change in regular ways and that the rate of lexical change remains constant over time. Swadesh’s model of lexical change is thus analogous to the scientific model of radioactive decay (Embleton 1986). Just as the presence of a radioactive isotope diminishes at a fixed rate over time, so does the content of a language’s “original” lexicon. In reality, words do not disappear when, statistically speaking, their time runs out. Instead, new lexical forms constantly compete with old forms, rendering them obsolete and eliminating them from a speech community’s lexical repertoire. Though lexical change may seem like an uncontrollable process dependent on a wide range of variables, Swadesh nevertheless hypothesized that vocabulary turnover

DLLS 2000
occurs at a diachronically fixed rate. The third assumption of lexicostatistics has to do with the actual lexical forms under analysis—the data used to determine relatedness among languages. Swadesh supports the distinction made between “core” and “peripheral” vocabularies, and incorporates this distinction into his theory. According to Crowley (1997), core vocabulary consists of forms like “tooth” and “tongue” that are common to the human experience and therefore show up almost universally in the world’s languages. Peripheral vocabulary forms, like “antelope” and “cilantro” cannot be found in every language. The presence of peripheral forms like these is directly tied to the unique characteristics of a linguistic community’s physical environment and material culture.

Swadesh believes that core vocabulary words most accurately represent the original lexical stock of a language because of their universality among languages and because of the decreased possibility that borrowed lexical items will replace them. For example, it is likely that the native Romani words associated with metalworking were replaced by Romanian terms when the Roma reached the Balkans, while it is much less likely that the Romani word for “and” would be replaced by a nonnative term. Likewise, it is not profitable to factor the peripheral term for “automobile” into a lexicostatistical comparison of Romani and Hindustani, as both languages have probably copied the term from English, albeit with various phonological adaptations. By contrast, a comparison of the core word for “rope” in both languages may reveal a shared genetic relationship.

With these theoretical considerations in mind, our focus can now turn to the techniques involved in lexicostatistical analysis. If one is going to compare core vocabulary items from two or more languages and thereby determine the relatedness of these languages, it is necessary to decide first which words may be reliably classified as “core” and which ones must be discarded as “peripheral.” Swadesh’s catalog of roughly two hundred lexical items is generally accepted as the standard list of core vocabulary (Crowley 1997). This so-called “Swadesh list” is expansive enough that some words may be deleted if they are determined functionally peripheral in a specific context (e.g., “snow” and “frost” may be excluded when working with languages native to equatorial regions).

Once the linguistic equivalent for each word is posited, the analyst must then decide which words are cognates and which words are not. I will discuss the problematic nature of this process later on in my own case study of Romani, Czech, and Hindustani. After counting the cognates, determining the lexicostatistical relationship of the languages is only a matter of arithmetic. Dividing the total number of cognate pairs (e.g., Hindustani [tang] and Romani āang, “foot”) by the total number of words in the Swadesh list will produce a percentage (“shared cognate percentage” or SCP) of linguistic relatedness. The SCP may then be correlated with the levels of

<table>
<thead>
<tr>
<th>Level of subgrouping</th>
<th>SCP in core vocabulary</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dialects of a language</td>
<td>81-100%</td>
</tr>
<tr>
<td>Languages of a family</td>
<td>36-81%</td>
</tr>
<tr>
<td>Families of a stock</td>
<td>12-36%</td>
</tr>
<tr>
<td>Stocks of a microphylum</td>
<td>4-12%</td>
</tr>
<tr>
<td>Microphyla of a mesophylum</td>
<td>1-4%</td>
</tr>
<tr>
<td>Mesophyla of a macrophylum</td>
<td>0-1%</td>
</tr>
</tbody>
</table>
linguistic subgrouping found in the following taxonomy (Swadesh cited in Crowley 1997, 173):

Once the SCP is calculated, a mathematical formula based on the equation "rate of change x time = total change" allowed Swadesh to posit the historical time-depth at which two related languages became linguistically distinct (or "branched apart," according to the family tree model). This technique is referred to as glottochronology. Swadesh's basic glottochronological formula states: where \( t \) represents the number of thousands of years that two languages have been linguistically distinct, \( C \) stands for the SCP, and \( r \) is the constant rate of lexical change (cited in Crowley 1997, 181). The logarithmic base for the equation is 0.368.

Before the formula could be of any practical use, Swadesh had to determine the constant rate of lexical change in language—the variable \( r \). Swadesh studied the vocabulary change in the written records of 13 languages with a literary tradition of at least one millennium. He discovered that after 1,000 years a language loses roughly one-fifth of its core vocabulary (Crowley 1997).

Though Swadesh's model won many adherents in the 1950s and 1960s, it has suffered a considerable barrage of criticism in more recent years (Embleton 1986). In his review of Swadesh's work, Crowley (1997) disagrees with several of the underlying assumptions and methods of lexicostatistics. First, he objects to the idea that lexical change occurs at a fixed rate in all languages. Second, he criticizes the lack of academic rigor that often marks the determination of what is a cognate and what is not. Crowley claims that he himself has performed lexicostatistical cognate counts on word lists previously analyzed by other linguists. In some cases, Crowley's SCP findings were 10–20% different from the findings of previous researchers (1997, p. 183). Finally, Crowley challenges the basic notion of how to define core vocabulary. Still, he does not disregard Swadesh's model entirely. In fact, Crowley devotes fifteen pages of his textbook on historical-comparative linguistics to a description of lexicostatistics and glottochronology, including numerous exercises that call for the manipulation of lexicostatistical data. Hence, it can be assumed that while lexicostatistics and glottochronology are by no means perfect instruments, they still constitute viable quantitative techniques for analyzing the genetic relationships among languages.

**STATEMENT OF PURPOSE AND HYPOTHESES**

In conducting the present lexicostatistical comparison of Romani with Hindustani and Czech, my purposes are twofold. First, I hope to vindicate Swadesh's theory of lexicostatistics, at least in part, by demonstrating that the application of his techniques can verify conventional linguistic knowledge about the subgrouping of languages. Hindustani, an Indic language spoken in the northern regions of India, has traditionally been classified as a sister language of Romani. Both are considered daughter languages of Sanskrit. Czech, on the other hand, may be considered Romani's linguistic cousin; the two languages share the reconstructed Proto-Indo-European language as a common ancestor. In the present study, Czech will be compared to Romani, and Romani will be compared to Hindustani; for the sake of focus and space, Czech will not be compared to Hindustani. My second purpose in conducting this study has to do with the debate over the real origins of Romani. I am interested in compiling Romani linguistic data to combat the unfortunate and anachronistic hypothesis that Romani is not an Indic language at all, but rather a "secret code" developed among European Gypsies and wandering criminals (Davidová 1995).
The following hypotheses, based on Swadesh's taxonomy (table 1 above) will be tested against the lexicostatistical data:

H_1. As languages of a common "stock" (they are Slavic and Indic descendants of Proto-Indo-European), Czech and Romani should have an SCP of between 12% and 36%.

H_2. As languages of a common "family" (they are both the daughter languages of Sanskrit), Romani and Hindustani should have an SCP of between 36% and 81%.

**METHODS**

To organize the relevant lexicostatistical data, I used Microsoft Excel to create a six-column table (see the sample in table 2). The first column of the table contains 216 English words duplicated from the original Swadesh list (Crowley 1997, 174). The next column contains the lexical equivalents of these words in Czech. The following comparison column, marked Cog C/R, contains a number that indicates the relationship of the Czech word to its Romani equivalent (the Romani gloss is found in the subsequent column). In this comparison column, 0 = no cognate relationship; 1 = cognate relationship; 2 = lexical borrowing; 3 = no gloss available (i.e., in a few instances, one of the dictionaries had no entry for the word in question). The Romani and Hindustani correspondences are set up according to the same numeric system, under the column heading Cog R/H, with the Romani and Hindustani glosses appearing on either side. Hence, the cross-linguistic analysis for the English word "sister" looks like the sample in table 2.

Czech *sestra* is not cognate with Romani *phen*, so the Cog C/R column contains a 0 (= no cognate relationship). The Romani term, however, is cognate with Hindustani [bva:n], so a 1 (= cognate relationship) is placed in the Cog R/H column.

Due to the character limitations of Microsoft Excel, the following modifications have been made to Czech, Romani, and Hindustani (romanized) orthography for use in the table. In Czech and Romani words, a (^) appears after the letter over which it would normally appear in the form of a haãek (Cz. hvizda = hve^zda). For the Hindustani words, a colon (:) is used to indicate length.

After the number of 1s (cognates), 2s (loans), and 3s (insufficient information) in one column were counted, the total number of cognates was divided by 216 (minus the number of loans and words for which sufficient information could not be located). The equation for finding the shared cognate percentage (SCP) may be stated simply as where C represents the SCP, Cog equals the total number of cognates, L is the total number of loans, and G stands for the total number of words for which no gloss was available. Crowley (1997) explains that it is necessary to reduce the denominator (i.e., the total number of words in the Swadesh list) by a factor of 1 whenever a loan word is encountered. If loan words are erroneously counted as cognate forms, they will strengthen the posited cognate relationship of the languages and thus skew the final results of the analysis. In the present study, the equation was applied to compare Czech with Romani and Romani with Hindustani. By plugging the numbers into the glottochronological formula mentioned earlier, the time-depths of linguistic divergence among these languages were also ascertained.

The lexical data in this study comes from Poldauf's *Anglicko-C^ešký C^esko-

<table>
<thead>
<tr>
<th>English</th>
<th>Czech</th>
<th>Cog C/R</th>
<th>Romani</th>
<th>Cog R/H</th>
<th>Hindustani</th>
</tr>
</thead>
<tbody>
<tr>
<td>sister</td>
<td>sestra</td>
<td>0</td>
<td>phen</td>
<td>1</td>
<td>bva:n</td>
</tr>
</tbody>
</table>

Table 2. Sample of Cross-Linguistic Analysis
Anglický slovník (1981), Húbschmannová et al.'s Romsko-Ceský C^esko-Romský kapesní slovník (1991), and Forbes' English-Hindustani, Hindustani-English Dictionary (1866). In the Discussion section, the information about Hindustani and Romani lexical items will be supplemented with glosses from dictionaries of Hindi (another sister language) and Sanskrit (Raker 1995; Benfey 1886). It should also be noted that the Romsko-C^eský C^esko-Romskýkapesní slovník represents the lexicon of Czechoslovak Romani, an individual dialect among many European variants of Romani. According to Davidová, the various dialects of Romani tend to diverge most ly in terms of their peripheral vocabulary (1995). Hence, for the purposes of this general inquiry into Romani core vocabulary, Czechoslovak Romani is considered to be representative of Romani as a whole.

**RESULTS**

After the lexical data into the Microsoft Excel spreadsheet was entered, the cognate/non-cognate forms were evaluated, and the numeric data was tabulated, the results proved the validity of Hypotheses 1 and 2. For Romani and Czech, the lexicostastical data produced an SCP of 23.077%, well within the 12-36% range that Swadesh posited to designate "families of a stock." The results were also encouraging with respect to Romani and Hindustani, which demonstrated an SCP of 53.774%. According to Swadesh, this numerical value designates Romani and Hindustani as "languages of a family." The glottochronological calculations indicated that Romani and Hindustani separated some 1,430 years ago. According to the study, the paths of Czech and Romani (i.e., Slavic and Indic) diverged some 3,380 years ago. Glottochronological results are traditionally rounded to the nearest thousand, thus we have 1,000 years and 3,000 years, respectively, for the Romani-Hindustani and Indic-Slavic splits.

**DISCUSSION**

Lexicostatistical analysis is not a flawless instrument for determining relationships among languages, but that does not mean that the flaws cannot be ameliorated with conscientious research. Crowley's criticisms of lexicostatistics become poignant when one actually starts assembling lexical data and trying to determine cognate relationships among words. The analyst is soon confronted with the responsibility of declaring relationships among words, which are not always easily identified as cognates. For example, Hindustani [sina:] and Romani sivel (to sew) have only two segments in common: [s] and [i]. Is it likely that [na:] and [vel] descended from some common protoform? How likely is a sound change that produces the phonetically dissimilar [n] in one language and [v] in another? Without some knowledge of Hindustani and Romani morphology, the analyst may be tempted to mark [sina:] and sivel as noncognates. However, if the analyst knows that [na:] is the Hindustani infinitive marker, and [vel] is its Romani counterpart (which is indeed the case), then the root of both words can be recognized as si-, and the cognate relationship between the two is hard to deny.

At this point, however, another problem arises: since the two forms are so similar, is it safe to conjecture that Romani adapted the root si- to its own verbal paradigm after borrowing it from Hindustani? Intuitively, this seems improbable, since loan words often fill in lexical or conceptual gaps in language. Was "sewing technology" imported to Romani culture because the Roma themselves had not developed it? Again, one intuitively reasons that, since sewing is prevalent among all the cultures of the
world, the Roma probably did not borrow the technology from their Hindustani neighbors, and that means they probably had their own lexeme for what to do with a needle and a thread. Despite its intuitive appeal, however, this line of reasoning is not altogether sound, as will be demonstrated shortly. To determine empirically whether or not si- is a lexical borrowing, it would be helpful to find the gloss for "to sew" in Sanskrit, the parent language of both Romani and Hindustani. The instantiation of Sanskrit si for sew proves the case.

Lexical borrowing is very important to the final results of lexicostatistical research. If not recognized for what they are, borrowings will indicate a false relationship between the languages under question and will ultimately skew the results of any lexicostatistical study. For example, Romani dychinel is probably a borrowing from Czech dýchat. This may come as a surprise, since both words mean "to breathe." Does this suggest that the Roma imported breathing from their Slavic neighbors? Obviously, the "lexical gap" argument that worked in the case of "to sew" does not work in the case of "to breathe." So how do we know that this is an example of borrowing and not an indication of the close linguistic relationship between Czech and Romani? First of all, one may observe that the match is "too good to be true" because of the [x] instantiated in both words. The voiceless velar fricative [x] (represented as ch orthographically) is a phoneme native to Slavic but not to Indic languages. It is unlikely that [x] would develop in Romani when it does not appear in Hindustani, Hindi, or Sanskrit—unless the forces of language contact intervened. Forced by social expediency to become bilingual in a Slavic language as well as Romani, the Roma of Eastern Europe learned how to articulate [x] in Slavic words. As an areal feature, [x] gradually became a native phoneme of Romani and Roma children acquired it naturally. However, to prove definitively that dychinel is in fact a loan word and not the product of a natural sound change, it is necessary to examine the variations of the root [di] in Sanskrit and Hindi. After observing that none of these words relate to breathing, it is safe to posit that dychinel is indeed a borrowing.

As this brief discussion has demonstrated, determining the cognate relationship between two words is not as simple as it sounds. The process quickly becomes complicated, requiring the analyst to consider multiple variables such as phonology, sound change, the probability that a lexeme would be copied from one language to the other, and the presence of similar words in parent or sister languages. Admittedly, this is an inherently messy task. The variables cannot be controlled with precision. It is safe to argue, however, that since Romani is by no means a pidgin or creole language, the number of its linguistic borrowings (especially in core vocabulary) will be limited. In fact, the results of this study indicate that the core vocabulary percentage of loans from Czech to Romani is only 8.333%. Due to the contrasting phonological systems of Slavic and Indic languages, it is much easier to identify Czech loans to Romani than it is to identify Hindustani loans to Romani.

Crowley (1997) identifies another valid criticism of lexicostatistics by alluding to the following scenario: Suppose one linguist calculates an SCP 35% for languages A and B while another linguist posits SCP 36% for the same pair of languages. In terms of Swadesh’s taxonomy of subgroupings (table 1), the first linguist must declare that A and B represent families of a stock, while the second must argue that A and B are really languages of a family. Can lexicostatistics be considered a valid technique if a 1% discrepancy generates totally different subgroupings in the two nearly identical studies? What if that 1% difference was the result of a lexical borrowing incorrectly classified as
a cognate? In light of this possibility, Swadesh's rigid, percentage-based taxonomy of subgrouping may need to be revised. One possibility for revision may include adding a margin of error (±5%) to the percentages Swadesh has laid out to group language relationships. After all, it should be remembered that even languages within the same level of subgrouping (e.g., families of a language) share different SCPs because they have different histories of language contact and language change. It may be profitable to visualize Swadesh's taxonomy as a spectrum composed of gradations in language relatedness, and not as a hierarchy arranged around discrete values. Whatever the case, in the present study, each pair of words marked as cognates increases the overall SCP by only 0.463%.

In addition to providing enlightening insights into the theory and method of lexicostatistics, the present study revealed some interesting facts about the relationship between Romani, Czech, and Hindustani. The remainder of this section will concentrate on these linguistic findings.

Perhaps the most interesting lexical relationship in the study is represented by the Os (= no cognate relationship) in the column between Romani and Hindustani words. The Os in this column indicate Romani words that are neither related to Hindustani nor borrowed from Czech. They account for 37.983% of Romani's core vocabulary. Such a large percentage raises the question: Where did these words come from? The origins of these lexemes (the "O-category" words) exemplify the important role phonological change and semantic shift have played in the lexical development of Romani.

For example, several Romani words begin with the phoneme [ts] (represented orthographically as c). Like [x], [ts] is a common segment in Slavic, but it is not native to the Indic languages. Nevertheless, the three Romani words in the Swadesh list that start with [ts] are clearly not Slavic borrowings: Romani cirdel vs. Czech tahnout (to pull), Romani cindo vs. Czech mokry (wet), and Romani cipa vs. Czech kura (bark). The origin of these words becomes even more obscure upon consideration of the seemingly unrelated Hindustani glosses: [k'i:ona:], [c'd'al], and [b'i:ga]. Discovering the origin of the Romani words with initial [ts] depends on a wider linguistic inquiry into Hindi. Though Romani [ts]-initial words correspond to a set of Hindi words that begin in [tar]: [ta:rna:] (to pull) and [tar] (wet). By applying the comparative method (i.e., by collecting and comparing data from several other Indic languages like Bengali and Punjabi), it may be possible to posit whether or not Romani word-initial [ts] developed from a protoform like *[tar]. The evidence so far suggests that there is a systematic connection between some ancient phoneme and the modern Romani [ts]. (Incidentally, the only Hindi lexeme for "bark" that seems phonetically similar to Romani cipa is [dhup] "incense.")

Several of the O-category words demonstrate the influential role of semantic shift in the lexical development of Romani. For example, Romani drom ("road" or "journey") has an interesting history. Drom descends from Sanskrit [dram] "to run" and is probably unrelated to Hindustani [rach]. In Hindi, the few lexical forms with word-initial [dr] have to do with speed, the act of running, or the act of melting. Sanskrit [dravaya] means "to put to flight" and [anudruta] means "pursued." Apart from its very general sense of "path," "road," or "journey," drom is used by the Roma to refer specifically to their centuries-long travels across Asia, Europe, and the Americas. Since the Roma have often been "put to flight" and "pursued" by angry townsfolk throughout their long sojourn, it is interesting to note the meanings of Indic [dr]-initial words.

Sanskrit [dram] "to run" corresponds with Greek dromos [dromos], "a race" or "a run." The meaning of dromos was
polysemous: it could signify both the race and the place where the race was run. In the New Testament, dromos is used in 2 Timothy 4:7 (“I have fought a good fight, I have finished my course, I have kept the faith”) where it metaphorically means “the course of life” (Zodhiates 1993). The Roma, who may have kept Indic [dram] as the cover-term for “running” until they arrived in the Peloponnesus, probably broadened the meaning of the word to signify the concrete nouns “course,” “race,” and “road” due to the influence of Greek dromos. Romanian drum (path) may also have facilitated the category change from the verb [dram] to the noun [drom]. The Greek pronunciation of [dromos] probably caused the Romani vowel shift from [a] in [dram] to [o] in modern [drom]. When Romani [dram] changed to [drom], a lexical gap appeared in the semantic domain of “to run.” To fill the lacuna, the Roma started using the verb prastal, perhaps a distant relative of Hindi [prastha:n] (march).

Another interesting phenomenon that involves phonological change occurred in the case of the modern Romani homonymous pair rat (blood) and rat (night). The presence in Hindi of [rakt] (blood) and [rat:] (night) seems to indicate that the two words were not originally homophones. Sanskrit [ranj] (blood) somehow acquired a [k] in the word-building process—this is evidenced in Sanskrit [raktapala] (blood drinker) and [raktapayan] (leech). Hindi preserves [kt] in [rakt] while Romani simplified the cluster by dropping the [k]. Without the [kt] cluster or vowel length to differentiate the meanings of “blood” and “night,” the two words have become homonyms in Romani.

In the case of Romani pachinel (to smell), a curious semantic shift has taken place. Whereas the Hindustani word for “to smell” [su:nghna:] seems to have a history unrelated to Sanskrit, the Romani word appears to be a semantic extension of Sanskrit [paka], the verbal noun meaning “cooking” or “baking”. Apparently, the Roma came to associate the pleasant aromas of baking with the act of smelling them, thus encoding the verb for smelling with information about what was originally being smelled. In terms of phonological shift, the [k] in [paka] simply lenited to [x] in [pachinel] due to the influence of Slavic phonology as well as the intervocalic environment in which [k] is instantiated.

Similar expositions could be conducted for all of the 0-category Romani words found in the Swadesh list. For the sake of space, however, only a few are listed in table 3. Notice how, even though the Romani words in the table have no Hindi, Sanskrit, or Hindustani equivalents that match them both phonetically and semantically, the Hindi and Sanskrit words in the middle column share some of the same semantic and phonetic space as the Romani 0-category words. This demonstrates the fact that semantic shift has gone on in Romani over time, along with sound changes that additional

<table>
<thead>
<tr>
<th>Romani</th>
<th>Hindi / Sanskrit</th>
<th>Hindustani</th>
</tr>
</thead>
<tbody>
<tr>
<td>khelel ‘to dance’</td>
<td>Hindi [k'elna:] ‘to play’, ‘to stage’</td>
<td>[na:c'na:] ‘to dance’</td>
</tr>
<tr>
<td>tavel ‘to cook’</td>
<td>Hindi [tav] ‘heat’</td>
<td>[paka:] ‘to cook’</td>
</tr>
<tr>
<td>men ‘neck’</td>
<td>Sanskrit [manyaa:] ‘back of the neck’</td>
<td>[gardan] ‘neck’</td>
</tr>
<tr>
<td>briyya ‘rain’</td>
<td>Sanskrit [varya] ‘the rainy season’</td>
<td>[men] ‘rain’</td>
</tr>
<tr>
<td>cacc’o ‘right’, ‘correct’</td>
<td>Hindi [ac^c'h'a] ‘good’</td>
<td>[durust] ‘right’, ‘correct’</td>
</tr>
<tr>
<td>kaif ‘tree’</td>
<td>Hindi [ka:st] ‘cultivation’</td>
<td>[darak’t] ‘tree’</td>
</tr>
<tr>
<td>parno ‘white’</td>
<td>Hindi [pa:rdarsi:] ‘transparent’</td>
<td>[safed] ‘white’</td>
</tr>
</tbody>
</table>
studies might undertake to identify in detail.

Finally, a brief discussion of the glottochronological results of the study is warranted. If the lexicostatistical results were fairly accurate, it stands to reason that the glottochronology of Romani, Hindustani, and Czech should also be valid. According to Cole (1998), archaeological evidence places the Slavs between the Vistula and Dnieper rivers around the seventh century B.C. According to her, it is during this time period that the Slavs may be considered a distinct ethnic group for the first time. The glottochronological data in the present study indicate that the split between the Slavic and Indic language families occurred 3,000 years ago. This figure would give the Slavs plenty of time to trek from the Eurasian Indo-European homeland and set up camp between the Vistula and Dnieper in the seventh century.

Romani historians like Courbet (1996) agree that the first diaspora of the Roma occurred around A.D. 1000. At this time, the Sindh and Panjab regions of India were invaded some seventeen times by an army of Islam warriors led by King Mahmud of Ghazni (present-day Iran). Though the Indian resistance was fierce, the records indicate that King Mahmud eventually triumphed, carrying off a million slaves as human booty. These slaves are generally considered to be the ancestors of the Roma. Again, if this catastrophe took place in the year 1000, the event correlates well with the glottochronological data that splits Romani from Hindustani roughly 1,000 years ago. While certainty about early events in the development of Slavic and Romani may never be achieved, it is important to note that Swadesh’s glottochronology corresponds well with the limited information we do have about the early history of the Slavs and the Roma.

**CONCLUSION**

The purpose of this study was to test the reliability of Swadesh’s lexicostatistical model of language change by determining whether or not his statistical technique could verify conventional wisdom about the subgrouping of three Indo-European languages: Romani, Czech, and Hindustani. It was hypothesized that the results of a classic lexicostatistical study would confirm that Romani and Hindustani are languages of the same family, and that Romani and Czech represent families of the same stock. Both hypotheses proved to be correct. In addition, Swadesh’s glottochronological formula was found to be a fairly reliable instrument for determining at what time-depth Romani separated from Hindustani and at what time-depth the Slavic and Indic language families broke off from Proto-Indo-European.

Besides producing statistical results that indicate the relationship between languages, lexicostatistical analysis can also serve as a point of departure for more general lexicography. By lining up core vocabulary terms in various languages and analyzing them methodically with the aid of bilingual dictionaries, the analyst may discover a wealth of information about the histories of the words. In this study, particular attention was paid to those Romani lexemes that do not appear to have exact phonetic and semantic equivalents in the Indic languages (the so-called O-category words). It was demonstrated that such words do bear a phonetic and semantic resemblance to noncognate words in the sister and parent languages, though semantic and phonological shifts have made these relationships harder to define. In effect, it is this network of synonyms, gradually changing over time, that bespeaks the truly ancient history of the Romani language and challenges anyone who proposes that Romani is little more than a criminal jargon that has been relexified.
with an Indic vocabulary. Romani is, in fact, a rich and diverse language whose history and lexical structure are as intriguing as its enigmatic speakers.

NOTES

1. In his Sanskrit dictionary, Benfey adds another interesting twist to the semantic domain of [paka]. He refers to [kumbhir-paka] as a level of hell in which the wicked “are baked like potter’s vessels.”

REFERENCES