

On Nostratic: The Search for Linguistic Roots

by Brian D. Joseph, Ph.D., Professor and Chair of Linguistics, The Ohio State University

As humans, we have always been fascinated by origins: the origins of our institutions, the origins of our families, the origins of ourselves. This apparently quite basic drive to understand our roots is evident in, among other things, the popularity of genealogy, tracing family trees and family history.

When it comes to that most human of institutions, namely Language, the same holds true, both for the origins of language in general and the origins of particular languages. Barring time-travel of course, the ultimate origin of human language will never be known and will always remain speculative since it lies in the distant past (perhaps as much as 200,000 years, or even more, by some estimates). Still, we are in much better shape, up to a point, when it comes to understanding where particular languages come from, and what their closest and more distant relatives are.

Scholars and others have pondered these matters for centuries. The Italian poet Dante Alighieri, for instance, in the late 13th and early 14th centuries, was among the first to recognize a relationship among French, Italian, and Spanish, languages now universally agreed upon by linguists to form a linguistic grouping (the "Romance" languages), and the 17th and 18th century German philosopher Gottfried Wilhelm von Leibniz in 1710 attempted a full genealogy of then-known languages, using no apparent method and thus achieving only mixed success at best, as judged from later perspectives. By the late 19th century, spurred in part by methodological advances (especially the Comparative Method as formulated and practiced by the German scholar Franz Bopp in the early 1800s) and by an increasing scholarly interest in analyzing language in general and in classifying specific

languages, the origins of most of the languages of Europe and the Middle East were worked out: English, German, the Romance languages, Latin, Greek, Persian, Sanskrit, and others formed a language "family", usually labelled "Indo-European", whereas Hebrew, Aramaic, Arabic, and others formed the Semitic family, and Finnish, Estonian, Hungarian, and others formed the Finno-Ugric family. Similarly, numerous other groups of related languages elsewhere in the world were identified, e.g. the Kartvelian language family (Georgian, Mingrelian, Laz, etc.) in the Caucasus, the Altaic languages (Turkic, Mongolian, Manchu, maybe Korean, etc.) in Central Asia, various Pacific languages (e.g. Indonesian and Polynesian, recognized by Wilhem von Humboldt in the late 1830s), and numerous families in North and South America, in East Asia, and so on.

For many people, such groupings simply raised the further question of what the origins of each of these families are, and the desire to get at their roots has led to various hypotheses about how -- if at all -- some (or all?) of these families are related to one another.

Although there have been many schemes for larger, longer-range, groupings of languages, and although some of these were proposed in the mid-1800's (e.g. between Indo-European and Semitic by Rudolf von Raumer in 1863), perhaps the best-known such hypothesis is the NOSTRATIC hypothesis, put forth first in 1903 by Holger Pedersen, which proposes that the Indo-European, the Semitic, the Finno-Ugric, and the Altaic language families, as well as some others, are themselves part of a larger linguistic grouping, which he called "Nostratic". In this view, therefore, these language families all stem from a common source, so that ultimately, if one goes far enough back in time, according to this hypothesis Indo-European languages such as English, French, German, Spanish, and Greek, would be related to Turkish and to Finnish and to Hebrew, along with many other languages as well, and that all these languages share a common origin.

The notion of languages stemming from a common source implies, as does the "family" metaphor so commonly employed for describing language groupings, that at some point in the past, languages that are currently differentiated from one another formed a unity and were part of one and the same speech community. Such a view, in turn, can be interpreted to mean cultural, ethnic, perhaps even racial commonality for the original speakers of these languages, and can even fuel discussion of the original location of such a speech community. Still, it is crucial here to realize that humans of any race or culture or ethnicity can come to speak any language, in part through the historical accident of the community into which one is born or adopted, and that languages can spread far beyond their original geographic boundaries. Thus it is important not to make too much of a connection between linguistic origins and racial/ethnic/cultural origins, however tantalizing such speculation might be.

Nostratic, therefore, as with all long-range comparison of languages, is a bold hypothesis that has the potential to captivate and fire the imagination and to command the serious attention of linguists, historians, and the general public alike.

As a serious scientific hypothesis advanced within the discipline of historical linguistics, a question concerning Nostratic that repeatedly arises is that of proof, in particular, how might one try to prove or disprove the proposal. While there are several varieties of the Nostratic proposal -- e.g. Pedersen's original conception or the Nostratic of the Russian scholar Vladislav Illich-Svitych (working in the 1960s before his untimely death), among others -- that differ in terms of which languages are included as well as in details of the structures reconstructed as part of the "Proto-Nostratic" starting point, the basic question of proof is inescapable; the same holds, for that matter, for all hypotheses of language relationship, whether involving long-range connections or not.

In working to prove language relationships, linguists are guided by several well-worked out principles. First, one must isolate various "matchings" among the languages in question, matchings which can involve the occurrence of sounds in particular words or grammatical forms, a structural pattern for the formation of new words, a syntactic construction, or the like. Second, given what is known about the way sounds change, matchings involving sounds must fall into regular sets of correspondences that do not conflict with one another and do not necessarily involve identical or even similar sounds (though they often do). Third, if one can systematically exclude chance, borrowing, and universality as causes for these matchings, then the only possible cause that remains is that these languages are related and thus stem from a common source.

Although the importance of these principles is generally agreed upon by linguists, it is often hard to apply them, a situation which leads to considerable controversy, claims and counter-claims, and so forth. For example, having m- as the initial sound in a word for 'mother' is probably not significant, since this sound is a relatively stable one that children acquire early and this meaning is one that is a natural focus of a young child's world; thus such a word could be created -- or re-created -- in any stage of any language relatively easily, so that it violates the "universality" criterion mentioned above. However, initial m- in the same pronoun (indexing a person as referent) across various languages could possibly be significant, and initial m- in a totally arbitrary and nonindexical word like 'spider' across many languages would have a much better chance of being significant. Thus the occurrence of m- in first person singular pronouns and grammatical markers in various putative Nostratic languages, e.g. Finnish *minut* 'me'; Greek *me*, Latin *me*., German *mi(-ch)* (objective form); Turkish *-Im* (possessive form), Mongolian *minu* 'my'; Georgian *me T*, etc., is enough to make one pay attention, but given the prevalence of nasal sounds in pronominal reference across the languages of the world, non-Nostratic as well as putative Nostratic, the

criterion of universality means that this feature may not be a specifically Nostratic one, and thus most likely of no probative value for establishing the relationships claimed for Nostratic.

Similar considerations hold when it comes to sorting out potential borrowings. It is well-known that speakers often adopt words from languages other than their own, a process linguists refer to as "borrowing" (even though the "donor" is not deprived of anything and nothing is ever "returned"). Clearly, if a word in language X is a borrowing from language Y, that tells us nothing about their ancestral sources, though it does of course give information about historical or even prehistorical contacts among speakers of the languages involved. Problematic for the Nostratic hypothesis is the occurrence of similar-looking words in different branches of the putative family, for in some instances it is not clear if they represent borrowings or not. A famous case is the word for 'water', which is based on a root *wed-* in Indo-European and *wete-* in Uralic — this matching is presumably not due to universality, for the word for 'water' takes many other forms in languages of the world (unlike the relatively high frequency of *m-* in words for 'mother'), but could it be the result of a prehistoric borrowing? Is such a word likely to be borrowed? Some scholars think not, but Latin *aqua* has been borrowed into English. Admittedly *aqua* (as a designation of a liquid, not a color) primarily occurs in a somewhat specialized usage in pharmaceutical and chemical names, e.g. *aqua regia* 'a mixture of hydrochloric and nitric acids', but *The American Heritage Dictionary of the English Language* (3rd edition, 1992), at least, glosses it simply as 'water' or more generally 'an aqueous solution', and *water* itself can mean not just H₂O but more generally 'any bodily liquid' (note how the onset of labor is signalled by a pregnant woman's water, i.e. amniotic fluid, breaking). So, is this matching significant for showing that Indo-European and Uralic sprang from a common source? Maybe, or could it be the case that Indo-European or Uralic 'water' was borrowed with a specialized meaning and then broadened semantically? Perhaps, and though it is hard to decide, questions like

this need to be asked, and they make it tricky to evaluate matchings involving putative Nostratic languages. And, every matching proposed among all the language groups being considered must be scrutinized in these ways, another way in which the task of potentially proving Nostratic becomes daunting, to say the least.

Even harder to evaluate is the extent to which chance alone can lead to matchings, especially since the long-range comparisons made are generally at a great time depth, where one or two chance matchings could potentially skew the outcome. Classic cases such as Modern Greek *mati* and Malay *mata*, both meaning 'eye' but demonstrably quite different from one another in earlier linguistic stages (e.g. the earlier Greek word was *ommaton*, where the root was *om-*, whereas 'eye' in earlier Malay was *mata*, where the root is *-ta*), show that chance matchings involving convergences of form and meaning are possible. While there have been some recent attempts to apply methods from the mathematics of probability to the question of how chance matchings might get in the way of determining the truth about common linguistic origins, there is much controversy in this area, and as yet no consensus has emerged.

None of these considerations in and of themselves disprove Nostratic, but they give an idea of how difficult it might be to verify this hypothesis.

Most of what I have been saying here is factual, even didactic, in nature, and I have tried to steer a fairly neutral course, giving a view of what is potentially interesting about Nostratic and where some of the pitfalls may be for the hypothesis. Personally, I remain with what Daniel Goldin of NASA, when contemplating the possibility of organic material in a Mars rock, called a "skeptical fascination", straddling acceptance and rejection, but doing so from the enlightened perspective that scientific historical linguistics offers, and many other linguists share this view. Indeed, Joseph Salmons of the University of

Wisconsin and I have adopted that phrase to characterize the state-of-the-art regarding Nostratic in the volume "Nostratic: Sifting the Evidence" which we are editing (to appear, 1997, published by John Benjamins Publishing Company, Amsterdam & Philadelphia). After all, Indo-European must have come from somewhere and Nostratic is an interesting hypothesis as just what other languages were part of that "somewhere".

Let me end then with some frank opinions where I am not sitting on the fence, views not on Nostratic so much as on the related matter of language genesis.

As we have seen, Nostratic, if correct, could be a first step towards the recognition of a "Proto-World", and thus it leads naturally into the ultimate issue of the origin of ALL languages. I would argue, however, that this issue is really a nonissue for a few reasons:

First, "Proto-World" is really an hypothesis only about oral languages, and leaves out the numerous manually-based sign languages around the world, such as American Sign Language, which have been shown to be fully functional and expressive languages in their own right, unconnected with any oral language.

Similarly, language "birth" through the process of creolization that occurs under some conditions of extreme language contact (as has happened in parts of the Caribbean, for instance) may not be covered in "Proto-World" hypothesis, and certainly the attempts to develop native speakers of artificial languages like Esperanto are similarly not covered. While it might be argued that creolization and Esperanto etc. draw on existing language resources, and so are offshoots of languages covered by the Proto-World hypothesis, this point makes the signed language argument more telling, since signed languages are NOT derivative from oral languages.

Finally, even if one could show that all the languages of the world really derive from a single language in the far distant past, that would still tell us nothing about the question of how the currently observable linguistic diversity on Earth came into being; that is, there is still much to learn and explore about the ways language in general evolves through time, and the ways specific languages have evolved through time, quite apart from what their "original evolution", their point of origin, might have been.

Thus, Nostratic, and any "Proto-World" hypotheses, as intrinsically interesting and captivating as they might be, would nonetheless not provide all the answers and might not even be asking all the questions about language history and prehistory.

SUBSTITUTE ABOVE IN third ¶ up:

Similarly, language "birth" has occurred through the process of creolization, the creation of "contact languages" (often referred to as "pidgins" or "creoles") under extreme conditions in which speakers of different languages are thrown together (as happened in various plantation settings in the Caribbean or Hawaii), and these newly emerging languages may not be covered in the "Proto-World" hypothesis.

BRIAN D. JOSEPH received his Ph.D. in Linguistics from Harvard University in 1978, and has been teaching at The Ohio State University since 1979. A specialist in historical linguistics, focusing on Indo-European and particularly Greek, Latin, and Sanskrit, he has authored or co-authored over 100 articles and 5 books in these and related areas. He is co-editor, along with Joseph C. Salmons of the University of Wisconsin of "Nostratic: Sifting the Evidence" (to appear, 1997, published by John Benjamins Publishing Company, Amsterdam & Philadelphia).

On Nostratic: The Search for Linguistic Roots

by Brian D. Joseph, Ph.D., Professor and Chair of Linguistics,
The Ohio State University

As humans, we have always been fascinated by origins: the origins of our institutions, the origins of our families, the origins of ourselves. This apparently quite basic drive to understand our roots is evident in, among other things, the popularity of genealogy, tracing family trees and family history.

When it comes to that most human of institutions, namely Language, the same holds true, both for the origins of language in general and the origins of particular languages. Barring time-travel of course, the ultimate origin of human language will never be known and will always remain speculative since it lies in the distant past (perhaps as much as 200,000 years, or even more, by some estimates). Still, we are in much better shape, up to a point, when it comes to understanding where particular languages come from, and what their closest and more distant relatives are.

Scholars and others have pondered these matters for centuries. The Italian poet Dante Alighieri, for instance, in the late 13th and early 14th centuries, was among the first to recognize a relationship among French, Italian, and Spanish, languages now universally agreed upon by linguists to form a linguistic grouping (the "Romance" languages), and the 17th and 18th century German philosopher Gottfried Wilhelm von Leibniz in 1710 attempted a full genealogy of then-known languages, using no apparent method and thus achieving only mixed success at best, as judged from later perspectives. By the late 19th century, spurred in part by methodological advances (especially the Comparative Method as formulated and practiced by the German scholar Franz Bopp in the early 1800s) and by an increasing scholarly interest in analyzing language in general and in classifying specific languages, the origins of most of the languages of Europe and the Middle East were worked out: English, German, the Romance languages, Latin, Greek, Persian, Sanskrit, and others formed a language "family", usually labelled "Indo-European", whereas Hebrew, Aramaic, Arabic, and others formed the Semitic family, and Finnish, Estonian, Hungarian, and others formed the Finno-Ugric family. Similarly, numerous other groups of related languages elsewhere in the world were identified, e.g. the Kartvelian language family (Georgian, Mingrelian, Laz, etc.) in the Caucasus, the Altaic languages (Turkic, Mongolian, Manchu, maybe Korean, etc.) in Central Asia, various Pacific languages (e.g. Indonesian and Polynesian, recognized by Wilhem von Humboldt in the late 1830s), and numerous families in North and South America, in East Asia, and so on.

For many people, such groupings simply raised the further question of what the origins of each of these families are, and the desire to get at their roots has led to various hypotheses about how -- if at all --

some (or all?) of these families are related to one another.

Although there have been many schemes for larger, longer-range, groupings of languages, and although some of these were proposed in the mid-1800's (e.g. between Indo-European and Semitic by Rudolf von Raumer in 1863), perhaps the best-known such hypothesis is the NOSTRATIC hypothesis, put forth first in 1903 by Holger Pedersen, which proposes that the Indo-European, the Semitic, the Finno-Ugric, and the Altaic language families, as well as some others, are themselves part of a larger linguistic grouping, which he called "Nostratic". In this view, therefore, these language families all stem from a common source, so that ultimately, if one goes far enough back in time, according to this hypothesis Indo-European languages such as English, French, German, Spanish, and Greek, would be related to Turkish and to Finnish and to Hebrew, along with many other languages as well, and that all these languages share a common origin.

The notion of languages stemming from a common source implies, as does the "family" metaphor so commonly employed for describing language groupings, that at some point in the past, languages that are currently differentiated from one another formed a unity and were part of one and the same speech community. Such a view, in turn, can be interpreted to mean cultural, ethnic, perhaps even racial commonality for the original speakers of these languages, and can even fuel discussion of the original location of such a speech community. Still, it is crucial here to realize that humans of any race or culture or ethnicity can come to speak any language, in part through the historical accident of the community into which one is born or adopted, and that languages can spread far beyond their original geographic boundaries. Thus it is important not to make too much of a connection between linguistic origins and racial/ethnic/cultural origins, however tantalizing such speculation might be.

Nostratic, therefore, as with all long-range comparison of languages, is a bold hypothesis that has the potential to captivate and fire the imagination and to command the serious attention of linguists, historians, and the general public alike.

As a serious scientific hypothesis advanced within the discipline of historical linguistics, a question concerning Nostratic that repeatedly arises is that of proof, in particular, how might one try to prove or disprove the proposal. While there are several varieties of the Nostratic proposal -- e.g. Pedersen's original conception or the Nostratic of the Russian scholar Vladislav Illich-Svitych (working in the 1960s before his untimely death), among others -- that differ in terms of which languages are included as well as in details of the structures reconstructed as part of the "Proto-Nostratic" starting point, the basic question of proof is inescapable; the same holds, for that matter, for all hypotheses of language relationship, whether involving long-range connections or not.

In working to prove language relationships, linguists are guided by

several well-worked out principles. First, one must isolate various "matchings" among the languages in question, matchings which can involve the occurrence of sounds in particular words or grammatical forms, a structural pattern for the formation of new words, a syntactic construction, or the like. Second, given what is known about the way sounds change, matchings involving sounds must fall into regular sets of correspondences that do not conflict with one another and do not necessarily involve identical or even similar sounds (though they often do). Third, if one can systematically exclude chance, borrowing, and universality as causes for these matchings, then the only possible cause that remains is that these languages are related and thus stem from a common source.

Although the importance of these principles is generally agreed upon by linguists, it is often hard to apply them, a situation which leads to considerable controversy, claims and counter-claims, and so forth. For example, having m- as the initial sound in a word for 'mother' is probably not significant, since this sound is a relatively stable one that children acquire early and this meaning is one that is a natural focus of a young child's world; thus such a word could be created -- or re-created -- in any stage of any language relatively easily, so that it violates the "universality" criterion mentioned above. However, initial m- in the same pronoun (indexing a person as referent) across various languages could possibly be significant, and initial m- in a totally arbitrary and nonindexical word like 'spider' across many languages would have a much better chance of being significant. Thus the occurrence of m- in first person singular pronouns and grammatical markers in various putative Nostratic languages, e.g. Finnish *minut* 'me'; Greek *me*, Latin *me*., German *mi(-ch)* (objective form); Turkish *-Im* (possessive form), Mongolian *minu* 'my'; Georgian *me T*, etc., is enough to make one pay attention, but given the prevalence of nasal sounds in pronominal reference across the languages of the world, non-Nostratic as well as putative Nostratic, the criterion of universality means that this feature may not be a specifically Nostratic one, and thus most likely of no probative value for establishing the relationships claimed for Nostratic.

Similar considerations hold when it comes to sorting out potential borrowings. It is well-known that speakers often adopt words from languages other than their own, a process linguists refer to as "borrowing" (even though the "donor" is not deprived of anything and nothing is ever "returned"). Clearly, if a word in language X is a borrowing from language Y, that tells us nothing about their ancestral sources, though it does of course give information about historical or even prehistorical contacts among speakers of the languages involved. Problematic for the Nostratic hypothesis is the occurrence of similar-looking words in different branches of the putative family, for in some instances it is not clear if they represent borrowings or not. A famous case is the word for 'water', which is based on a root *wed-* in Indo-European and *wete-* in Uralic — this matching is presumably not due to universality, for the word for 'water' takes many other forms in languages of the world (unlike the relatively high frequency of m- in words for 'mother'), but could it

be the result of a prehistoric borrowing? Is such a word likely to be borrowed? Some scholars think not, but Latin *aqua* has been borrowed into English. Admittedly *aqua* (as a designation of a liquid, not a color) primarily occurs in a somewhat specialized usage in pharmaceutical and chemical names, e.g. *aqua regia* 'a mixture of hydrochloric and nitric acids', but *The American Heritage Dictionary of the English Language* (3rd edition, 1992), at least, glosses it simply as 'water' or more generally 'an aqueous solution', and water itself can mean not just H₂O but more generally 'any bodily liquid' (note how the onset of labor is signalled by a pregnant woman's water, i.e. amniotic fluid, breaking). So, is this matching significant for showing that Indo-European and Uralic sprang from a common source? Maybe, or could it be the case that Indo-European or Uralic 'water' was borrowed with a specialized meaning and then broadened semantically? Perhaps, and though it is hard to decide, questions like this need to be asked, and they make it tricky to evaluate matchings involving putative Nostratic languages. And, every matching proposed among all the language groups being considered must be scrutinized in these ways, another way in which the task of potentially proving Nostratic becomes daunting, to say the least.

Even harder to evaluate is the extent to which chance alone can lead to matchings, especially since the long-range comparisons made are generally at a great time depth, where one or two chance matchings could potentially skew the outcome. Classic cases such as Modern Greek *mati* and Malay *mata*, both meaning 'eye' but demonstrably quite different from one another in earlier linguistic stages (e.g. the earlier Greek word was *ommaton*, where the root was *om-*, whereas 'eye' in earlier Malay was *mata*, where the root is *-ta*), show that chance matchings involving convergences of form and meaning are possible. While there have been some recent attempts to apply methods from the mathematics of probability to the question of how chance matchings might get in the way of determining the truth about common linguistic origins, there is much controversy in this area, and as yet no consensus has emerged.

None of these considerations in and of themselves disprove Nostratic, but they give an idea of how difficult it might be to verify this hypothesis.

Most of what I have been saying here is factual, even didactic, in nature, and I have tried to steer a fairly neutral course, giving a view of what is potentially interesting about Nostratic and where some of the pitfalls may be for the hypothesis. Personally, I remain with what Daniel Goldin of NASA, when contemplating the possibility of organic material in a Mars rock, called a "skeptical fascination", straddling acceptance and rejection, but doing so from the enlightened perspective that scientific historical linguistics offers, and many other linguists share this view. Indeed, Joseph Salmons of the University of Wisconsin and I have adopted that phrase to characterize the state-of-the-art regarding Nostratic in the volume "Nostratic: Sifting the Evidence" which we are editing (to appear, 1997, published by John Benjamins Publishing Company,

Amsterdam & Philadelphia). After all, Indo-European must have come from somewhere and Nostratic is an interesting hypothesis as just what other languages were part of that "somewhere".

Let me end then with some frank opinions where I am not sitting on the fence, views not on Nostratic so much as on the related matter of language genesis.

As we have seen, Nostratic, if correct, could be a first step towards the recognition of a "Proto-World", and thus it leads naturally into the ultimate issue of the origin of ALL languages. I would argue, however, that this issue is really a nonissue for a few reasons:

First, "Proto-World" is really an hypothesis only about oral languages, and leaves out the numerous manually-based sign languages around the world, such as American Sign Language, which have been shown to be fully functional and expressive languages in their own right, unconnected with any oral language.

Similarly, language "birth" through the process of creolization that occurs under some conditions of extreme language contact (as has happened in parts of the Caribbean, for instance) may not be covered in "Proto-World" hypothesis, and certainly the attempts to develop native speakers of artificial languages like Esperanto are similarly not covered. While it might be argued that creolization and Esperanto etc. draw on existing language resources, and so are offshoots of languages covered by the Proto-World hypothesis, this point makes the signed language argument more telling, since signed languages are NOT derivative from oral languages.

Finally, even if one could show that all the languages of the world really derive from a single language in the far distant past, that would still tell us nothing about the question of how the currently observable linguistic diversity on Earth came into being; that is, there is still much to learn and explore about the ways language in general evolves through time, and the ways specific languages have evolved through time, quite apart from what their "original evolution", their point of origin, might have been.

Thus, Nostratic, and any "Proto-World" hypotheses, as intrinsically interesting and captivating as they might be, would nonetheless not provide all the answers and might not even be asking all the questions about language history and prehistory.