Um Tratamento Quantitativo para a Classificação das Línguas Indo-Européias

(Quantitative Approach to the Classification of Indo-European Language)

Lincoln Almir Amarante Ribeiro
(Grupo de Investigação Científica de Línguas Indígenas da Universidade Estadual de Goiás e Universidade Federal de Minas Gerais)

Abstract: *This work applies the biology cladistic methodology in the analysis of the classification of the Indo-European family of languages, through phonologic and morphologic characters of 12 linguistic families. For the phylogenetc analysis of the phonologic and morphologic data we use the method of maximum parsimony. Our results point to the existence of the traditional Indo-European languages sub-groups certified in the literature and indicate that one can exist an Italic-Celtic-Germanic supergroup.*
Key-words: *Indo-European Languages; Cladistics; Maximum Parcimony Method.*

Resumo: *Este trabalho tem por objetivo demonstrar a aplicação de uma metodologia cladística da Biologia Evolucionária na análise da classificação das línguas Indo-Européias por meio de caracteres fonológicos e morfológicos de 12 famílias lingüísticas. Para a análise filogenética dos dados, foi usado o tratamento de máxima parcimônia. Os resultados da análise aqui descrita apontam para a existência dos subgrupos tradicionais das línguas Indo-Européias atestados na literatura e indicam, ainda, que pode existir um super grupo de línguas Ítalo-Celtico-Germânico.*
Palavras-chave: *Línguas Indo-Européias; Cladística; Método da Máxima Parcimônia.*

Introdução

O presente trabalho é uma tentativa de aplicar a metodologia cladística na análise da classificação das línguas Indo-Européias, através de caracteres fonológicos e morfológicos de 12 famílias lingüísticas. É preciso ressaltar que uma tentativa dessa natureza já foi feita por Kroeber & Chretien (1937) em uma versão reduzida que contemplou apenas 9 línguas. Contudo, por não ter usado os métodos mais modernos da Biologia Evolucionária, não disponíveis na época, os resultados apresentados por esses autores deixam dúvidas quanto à consistência do método estatístico utilizado. Buscando então sanar problemas desse tipo na análise filogenética dos dados fonológicos e morfológicos aqui apresentada, usaremos o tratamento de Máxima Parcimônia. Assim, o trabalho terá como base as seguintes premissas: a) as línguas, estão sujeitas à evolução de uma maneira puramente cultural (estamos desprezando qualquer ligação biológica com essa evolução) e individualmente preservam sua continuidade através de grandes escalas temporais; b) a evolução das línguas é divergente e c) a língua é transmitida como um todo e a freqüência de empréstimos, isto é, a transmissão horizontal dos caracteres entre as línguas é baixa.

Preliminares

Desde a época de Darwin (1871), acredita-se que a evolução das línguas e das espécies ocorre de maneira paralela bem como a demonstração de que esse fenômeno se desenvolve mediante um processo gradual. De fato, processos biológicos fundamentais de evolução como, por exemplo, a cladogênese, a seleção, a flutuação aleatória e a mutação, têm análogos lingüísticos (Pagel, 2000). Isto é, assim como as espécies estão sujeitas à seleção natural, as línguas estão sujeitas à seleção social. Os processos de flutuação aleatória e mutação atuam nos "linguemas" da mesma maneira que nos genes (Croft, 2000). E o que é mais fundamental: do mesmo modo que as linhagens biológicas separam-se e divergem em árvores de famílias, também as línguas devem se comportar.

Naturalmente, todo esse paralelismo indica que tanto a Biologia Evolucionária quanto a Lingüística Histórica buscam respostas para questões semelhantes, que ambas as ciências encontram as mesmas dificuldades e que, muitas vezes (como no presente estudo, aliás), costumam usar métodos semelhantes para se chegar a uma solução de problema.

Muitas línguas naturais se agrupam em famílias geneticamente relacionadas. De acordo com a Lingüística Histórica, um grupo é geneticamente relacionado lingüisticamente se: a) todas as línguas desse grupo foram em alguma época do passado uma só língua; b) esta língua comum do passado (chamada protolíngua) transformou-se nas diversas línguas do grupo; c) essa transformação foi realizada por meio da transmissão de L1 (a criança adquire sua primeira língua através do contato com os adultos de sua comunidade lingüística).

Assim, acredita-se que as línguas possam ser descritas por uma estrutura evolucionária hierárquica na forma de uma árvore, ou seja, uma representação gráfica da evolução do grupo de línguas a partir do ancestral comum mais próximo. As línguas dão nomes aos ramos da árvore enquanto os nós internos indicam os ancestrais comuns. O nó de partida é freqüentemente denominado raiz. Um nó é rotulado como "pai" de outro nó se está situado imediatamente acima na hierarquia e mais próximo da raiz. Nós "irmãos" são aqueles que compartilham o mesmo nó "pai". O nó que está conectado a todos os nós mais baixos na hierarquia é chamado de "ancestral". A raiz é a ancestral de todos os nós. Notemos que o desenho, ou seja, a representação gráfica de uma árvore na Lingüística assim como na Biologia é o de uma árvore invertida. A linguagem de árvores foi introduzida no meio científico por Schleicher (1861).

Um aumento da ênfase no contato lingüístico tem desafiado a validade do modelo de árvores como uma descrição satisfatória para o relacionamento genético entre línguas. Já no passado visões diferentes para a evolução das línguas naturais - entre elas, o modelo de ondas de Schmidt (1872) - questionaram essa visão e deram origem à ciência da Dialetologia. Mais recentemente, Dixon (1997) criticou a validade do modelo em certas circunstâncias. Apesar disso, o tratamento da classificação de línguas por meio de árvores continua tendo muita aceitação nos trabalhos de descrição de sua evolução histórica.

Um problema antigo e muito controverso na Lingüística Histórica é o de se obter a árvore evolucionária para a família de línguas Indo-Européias (doravante IE), aquela que inclui a maioria dos idiomas ainda falados ou já extintos da Europa. O domínio de falantes da família IE se estende da Irlanda, no oeste da Europa, ao Turquestão Chinês, no este da Ásia, e, da Escandinávia, no norte da Europa, até a Itália e a Grécia, no sul desse continente.

A existência de 12 subfamílias (Albanesa, Anatoliana ou Anatólica, Armênica, Báltica, Céltica, Germânica, Helênica, Índica, Iraniana, Itálica, Eslávica e Tocariana ou Tocárica[1]) não é colocada em dúvida atualmente. Acredita-se que as línguas da família IE possam ser descritas por um modelo de árvores e assim seja possível aplicar os métodos da Biologia Cladística para a determinação dessa árvore. Segundo Rexová (2003), a barreira entre as humanidades e as ciências tem impedido o uso da metodologia cladística na Lingüística Comparativa. Exceções a isso são a própria Rexová (2003) com o IE, Gray & Jordan (2000) com as línguas Malaio-Polinésias, Gray & Atkinson (2003) para as línguas IE, Holden (2002) com as línguas Bantu e Bantóides e Foster & Toth (2003) com as línguas célticas. Algoritmos filogenéticos foram também usados por Warnow (1977) e Ringe, Warnow & Taylor (2002) que trataram o IE de uma maneira cladística.

Recentemente, Dunn et alli (2005) aplicaram métodos cladísticos nas línguas Papua que se separaram há mais de 10 mil anos e, usando 125 traços gramaticais, obtiveram uma árvore na qual as línguas se agrupavam de acordo com sua proximidade geográfica. Esse resultado, embora ainda carecendo de maiores confirmações, traz um grande alento ao uso de métodos da Biologia Genética para a Lingüística Histórico-Comparativa.

Métodos léxico-estatísticos como os de Swadesh (1952) também podem ser usados na determinação de árvores, mas seus objetivos são outros. Esse autor, por exemplo, utiliza a analogia do modelo biológico do relógio molecular para determinar data de separação de línguas, ou seja, os tempos dos nós.

Os Dados

Os dados analisados neste artigo têm como base uma lista de 74 itens elaborada por Kroeber & Chretien (1937). Essa lista foi preenchida com dados da família IE, mais especificamente, das línguas Armênicas, Bálticas, Itálicas, Célticas, Eslavas, Germânicas, Índicas, Helênicas e Iranianas. Os dados foram coletados por Kroeber e revisados posteriormente por seu

[1] Abreviaturas usadas neste texto dos nomes das famílias lingüísticas: Al, Albanesa, Ar, Armênica; Ba, Báltica; Ce, Céltica; Ge, Germânica; Gr, Grega; It, Itálica; Hi, Anatoliana ou Anatólica; Ir, Iraniana; Sk, Índica; Sl, Eslávica e To, Tocariana ou Tocárica.

colaborador Chretien com base nas seguintes fontes: Brugmann & Delbrück (1897-1916), Kieckens (1933), Meillet (1922,1934) e Sommer (1914). Posteriormente, Kroeber & Chretien (1939) acrescentaram à lista dados do Hitita.

Aos dados de Kroeber e Chretien (1937), acrescentamos o Albanês utilizando informações de Mann (1941, 1950, 1952) para os dados fonológicos e Mann (1977), para os morfológicos. Também houve acréscimo de nossa parte à lista com dados do Tocário, os quais foram obtidos junto a Adams (1984, 1988). A lista original de Kroeber & Chretien (1937.) é a seguinte:

1. Assimilação de ***e** e ***a**;
2. assimilação de ***ə** e ***i**, em lugar ***ə** e ***a**;
3. as consoantes oclusivas vozeadas aspiradas e as desvozeadas aspiradas tornam-se, respectivamente, oclusivas vozeadas não-aspiradas e aspiradas vozeadas (Lei de Bartholomae);
4. genitivo e ablativo singular de genitivo e ablativo de temas em **ā** em **- āyā- (-(i)y-)**;
5. genitivo plural ***-ōm** substituído em temas vocálicos por **–n-ām**;
6. terceira pessoa do imperativo em **-u**;
7. assimilação de ***r̥ *l̥** e ***°r *°l**;
8. assimilação de ***m̥ *n̥** e ***°m *°n**;
9. ***p...k**^{W} torna-se **k**^{W}**...k**^{W};
10. genitivo singular de temas em o em **ī**;
11. sufixo formativo ***tjjen**, ***tijen**;
12. superlativo em ***sm̥mo-**,***ism̥mo-**;
13. futuro em **-bō**;
14. passiva em **-r-**;
15. subjuntivo em **-ā-**;
16. subjuntivo em **-s-**;
17. uso de adjetivo verbal em ***-to-** como particípio passado;
18. simplificação de consoantes geminadas;
19 . particípio presente masculino e neutro com ***-jo-**, flexão análoga ao feminino ***-jā-**;
20. *centum* torna-se *satem*;
21. lábio-velares tornam-se velares;
22. assimilação de ***o** e ***a**;

23. ***ā*** e ***ō*** não assimilados;

24. assimilação de ***ā*** a ***ō***;

25. assimilação de ***ō*** a ***ā***;

26. deslocamento de ***-tt-*** para ***-st-***;

27. perda de ***ə*** medial antes de uma consoante;

28. perda de ***ə*** medial antes de uma consoante e seguida de uma sílaba contendo ***ō***;

29. perda de ***ə*** medial antes de uma consoante e após uma vogal mais ***i***;

30. deslocamento de ***-wj*** para ***uj***;

31. consoantes oclusivas aspiradas vozeadas ficam inalteradas.

32. consoantes oclusivas aspiradas tornam-se oclusivas vozeadas;

33. consoantes oclusivas aspiradas tornam-se fricativas;

34. consoantes oclusivas desvozeadas e oclusivas desvozeadas mantêm-se distintas;

35. consoantes oclusivas aspiradas desvozeadas são parcialmente assimiladas a oclusivas desvozeadas;

36. consoantes oclusivas aspiradas desvozeadas são totalmente assimiladas pelas oclusivas desvozeadas;

37. consoantes oclusivas aspiradas desvozeadas tornam-se fricativas;

38. ***s*** torna-s **š** depois de **i**, **u**, **r**, **k** regularmente.

39. ***s*** torna-s **š** depois de **i**, **u**, **r**, **k** somente quando seguida de vogal na mesma palavra;

40. ***s*** torna-se **h** quando se encontra em posição: (a) inicial, (b) intervocálica, ou (c) antes ou após uma consoante não oclusiva;

41. consoantes oclusivas aspiradas tornam-se fricativas;

42. consoantes oclusivas vozeadas tornam-se oclusivas desvozeadas;

43. aumento;

44. perfeito regular reduplicado;

45. reduplicação ocasional;

46. nenhuma reduplicação;

47. o pretérito perfeito é preservado;

48. o pretérito é derivado parcialmente do perfeito e parcialmente do aoristo;

49. o pretérito é derivado inteiramente do aoristo;

50. o sufixo verbal **-je-/-jo-** é usado para derivados, sufixo ***-i-*** para estados;

51. o sufixo verbal **-je-/-jo-** e ***-i-*** são usados para derivados;

52. o sufixo verbal **-je-/-jo-** usado para derivados e estados, o sufixo ***-i-*** não é usado;

53. abstratos verbais do tipo raiz mais ***o/*ā** freqüentes;

54. abstratos verbais do tipo raiz + ***o/*ā** não freqüentes, mas mais do que esporádicos;

55. abstratos verbais do tipo raiz + *o/*ā esporádicos;

56. comparativo em ***-jes-**, ***jos-** , ***-is-**;

57. comparativo em ***-isen-**, ***-ison-**;

58. sufixo ***-tero-**, ***-toro-**, ***-tro-** usado para comparativo;

59. sufixo ***-tero-**, ***-toro-**, ***-tro-** usado em certas palavras originalmente como comparativo, mas perdeu a força comparativa;

60. particípios formados pelo sufixo ***-lo-**;

61. temas em ***o** são femininos bem como masculinos;

62. sufixo ***-tūt-** forma nomes abstratos comumente;

63. sufixo ***-tūt-** forma nomes abstratos raramente;

64. sufixo ***-tūt-** não usado;

65. números coletivos em ***-o-** (Sânscrito: **traya´h**);

66. números coletivos em ***-no-** (Latim: **trini**);

67. comparativo em ***-jes-**, ***-jos-**, ***-is-** não tem feminino;

68. sufixo de caso ***-bh-** substituído pelo de caso em ***-m-**;

69. locativo plural em ***-su**;

70. destruição do sistema de casos original e amalgamação das funções do dativo, ablativo, locativo e instrumental;

71. nominativo plural de temas em ***o** em **-oi** sob a influência do demonstrativo;

72. nominativo plural de temas em ***ā** em **-āi** em analogia aos temas em ***o**;

73. genitivo plural de temas em ***ā** usados na forma demonstrativa (***-āsom**);

74. formas de ***bhewə** 'crescer' parcialmente substitui ***es-** 'ser'.

Os caracteres usados na Lista de Kroeber & Chritien consistem em um conjunto de 30 caracteres fonológicos, sendo 12 vocálicos e 18 consonantais; 44 caracteres morfológicos, sendo 17 nominais, 8 adjetivais e 19 verbais.

No Quadro 1, a seguir, o sinal "+" expressa a presença e o "-" a ausência de um determinado caractere. O sinal de interrogação (?) significa que ou há dúvidas sobre a presença ou ausência do dado ou que absolutamente não se conseguiu obtê-lo. Quando o sinal "+" ou "-" está colocado entre parêntesis ou seja (+) e (-) respectivamente, indica-se que o estado ocorre, mas nem sempre. Em nosso trabalho, contamos (+) como "+" e (-) como "-".

Quadro 1: Matriz Obtida pela Análise da Lista de Kroeber & Chretien (1939) para as Línguas Indo-Européias.

	Ce	It	Gr	Ar	Ir	Sk	Sl	Ba	Ge	Hi	To	Al		Ce	It	Gr	Ar	Ir	Sk	Sl	Ba	Ge	Hi	To	Al
1	-	-	-	-	+	+	-	-	-	-	-	-	**38**	-	-	-	?	+	+	-	-	-	-	-	-
2	-	-	-	-	+	+	-	-	-	-	-	-	**39**	-	-	-	?	-	-	+	(+)	-	-	-	-
3	-	-	-	-	+	+	-	-	-	?	?	-	**40**	-	-	+	+	+	-	-	-	-	-	-	-
4	+	-	-	+	+	+	-	-	-	?	-	-	**41**	-	-	-	+	-	-	-	-	+	-	-	-
5	-	-	-	-	+	+	-	-	(+)	-	-	-	**42**	-	-	-	+	-	-	-	-	+	+	+	-
6	-	-	-	-	+	+	-	-	-	+	-	-	**43**	-	-	+	+	+	+	-	-	-	-	-	-
7	-	-	+	+	-	-	-	+	+	+	+	-	**44**	-	-	+	?	+	+	-	-	-	-	-	-
8	+	+	-	+	-	-	-	+	+	-	+	-	**45**	+	+	-	-	-	-	-	-	+	+	-	-
9	+	+	-	-	-	-	-	-	-	?	-	-	**46**	-	-	-	-	-	-	+	+	-	-	+	+
10	+	+	-	-	-	-	-	-	-	-	?	-	**47**	-	-	+	(+)	+	+	-	-	-	+	-	-
11	+	+	-	-	-	-	-	-	-	-	-	-	**48**	+	+	-	-	-	-	-	-	+	?	+	-
12	+	+	-	-	-	-	-	-	-	-	-	-	**49**	-	-	-	-	-	-	+	+	-	-	-	+
13	+	+	-	-	-	-	-	-	-	-	-	-	**50**	-	-	-	+	-	-	+	+	-	-	-	-
14	+	+	-	-	-	-	-	-	-	+	+	-	**51**	+	+	-	-	-	-	-	-	+	-	-	+
15	+	+	-	-	-	-	-	-	-	?	+	-	**52**	-	-	+	-	+	+	-	-	-	+	+	-
16	+	+	-	-	-	-	-	-	-	?	+	-	**53**	-	-	+	-	+	+	+	+	-	-	-	?
17	+	+	-	-	+	+	+	+	-	-	+	-	**54**	-	-	-	-	-	-	-	-	+	-	+	?
18	-	-	-	+	-	-	+	+	-	-	-	+	**55**	+	+	-	-	-	-	-	-	-	+	-	?
19	-	-	-	-	-	-	+	+	+	-	?	-	**56**	+	+	+	-	+	+	+	-	-	-	-	?
20	-	-	-	+	+	+	+	+	-	-	-	+	**57**	-	-	+	-	-	-	-	+	+	-	-	-
21	-	-	-	+	+	+	+	+	-	-	-	+	**58**	+	-	+	-	+	+	-	-	-	-	?	-
22	-	-	-	-	+	+	+	+	+	-	+	+	**59**	-	+	-	-	-	-	-	-	+	-	-	-
23	-	+	+	+	-	-	-	-	-	-	-	+	**60**	-	-	-	+	-	-	+	-	-	-	-	-
24	-	-	-	-	-	-	-	+	+	+	-	+	**61**	-	+	+	+	-	-	-	-	-	-	+	-
25	+	-	-	-	+	+	+	-	-	+	+	-	**62**	+	-	-	-	-	-	-	-	-	-	-	-
26	-	-	+	?	+	+	+	+	-	+	-	-	**63**	-	-	-	-	+	-	-	-	+	-	-	-
27	-	-	-	+	+	-	+	+	+	?	+	+	**64**	-	-	+	+	-	+	+	+	-	+	+	-
28	?	+	+	(+)	(+)	-	(+)	(+)	(+)	?	?	+	**65**	-	-	-	-	+	+	+	+	-	-	_	-
29	-	-	-	-	-	+	-	-	-	?	?	+	**66**	-	+	-	-	-	-	-	+	+	-	-	+
30	-	-	-	-	+	-	+	+	+	?	-	-	**67**	+	+	+	-	-	-	-	-	-	?	?	-
31	-	-	-	-	-	+	-	-	-	-	-	-	**68**	-	-	-	(-)	-	-	+	+	+	?	?	-
32	+	-	-	+	+	-	+	+	-	+	-	+	**69**	-	-	-	+	+	+	+	+	-	-	-	-
33	-	+	+	-	-	-	-	-	+	-	-	-	**70**	+	+	+	-	-	-	-	-	+	-	+	+
34	-	-	+	(+)	+	+	(-)	-	-	-	-	?	**71**	+	+	+	-	-	-	-	+	+	-	-	-
35	-	-	(-)	+	-	-	(+)	-	-	-	-	?	**72**	-	+	+	-	-	-	-	-	-	-	+	-
36	+	+	(-)	-	-	-	(+)	+	+	+	+	-	**73**	-	+	+	-	-	-	-	-	-	-	-	-
37	-	-	+	(-)	+	-	-	-	+	-	-	-	**74**	+	+	-	-	+	+	+	+	+	-	-	-

O Método

Em ciência, de uma maneira geral, denominamos "parcimônia" a preferência pela explicação menos complicada para a observação de um fato experimental. Esta atitude é considerada adequada geralmente no julgamento de hipóteses alternativas.

Em Sistemática, a Máxima Parcimônia (doravante MP) é um critério cladístico de otimização, baseado no principio da parcimônia acima enunciado (Fitch,1971). Trata-se, assim, de uma técnica usada na cladística com a finalidade de obter árvores filogenéticas para um conjunto de táxons (conjunto de espécies na Biologia; conjunto de línguas na Lingüística). Os

dados de entrada numa análise MP são os caracteres para um certo grupo de táxons. Um caractere pode ser um valor binário para a presença ou ausência de uma propriedade (como em nosso trabalho) ou, no caso da Biologia, uma proteína ou um ácido nucléico do genoma do táxon.

As árvores usadas na análise MP são em geral árvores sem raiz (não há indicação na árvore, só relações entre os táxons). Se um caractere sofre uma mudança, dizemos que ele sofreu uma transição. É a árvore MP que explica os caracteres e que tem menor ou igual número de transições que qualquer outra árvore possível que explique os caracteres. Para se implementar isso na prática, a todas as árvores é atribuído um comprimento igual ao número de transições que explica a árvore. A árvore que tem menor comprimento é a MP.

Embora sem sempre seja muito preciso, o método de análise MP é muito simples e, por isso, muito usado pelos biólogos. O maior problema com a MP reside no fato de que, em seus fundamentos, exigi-se que duas espécies que compartilhem um mesmo caractere sejam geneticamente relacionadas. Isso, porém, nem sempre é verdadeiro. Por exemplo, ambos os animais, as aves e os morcegos, têm asas, o que não ocorre com o homem e o crocodilo. Baseando-se nesses dados, o método de análise MP tende a agrupar crocodilos com homens e aves com morcegos. Todavia, é sabido que o homem está muito mais próximo do morcego (por serem ambos mamíferos) do que de aves e crocodilos.

Um problema semelhante ocorre na *back mutation*, ou seja, o processo no qual uma pequena mutação produz a perda de um caractere por parte de uma espécie e mais tarde esse caractere é recuperado novamente. Devido ao fato de a análise MP usar árvores sem raiz, a inovação independente e a *back mutation* são matematicamente equivalentes. Outras falhas desse método são apresentadas em Felsenstein (1978).

Aplicação do Método MP

De posse dos dados do Quadro 1, construímos então uma matriz de caracteres que será usada para a aplicação do método MP. Cada caractere foi codificado como ausente (0) ou presente (1). A matriz resultante apresenta 888 elementos (12 linhas por 74 colunas).

Para reconstruir as relações lingüísticas do Indo-Europeu, foram realizadas análises filogenéticas evolucionárias por meio do programa de computador MEGA, versão 3.1, de Kumar et alli (2004) porque, dentre suas várias utilidades, esse programa implementa o método de MP.[2]

Em nossa análise, todos os caracteres foram tratados como não ordenados e com pesos iguais. Pelo menos 10.000 ciclos de *bootstrapping* foram executados. Todas as colunas com *gaps* foram desprezadas, ou seja, usamos o *pairwise delection.* O número de táxons é 12, o de *sites* 74 e o CNI (*Close-Neighbor-Interchange*) com nível de procura igual a 3. A adição aleatória de árvores com 50 réplicas; a semente aleatória para o teste de filogenia foi de 24054.

Um fato que deve ser relatado aqui é que, ao contrário de alguns outros programas disponíveis para execução de análises filogenéticas, o MEGA não tem uma implementação para matrizes binárias utilizando-se código (1/0). Isso porque o programa foi construído para trabalhar exclusivamente com dados biológicos, ou seja, com proteínas ou nucleotídeos. Assim, foi necessário idealizar uma adaptação desse programa, a qual consiste em substituir os dados binários codificados numericamente por dados de bases. Isto é, substituímos 1 por, digamos, A e 0 por G. Outras escolhas são possíveis e não alteram o resultado. Seria como se toda a codificação genética da natureza utilizasse duas bases em vez de quatro.

Resultados e Conclusões

A árvore de consenso é apresentada na Figura 1, a seguir. Notemos que o Hitita foi colocado como raiz das árvores de acordo com a evidência lingüística em consonância com Gamkrelidze & Ivanov (1995) e Rexová et alli (2003). Em outras palavras, estamos admitindo que a Hipótese do Indo-Hitita seja válida, embora isso não cause nenhum problema na aplicação do método MP, porque a introdução do Hitita como grupo externo e como raiz só pode ser implementado após a escolha da árvore de consenso.

[2] Este programa está disponível gratuitamente em http://www.megasoftware.net/, com o fornecimento de um excelente manual de instrução de uso *online*.

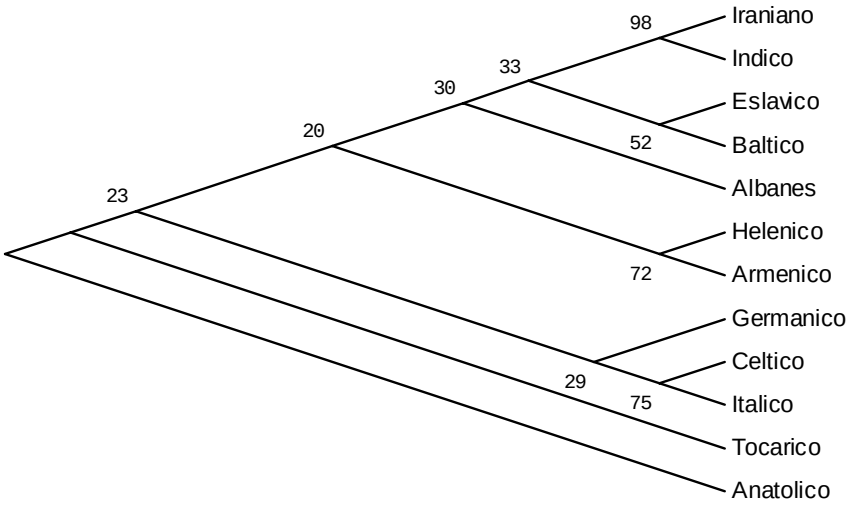

Figura 1: Árvore de Máxima Parcimônia das Línguas Indo-Européias.

A topologia da árvore, na Figura 1, é consistente com os grupos Indo-Europeus tradicionais: Indo-Iraniano e Balto-eslávico (Campbell, 1998). Todos os grupos apresentados nessa árvore são monofiléticos. Estudos recentes de parcimônia e compatibilidade também detectam os grupos acima e ainda o "super grupo" Ítalo-Celta-Germânico e o Heleno-Armênico em concordância com Rexová et alli (2003) e Gray & Atkinson (2003).

A árvore da Figura 1 também nos mostra a posição basal do Tocário, tal como expresso em Ringe et alli (2002). Essa árvore mostra a existência de um "super grupo *satem*", mas não indica claramente um "grupo *centum*". A figura arbórea mostra ainda que o Grego tem mais afinidade com as línguas do "grupo *satem*". Isso indica que a divisão em línguas *centum* e *satem* é puramente arbitrária e não uma divisão orgânica. Aliás, a Lingüística Histórico-Comparativa reconhece as dificuldades de uma classificação *centum-satem* e nosso método confirma essa opinião. Meillet (1922) dedica um capítulo inteiro dessa obra, o Capítulo XX, ao grupo Germano-Ítalo-Céltico e, em sua conclusão, o descreve como um grupo natural. A árvore expressa na Figura 1 confirma essa hipótese de Meillet (1922).

No trabalho de Kroeber & Chretien (1937), encontramos o Germânico mais próximo do grupo Balto-Eslávico, contrariando a hipótese de Meillet (1922). Como os dados que usamos são essencialmente os mesmos,

exceto pela introdução do Albanês, do Hitita e do Tocário, é provável que o método estatístico usado por Kroeber & Chretien (1937) seja falho, conforme os mesmos chegam a conjeturar em seu trabalho.

Na Figura 2 e na Figura 3, a seguir, apresentamos os resultados dos cálculos usando, para tanto, os métodos alternativos de mínima evolução (*minimum evolution*) e junção de vizinhos (*neighbor-joining*), respectivamente. Para a obtenção das árvores expressas nas Figuras 2 e 3, usamos os mesmos dados e parâmetros do método MP. É muito importante ressaltar que Gray & Atkinson (2003), ao estudarem as línguas IE, utilizaram métodos puramente lexicais, ou seja, listas de Swadesh de 200 palavras para as línguas IE e obtiveram os mesmos subgrupos que os de nossa pesquisa. Isso, a propósito, aponta para um resultado muito encorajador quando se trata de classificação de línguas indígenas, pois dados gramaticais de línguas extintas desse tipo são escassos, isso quando não inexistentes. Para essas línguas, temos disponíveis apenas listas de palavras. A equivalência de nosso trabalho usando dados morfológicos e fonológicos com um trabalho que usa dados lexicais para uma família bem conhecida, ou seja, a Indo-Européia, nos encoraja a usar para a classificação das línguas indígenas, métodos cladísticos da Biologia, usando apenas dados lexicais.

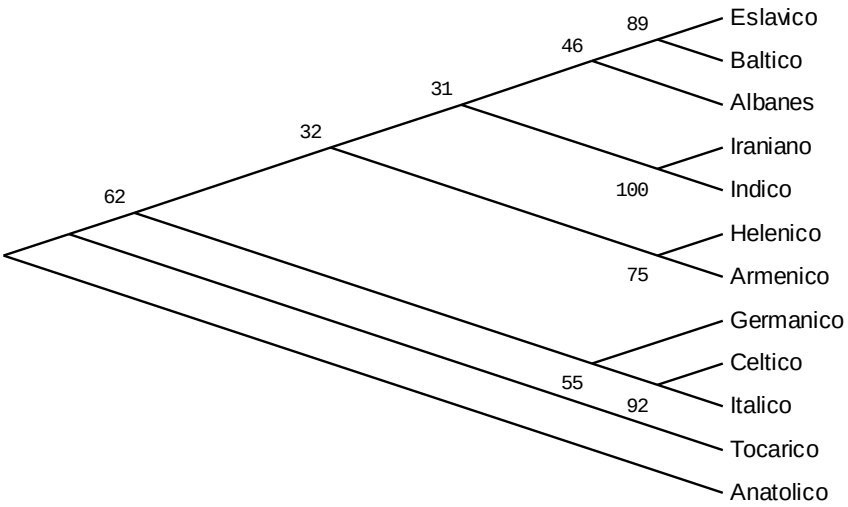

Figura 2: Árvore de Evolução Mínima das Línguas Indo-Européias.

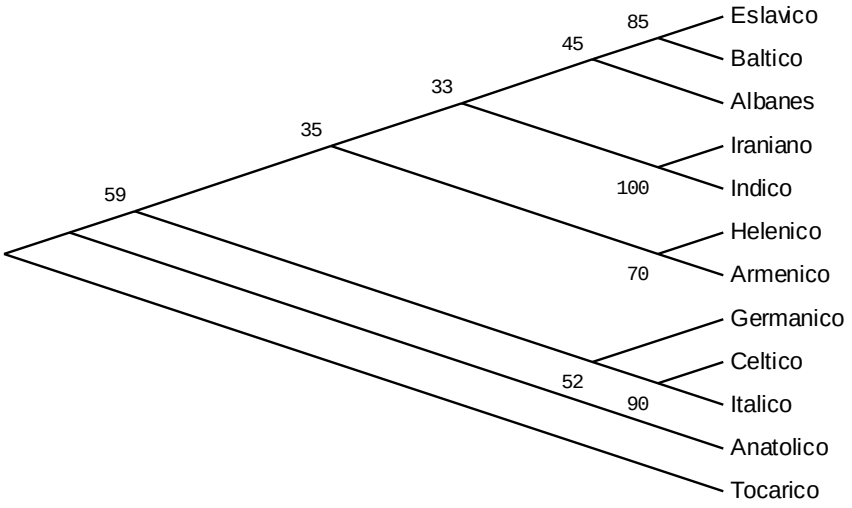

Figura 3: Árvore de Junção de Vizinhos Línguas Indo-Européias.

Em um próximo trabalho, pretendemos elaborar uma lista de itens mista na qual teremos dados lexicais, fonológicos, morfológicos e sintáticos para as línguas IE. Outra direção poderá ser também a utilização de um método mais moderno de obtenção de árvores, a saber, o Método de Máxima Verossimilhança com Inferência Bayesiana da Filogenia (Huelsenbeck & Ronquist, 2001), o qual utiliza como ponto de partida árvores de MP e um algoritmo baseado no método de Monte-Carlo para limitar o universo imenso das árvores produzidas pelo método original a algumas amostras mais significativas.

Esse método foi o usado por Gray & Atkinson (2003) para as línguas IE com dados lexicais. De qualquer forma, acreditamos que os métodos de tratamento de dados na Biologia vão evoluir e talvez seja oportuno que lingüistas que tenham formação matemática e computacional comecem a se unir e produzir programas que sejam específicos para o uso na Lingüística. Urge realizar essa tarefa, deixando de lado barreiras artificiais que a ciência não permite. Que a afirmação de Kroeber & Chretien (1937) seja levada em consideração, diante dos fatos que ora apresentamos, ou seja, as análises estatísticas podem contribuir no sentido de *"validate and correct insight,or, where insight judgments are in conflit, help to decide between them. In short, it increases objectivity, sharpens findings, and sometimes forces new problems".*

Agradecimentos

Expressamos nossa imensa gratidão ao professor Douglas Q. Adams, da Universidade de Idaho, que de maneira muito gentil corrigiu os dados do Tocário apresentados no presente texto.

Recebido em outubro de 2005
Aprovado em abril de 2006
E-mail: almir.bh@terra.com.br

Referências Bibliográficas

Adams, D. Q. 1984. The position of Tocharian among other Indo-European Languages. *Language* 3. Vol. 104: 395-402.

_____. 1988. *Tocharian Historical Phonology and Morphology*. American Oriental Series. Eisenbrauns. pp. 199.

Brugmann, K. & B. Delbrück. 1897-1916. *Gundries der Vergleichenden Grammatik der Indogermanischen Sprachen*. 2 ed. Strasburg.

Campbell, L. 1998. *Historical Linguistics: an Introduction*. Cambridge, Massachusets: The MIT Press.

Croft, W. 2000. *Explaining Language Change - An Evolucionary Approach*. Singapure: Longman Linguistics Library/Pearsons Education Ltd.

Dixon, R. M. W. 1997. *The Rise and Fall of Languages*. Cambridge: Cambridge University Press.

Darwin, C. 1871. *The Descent of Man*. London: Murray.

Dunn, M., A. Terril, G. Reesink, R. A. Foley & S. C. Levinson. 2005. Structural Phylogenetics and the Reconstruction of Ancient Language History. *Science* 75743. Vol 309: 2072-2075.

Felsenstein, J. 1978. Cases in Which Parsimony or Compatibility Methods will be Positively Misleading. *Systematic Zoology* 27: 401-410.

Fitch, W. M. 1971. Towards Defining the Course of Evolution: Minimum Change for a Specific Tree Topology. *Systematic Zoology* 20: 406-416.

Foster, P. & A. Toth. 2003. Towards a Phylogenetic Chronology of Ancient Gaulish, Celtic, and Indo-European. *Proc. Ntl. Acad. Sci.* 15. Vol. 100: 9079-9084.

Gamkrelidze, T. V. & V. V. Ivanov. 1995. *Indo-European and the Indo-Europeans*. Trends in Linguistics 80. Berlim: Moulton de Gruyter.

Gray, R. D. & F. M. Jordan. 2000. Language Tree Supports the Espress-Train Sequence of Austronesian Expansion. *Nature* **405:** 1052-1055.

Gray, R. D. & Q. D. Atkinson. 2003. Language-Tree Divergence Times Support the Anatolian Theory of Indo-European Origin. *Nature* **426**: 435-438.

Holden, C. J., (2002). Bantu Language Tree Reflects the Spread of Faeming Across su-Saharan Africa: a Maximum Parsimony Analysis. *Proc. R. Soc. London* **269**: 793-799.

Huelsenbech, J. P. & F. Ronquist. (2001). MRBAYES: Bayesian Inference of Phylogeny. *Bioinformatics* **17**: 754-755.

Kiekens, E. 1933. *Einffühung in Die Indogermanische Sprachwissenschaft*. Vol. 1, Munich.

Kroeber, A. L. & C. D. Cheretien. 1937. Quantitative Classification of Indo-European Languages. *Language* **2**. Vol. 13: 83-103.

_____. 1939. The Statistical Technique and the Hittite**.** *Language* **2**. Vol. 15: 69-71.

Kumar, S., K. Tamura & M. Nei. 2004. MEGA3: Integrated Software for Molecular Evolutionary Genetics Analysis and Sequence Alignment. *Briefings in Bioinformatics* **5**: 150-163.

Mann, S. E. 1941. The Indo-European Semivowels in Albanian. *Language* **1**. Vol. 17: 12-23.

_____. 1950. The Indo-European Vowels in Albanian**.** *Language* **3**. Vol. 26: 379-388.

_____. 1952. The Indo-European Consonants in Albanian. *Language* 1. Vol. 28: 31-40.

_____. 1977. *An Albanian Historical Grammar*. Buske: 1 Aufl Edition. pp. 239

Meillet, M. 1922. *Lés Dialectes Indo-Européens.* Paris.

_____. 1934. *Introduction à l'Étude Comparative des Langues Indo-Européens*. 7 ed, Paris.

Pagel, M. 2000. Maximum-Likelihood Models of Glottochronology and for Reconstructing Linguistic Phylogenies. In: C. Renfrew, A. Mcmahon & L. Trask (eds.). *Time Depth in Historical Linguistics*. Cambridge: The McDonald Institute for Archaeological Research. p. 413-439.

Rexova, K., D. Frynta & J. Zrzavý. 2003. Cladistic Analysis of Languages: Indo-European Classification Based on Lexicostatistical Data*. Cladistics* **19**: 120-127.

RINGE, D., T. Warnow & A. Taylor. 2002. Indo-European and Computational Cladistics. *Trans. Philos. Soc.* **100**: 59-129.

SCHMIDT, J. 1872. *Die Verwandtschaftsverhältnisse der Indogermanischen Sprachen*. Weimar.

SCHLEICHER, A. 1861. *Compendium der Vergleichenden Grammatik der Indogermanischen Sprachen: Kurzer Abriss einer Laut-und Formenlehre der Indogermanischen Ursprache.* Weimar: Hermann Böhlau.

SOMMER, F. 1914. *Handbuch der Lateinischen Laut-und Formenlehr.* 2 e 3 edições. Heidelberg.

SWADESH, M. 1952. Lexico-Statistic Dating Prehistory Ethnic Contacts with Special Reference to North American Indins and Eskimos. *Proc. Am. Philos. Soc*. **95**: 453-462.

WARNOW, T. 1997. Mathematical Approaches to Comparative Linguistics. *Proc. Natl. Acad. Sci.* **94**: 6585-6590.