Sesión 5: Filolingüística

Carlos Ugarte

04.09.2025

Por favor, descarguen git, entren al software PyCharm CE y repitan el proceso visto en la anterior clase hasta que su directorio de trabajo sea la carpeta clonada de este taller. Para esto, estando en la carpeta Documents, sigan los siguientes comandos:

```
mkdir proyectos
cd proyectos
mkdir taller
cd taller
git clone https://github.com/MuffinLinwist/taller_pucp25.git
```

Para quienes ya tengan la carpeta en sus computadoras, por favor solo actualicen el repo a su última versión con el siguiente comando:

```
git fetch origin main
```

Por favor, entren a la página web de Densitree y de Figtree y descarguen los software

1 Introducción

Las filogenias lingüísticas son usadas como la base histórica en modelos cuantitativos de cambio lingüístico. A partir de ellas, podemos evaluar hipótesis de rutas de dispersión humanas, process de cambio cultural, evolución de otros subsistemas lingüísticos, etc. Los análisis filolingüísticos suelen estar basados en modelar el comportamiento histórico de set de cognados léxicos. Esto se hace típicamente a partir de la recolección data de listas de Swadesh en las lenguas bajo estudio. Sin embargo, muchos son los trabajos que también incorporan entradas de data gramatical para modelar relaciones filogenéticas entre lenguas.

El punto de partida para la comparación de cognados léxicos luce así:

		Hup	Yuhup	Dâw	Nadëb
1	I (1sg)	*~ʔáh*	*~ʔăh*	*ʔãh*	*ʔɨ̃h*
2	you (2sg)	*~ʔab*	*~ʔăb*	*ʔãm*	*ʔɔ̃m*
3	we (1pl)	*~ʔɨ́d*	*~ʔɨ̌d*	*ʔid*	*ʔəːr* (INCL)
8	not	*~pă*	--	*mæ̃h*	*péh*
13	big	*pŏg*	*pôg*	*peg (pôg)*	*poŋ*
14	long	*wʼăt*	*wʹət*	*wʹət*	*dawɨːt*
16	woman	*~taʔáj*	*~ʔâːj*	*ʔɑ̃j*	*ʔɨ̃ːɲ, ʔɨ̃ːj*
19	fish (n.)	*~hŏp*	*~hôp*	*hɑ̃p*	*tahɨ̰̃ːb*
20	bird	*wĕt* (dove sp.)	*wêt*	*təwăt*	*tawəːd*
22	louse	*~dĕb*	*~dêb*	*nɑ̃̂m*	*naːm*
23	tree	*tĕg*	*têg*	*tâg*	*təːg*
25	leaf	*cugʹĕt, -gʹæt*	*kʹêt*	*kĕt*	*gɑ̰ːd*
27	bark	*bʹɔ́k*	*bʹɔ̆k*	*bɨk*	*bɔ̰g*
29	flesh	*dʹáp*	*dʹăp*	*dæp*	*dab*
30	blood	*bijĭw*	*dîw*	*jɨ̂w*	*mãjɨːw*
33	egg	*tĭp*	*tîp*	*tɨ̆p*	*tɨb*
34	horn	*~căd'*	*~câd'*	*ʃɑ̃̆n'*	*ʃɑ̰ːn*
35	tail	*dʹŭb*	*dʼûb*	*dum*	*dõm*
37	hair	*~căd* ('pubic hair')	*~câd* ('pubic hair')	*ʃân* ('pubic hair')	*ʃəːn*
38	head	*~dúh*	*~dŭh*	*nũh*	*nuːh*
42	mouth	*~dɔ-gʹŏd*	*~dɔh-kʹôd*	*nɔ̃h*	*nɔːh*
43	tooth	*tág*	*tăg*	*təg*	*təg*
44	tongue	*~dɔ-gʹĕd*	*~dɔh-kʹêd*	*nɔ̃hkɑ̂d*	*nagɑ̰ːd*
46	foot	*ɟʼĭb*	*cʹîb*	*cîm*	*ɟɨːm*
54	drink	*ʔəg-*	*ʔəg*	*ʔəg*	*ʔəːk, ʔəŋ*

Figure 1: Lista de vocabulario Swadesh para la Familia lingüística Nadahup (tomada de Epps and Bolaños 2017)

Se pasa de lexemas a clases de cognados. Las formas reciben la misma codificación si comparten un ancestro en común. Se le asigna un carácter por significado. Así llegamos a la siguiente matriz:

	Meaning 1		Meaning 2	
	lexeme	class	lexeme	class
Language A	mhɨm	a	cɨŋ	x
Language B	mhɨm	a	kɨt	y
Language C	lɔ:t	b	kət, lpəc	y, z
Language D	?	?	lpət	z

Figure 2: Matrices multiestado de dos significados (tomadas de Dunn 2015)

El siguiente paso es convertir la matriz de multiestados en una matriz binaria, ya que estas suelen desempeñarse de mejor manera que las matrices de multiestados en diversos estudios filogenéticos. La matriz binaria es una matriz que marca la presencia o la ausencia de un carácter en las lenguas.

	Cognate set				
	1a	1b	2x	2y	2z
Language A	1	0	1	0	0
Language B	1	0	0	1	0
Language C	0	1	0	1	1
Language D	?	?	0	0	1

Figure 3: Codificación binario de data de cognados (tomada de Dunn 2015)

El carácter en este caso es la clase de cognado. A partir de estos resultados, ya podemos correr análisis filogenéticos.

2 Métodos para inferir filogenias

Una primera distinción a hacer es el uso de dos tipos de metodos. El primero comprende los métodos con modelos de cambio basados en la distancia. Este estima la cantidad de cambio entre dos lenguas a partir de la cantidad de diferencia entre ellas (Dunn 2015). Todos estos métodos utilizan alguna métrica de distancia para medir cuánto difiere cada taxon de los demás.

En la lingüística se utilizan más comúnmente dos: La proporción de cognados compartidos y la distancia de Levenshtein (una medida palabra por palabra de similitud fonética). El segundo tipo comprende los métodos con modelos de cambio basados en caracteres. Estos métodos estiman la relación entre dos lenguas infiriendo los caminos por los que ambas evolucionarion de un mismo ancestro (**Dunn**).

Nosotros en esta sesión entenderemos y compararemos los resultados de usar dos métodos de cada uno de los tipos en un subset del dataset NORTHPERULEX. Primero, utilizaremos el método de agrupamiento UPGMA (UPGMA artículo Wikipedia), que es un método basado en el promedio aritmético de sus distancias. Segundo, utilizaremos métodos bayesianos, en la versión del programa "MrBayes" (MrBayes sitio web oficial), que es un programa con inferencia bayesiana que calcula la plausibilidad de escenarios evolutivos.

Para poder correr todo el flujo de trabajo de esta sesión, vamos a crear un VENV para instalar todos los paquetes necesarios. Para eso, en nuestra carpeta TALLER, corremos los siguientes comandos.

```
python -m venv phylovenv
.\phylovenv\Scripts\activate
Set-ExecutionPolicy -Scope Process -ExecutionPolicy Bypass
.\phylovenv\Scripts\activate
pip install lingpy
pip install lingrex
```

Descargamos los datos de NORTHPERULEX para poder utilizarlos:

```
git clone https://github.com/lexibank/northperulex.git
cd northperulex
git checkout distance_analysis
cd analysis
```

Corremos el script "s_align.py" en esta carpeta con el siguiente comando:

```
python s_align.py
```

Este script recoge la data presente en NORTHPERULEX, así como extrae un subset de lenguas con mayor número de palabras para conceptos en común. Dos son los outputs al correr el script. El primero es un archivo con un árbol creado en base a las distancias entre palabras por COGID. El segundo es un archivo en formato NEXUS especial para correr filogenias bayesianas.

Con el segundo archivo generado, pasamos a trabajar con MrBeast.

3 Trabajar con MrBayes

En nuestras terminales, tenemos que inicializar MrBayes (ya descargado en nuestras computadoras) y ejecutar el archivo nexus para que el software pueda leer su contenido. Luego, tenemos que especificar el modelo evolutivo y correr el análisis. Para el modelo evolutivo, nosotros seguiremos la configuración default encontrada en el manual de documentación del software. Al final resumiremos las muestras.

Lo hacemos con los siguientes comandos:

```
mb
execute npl.nex
lset nst=6 rates=invgamma
mcmc ngen=20000 samplefreq=100 printfreq=100 diagnfreq=1000
```

Se para el análisis cuando la deviación estándar de frecuencias partidas está debajo de 0.01. Sino, añadimos otras iteraciones más. Por cuestión de tiempo, paremos ahora.

```
sump
sumt
```

El primer comando resume los valores de los parámetros. El programa generará una tabla con resúmenes de las muestras de los parámetros del modelo de sustitución, incluyendo la media, la moda y el intervalo de credibilidad del 95% (región de máxima densidad posterior, HPD) de cada parámetro.

Aquí lo importante son dos puntos. Hay que asegurarse que el factor de reducción de escala potencial (PSRF) sea cercano a 1.0 para todos los parámetros. El otro es el tamaño efectivo de la muestra (ESS). Todos los parámetros deben ser superiores a 100. Si cualquiera de estos puntos no se cumplen, hay que volver a correr el análisis por más tiempo.

References

Dunn, Michael (2015). "Language phylogenies". In: *The Routledge Handbook of Historical Linguistics*. Ed. by Claire Bowern and Bethwyn Evans. New York: Routledge, pp. 190–211.

Epps, Patience and Katherine Bolaños (2017). "Reconsidering the "Makú" Language Family of Northwest Amazonia". In: *International Journal of American Linguistics* 83.3, pp. 467–507.