Sesión 4: Tipología léxica

Carlos Ugarte

03.09.2025

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
```

1 Repaso de EDICTOR

Una vez que hayan clonado/actualizado la carpeta del taller en las computadoras, procedemos a abrir el archivo "data/quechua_modern_wordlist.tsv". También abrimos el servidor web de EDICTOR (https://edictor.org). Seguimos los siguientes pasos:

1. En la ventana que se abre por default "Home", le hacemos click a a la opción "GET STARTED". Se abrirá otra ventana en su servidor web.
2. Subimos el archivo de data quechua que descargamos previamente.
3. Agreguemos una columna llamada "Alignments".

Hemos subido el archivo exitosamente a EDICTOR para poder trabajar con él. La columna "Alignments" es necesaria porque, una vez que generemos automáticamente los alineamientos, el algoritmo tendrá una columna donde almacenar esa información.

4. Seleccionamos la cabecera "COMPUTE".
5. Computamos "ALIGNMENTS". Con esto se abrirá, abajo de la ventana de Wordlist, la ventana de "Compute Alignments".
6. Estará predeterminada la configuración default del algoritmo para hacer un alineamiento automático de la data. Le damos "SUBMIT".

Ahora, desde el servidor web, hemos generado automáticamente los alineamientos para nuestra data. Con esto, ya tenemos una buena base para empezar. Repetimos ahora el proceso para conseguir los patrones de correspondencias:

7. Seleccionamos la cabecera "COMPUTE".
8. Computamos "CORRESPONDENCE PATTERNS". Con esto se abrirá, abajo de la ventana de Wordlist, y abajo de la ventana de "Compute Alignments', la ventana de "Compute Correspondence patterns".
9. Nuevamente, estará predeterminada la configuración default del algoritmo para hacer un alineamiento automático de la data. Le damos "SUBMIT".

Aquí podemos observer los patrones de correspondencias de sonidos. A partir de esto, podemos conllevar muchos análisis y experimentos dependiendo de en qué estemos interesados.

2 Introducción: colexificaciones

La colexificación, término postulado por (François 2008), se refiere a una situación en la que dos o más significados están expresados en una lengua por un mismo elemento léxico. La colexificación implica una relación conceptual entre lexemas, pero no distingue entre polisemia o vaguedad semántica, ni se pronuncia sobre la relación diacrónica entre los significados que comprende un lexema.

En esta sesión, emplearemos la data presente en la última versión de la base de datos CLICS^4 (Tjuka et al. 2025) para ejemplificar cómo crear visualizaciones de redes. Hay muchos libros (Cairo 2016) y herramientas que sirven para tal propósito, pero nosotros nos concentraremos en uno: Cytoscape (https://cytoscape.org).

Las redes se utilizan, en general, para visualizar la relación entre dos elementos dentro de nuestra data. Para la tipología léxica, y más específicamente para la semántica léxica, el uso de redes está actualmente muy empleado para demostrar la relación entre dos conceptos a través de la colexificación.

3 Usando CLICS^4

Para powder 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 clicsvenv
.\clicsvenv\Scripts\activate
Set-ExecutionPolicy -Scope Process -ExecutionPolicy Bypass
.\clicsvenv\Scripts\activate
```

Ahora, descargamos los datos de CLICS^4 para poder utilizarlos:

```
git clone https://github.com/clics/clics4.git --branch=patch --depth=1
cd clics4
git checkout v0.5
```

Con esto, ya podemos usar la data de CLICS^4. Lo que necesitamos es ahora procesar la data y generar un archivo que nos de información sobre las colexificaciones halladas en esta base de datos. Para ello, escribí un pequeño script, que va a demorar un rato en correr, llamado "clicscolex.py" y que se encuentra en la carpeta "taller_pucp25/codigo/". Para correrlo, escribimos el siguiente comando en nuestra terminal:

```
python taller_pucp25/codigo/clicscolex.py
```

Esto habrá generado un archivo output con colexificaciones halladas en la base de datos con una ruta relativa: "taller/taller_pucp25/data/clicscolexificaciones.tsv".

Una vez generado el archivo, pasamos a abrir la aplicación de Cytoscape. Podrán ver que ya hay sesiones de ejemplo para explorar y familiarizarse un poco más con la aplicación. Como nosotros queremos crear una red desde un archivo, le damos click a la opción "Importar" (Import) y luego "Red desde un archivo" (Network from file).

Ya existen sesiones de ejemplo que puedes explorar. Si quieres crear tu red desde un archivo, puedes usar la opción "Importar red desde archivo".

Los formatos de archivo posibles incluyen archivos .gml o formatos de tabla como .csv. Al importar un archivo .gml, los nodos y las aristas ya están definidos, por lo que la red se importa inmediatamente. Si utiliza un formato de tabla simple (.csv o .tsv), se deberá definir los nodos de origen y destino. En este caso, "Concepto 1" sería el origen y "Concepto 2" el destino. La columna "Colexifica" incluye las frecuencias de colexificaciones entre los dos conceptos en los idiomas representados por las aristas de una red.

References

François, Alexandre (2008). "Semantic maps and the typology of colexification: Intertwining polysemous networks across languages". In: *From Polysemy to Semantic Change. Towards a typology of lexical semantic associations*. Ed. by Martine Vanhove. Studies in Language Companion Series 106. Amsterdam/Philadelphia: John Benjamins Publishing Company, pp. 163–215. URL: https://doi.org/10.1075/slcs.106.09fra.

Cairo, A. (2016). *The Truthful Art: Data, Charts, and Maps for Communication*. Voices That Matter. Pearson Education. ISBN: 9780133440539. URL: https://books.google.com.pe/books?id=8dKKCwAAQBAJ.

Tjuka, Annika et al. (2025). "CLICS: An Improved Database of Cross-Linguistic Colexifications [Dataset, Version 0.4]". In: URL: https://github.com/clics/clics4/.