Un corpus sintáctico del idioma uspanteko

Robert Henderson, Luis A. Irizarry-Figueroa, Francis Tyers

Introducción

En este proyecto se realizó la anotación de un corpus sintáctico del idioma uspanteko.

- Usamos el corpus para capacitar unos analizadores para el idioma.
- Demostramos que podemos utilizar datos de idiomas de la misma rama, como el k'iche', para mejorar la precisión del analizador.
- Estos analizadores híbridos podrían ayudar a los equipos pequeños a realizar anotaciones para un corpus sintáctico.

Uspanteko

El uspanteko (o Tz'únun Tziij) es un idioma Maya de la rama k'iche' hablado por alrededor de 1200-4000 personas en la región de San Miguel de Uspántan en Guatemala.

- Está en peligro ya que muchos niños en Uspantán están aprendiendo a hablar el k'iche o el español.
- Mientras que el sistema de flexión lleva muchas diferencias de lo que se encuentra en k'iche', y la pronunciación es muy diferente (incluyendo el uso de tono léxico), tienen muchas raíces compartidas, la estructura de la cláusula es similar, y en la forma escrita, se ven muy parecidos.
- Son importantes las similitudes porque nuestra idea es usar k'iche' para 'bootstrap' o impulsar el analizador uspanteko.

Bibliografía

[1] Can Pixabaj, T. A. (2007). Jkemiik yoloj li Uspanteko: Gramática Uspanteka. Cholsamaj.

[2] Alexis Palmer et al. (2010). Computational strategies for reducing annotation effort in language documentation. *Linguistic Issues in Language Technology*, 3(4):1–42.

[3] Tyers, F. and Henderson, R. (2021) A corpus of K'iche' annotated for morphosyntactic structure. *NAACL-HLT 2021*, pages 1–10.

[4] Strake et al. (2016) UDPipe: trainable pipeline for processing CoNLLU files performing tokenization, morphological analysis, POS tagging and parsing. *LREC'16*.

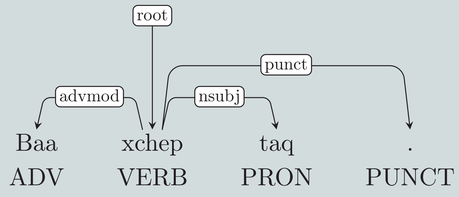
Sobre el corpus

El corpus del uspanteko utilizado fue tomado de Palmer (2010) en el que posteriormente se realizó el anotaje.

- El corpus de Palmer estaba en un formato XML para codificar glosas interlineales. Por esta razón, tuvimos acceso a las traducciones, así como a las partes de la oración para las palabras uspantekas.
- El primer paso fue convertir este corpus de su formato XML al formato CONLLU que se utiliza para representar, en texto, los grafos que constituyen un análisis UD.

El segundo paso fue anotar 5,000 tokens del corpus. Esto correspondía a 417 oraciones (∼12 tokens por oración), que dos coautores pudieron anotar durante dos meses trabajando a tiempo parcial.

```
# sent_id = 2.xml:22
# text = Baa xchep taq.
# text[spa] = Vaya lo agarraron.
# text[orig] = Baa xchep taq.
# labels = T16_P43
1       Baa     _       ADV     _       _       2       advmod  _       Gloss=voc
2       xchep   _       VERB    _       Aspect=Perf|Valency=2   0       root    _       Gloss=COM-vt
3       taq     _       PRON    _       Number=Plur     2       nsubj   _       Gloss=PL
4       .       _       PUNCT   _       _       2       punct   _       Gloss=.
```



Los experimentos

Entrenamos a dos modelos. El primero utilizó solo datos de uspanteko. El segundo utilizó una combinación de uspanteko (200 oraciones) y k'iche' (1430 oraciones), que se probó con 200 oraciones uspantekas retenidas.

- Usamos un modelo listo para usar—UDPipe 1.2 Straka et al. 2016—y el conjunto estándar de pruebas del proyecto UDpipe.

Modelo Uspanteko			
Tokens	75.45	Lemmas	75.63
Sentences	92.69	UAS	44.77
Words	75.63	LAS	37.87
UPOS	73.66	CLAS	42.94
XPOS	75.63	MLAS	40.06
UFeats	75.18	BLEX	42.94
AllTags	73.62		

Modelo Uspanteko más K'iche'			
Tokens	97.72	Lemmas	97.53
Sentences	98.37	UAS	82.09
Words	97.53	LAS	76.53
UPOS	84.92	CLAS	75.06
XPOS	97.53	MLAS	59.41
UFeats	90.27	BLEX	75.06
AllTags	82.73		

Columnas son la media del puntaje F_1 de validación cruzada 10 veces: '**Tokens**' la tokenización; '**Words**' separando paralabra sintácticas (e.g las contracciones); '**Lemmas**' la lematización; '**UPOS**' etiquetas universales de las partes-del-discurso; '**Feats**' rasgos morfológicos; '**UAS**' el puntaje adjunto no etiquetado (nucleos de dependencia); '**LAS**' el puntaje adjunto etiquetado (núcleos de dependencia y sus relaciones).

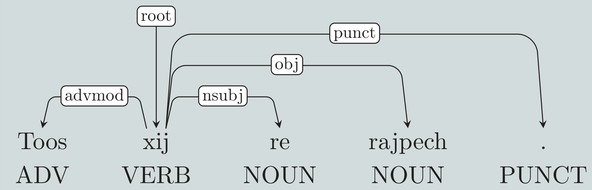
Interpretación

Podemos ver que el modelo mixto es mejor en todas las métricas.

- En particular, vemos que las métricas UAS y LAS, que corresponde a la selección de los dependientes de cada nodo y el nombre del tipo de la dependencia mejora mucho.
- Atribuimos las mejoras al hecho que la estructura de la cláusula del k'iche' es muy similar al uspanteko. También comparten muchas raíces nominales y verbales de alta frequencia.

Dependencias Universales

- El proyecto de dependencias universales consiste en la elaboración de corpus sintácticos anotados para una variedad de lenguages, con el propósito de faciliar el análisis sintáctico, el aprendizaje interlingüistico y la investigación sintática desde una perspectiva tipológica.
- En escencia el projecto intenta proveer un inventario de categorías y guías para facilitar el anotaje de construcciones sintácticas similares en las distintos idiomas pero proveyendo la oportunidad de realizar extensiones específicas de algún idioma en particular.



Conclusión

- El analizador sintáctico muestra la mejor eficiencia cuando se capacita o entrena con oraciones mezcladas del k'iche y del uspanteko.
- Para futura investigación se anotaran mas tokens para realizar el experimento únicamente con datos del uspanteko y determinar si la baja eficiencia se debe a la extensión del corpus actual.
- No obstante, a pesar de que los resultados del experimento con data del uspanteko no son muy altos, el haber realizado este proyecto nos facilita nuestro trabajo futuro, ya que al utilizar el analizador, este nos permite utilizar los análisis que produce para corregir y expandir y el corpus sin tener que comenzar desde cero.
- Igualmente este proyecto impacta directa e indirectamente la comunidad de académicos de los idiomas mayas, ya que esta información queda disponible a la comunidad para examinar y poder contrastarla con investigaciones concurrentes, entre otros usos de utilidad para la expansión del conocimiento lingüistico.