بەشی پراکتیک 1: ئانالیزی کۆرپس

خولی زمانەوانیی کۆمپیوتەری

https://sinaahmadi.github.io/KurdishCL/

سینا ئەحمەدی

ئەرکە پراکتیکەکانی ئەم بەشە لە بەشی دووی خولەکەدا باس کراون.

پێش لە هەموو شتێک، کۆرپسێک کە بە تایبەت بۆ ئەم خولە کۆ کراوەتەوە و نموونەیەکی بچووکە لێرەدا دابەزێنە: بەستەر. ئەم کۆرپسە پێنج بنزاراوەی کوردی لە خۆ دەگرێ: سۆرانی (ckb)، کورمانجی (kmr)، زازا (zza)، هەورامی (hac) و کوردیی باشوور (sdh). لە پاش داگرتنی کۆرپسەکان، بەرنامەی AntConc لەم بەستەرەوە دابەزێنە و کۆرپسێکی تێدا بکەوە.

ئەرکەکان

1.0 ڕێنووس و وشە

1. لە ڕێگای کۆرپسەوە، ئەمە بسەلمێنە: «هیچ وشەیەکی کوردیی سۆرانی نییە کە بە ر دەست پێ بکا»
2. لە ڕێگای کۆرپسەوە، ئەمە بسەلمێنە: «هیچ وشەیەکی کوردیی سۆرانی نییە کە بە ڵ دەست پێ بکا»
3. لە خشتەی خوارەوە هەندێک وشەی هاوواتا هێنراون کە بە شێوازی جیاواز دەنووسرێن. لە ڕێگای کۆرپسەوە، فراوانیی وشەکان بدۆزەوە:

وشە	فراوانی
ووڵات	
وڵات	
وەڵام	
وڵام	
بەتایبەتی	
بە تایبەتی	
ئەدا	
ئەدات	
دەدا	
دەدات	

2.0 دەربڕینە فرەوشەییەکان

دەربڕینە فرەوشەییەکان (multiword expression) بەو وشانەی دەوترێ کە لە چەند بەش پێک هاتوون وەک «ئاڵوگۆڕ»، «هەڵیت و پەڵیت»، «هێنانە دەر» و «لە کیسی خۆ دان». بۆ دۆزینەوەی دەربڕینە فرەوشەییەکان دەتوانین لە ئانالیزی کۆرپسەوە کەڵک وەرگرین. هەندێک لە دەربڕینەکان بە یەک شێوە دەردەکەون، وەک ئەوانەی کە بە شێوەی ناو و کردارێک (ناو + کردار) دەردەکەون وەک «کار کردن» یان «تووڕە بوون». نموونەیەکی بەربڵاوی کوردی بۆ وشەسازی، کەڵک وەرگرتن لە کردارێکی سووک لەگەڵ ناوێکە؛ کرداری سووک (light verb) کردارێکە کە لە دەربڕینەکەدا واتاکەی بەتەنیا زۆر گرینگ نییە و لەگەڵ وشەکانی دیکەدا مانا دەگرێ. هەندێک لە کرداڕە سووکەکان لە سۆرانیدا ئەمانەن:

- «**کردن**» وەک «بەش کردن» و «ڕاکردن»
- «**دان**» وەک «دادان» و «هان دان»
- «**بوون**» وەک «سوور بوون» و «ڕابوون»
- «**گرتن**» وەک «هەڵگرتن» و «ماسی گرتن»
- «**کەوتن**» وەک «سەرکەوتن» و «پەک کەوتن»
- «**چوون**» وەک «دەرچوون» و «لە دەست چوون»

1. پیتی پێشدانراو (preposition) وەک «لە، بە، دە، وە» و پاشدانراو (postposition) وەک «دا، ڕا، ەوە» بە چ شێوازێک دەردەکەون؟

2. «ەوە» وەک لە کرداری «کردنەوە» یان «هاتنەوە» چ ڕۆڵێکی هەیە بە پێی ئامارەکانی کۆرپسەکە؟

3. دەربڕینە فرەوشەییەکان کە بە کرداری سووکەوە ساز دەکرێن بدۆزەوە.

3.0 جیاوازیی نێوان بنزاراوەکان (پێشکەوتوو)

1. 20 وشەی پڕفراوانی کە لە سۆرانی، کورمانجی، هەورامی و زازادان چین؟ چۆنیان بەراوەرد دەکەی؟

2. قەبارەی مامناوەندیی وشەکان لە سۆرانی، کورمانجی، هەورامی و زازا چۆن هەڵدەسەنگێنی؟

3. وشەی هاوڕێ (collocation) بەو وشانەی دەکوترێ کە بە فراوانییەکی زۆرەوە پێکەوە دێن و مانایەکی جیاوازیان هەیە. بۆ نموونە، ئەمانە وشەی هاوڕێن: «چەپکێک گوڵ» نەک «کۆمەڵێک گوڵ»، «ئاو خواردنەوە» نەک «ئاو خواردن» و «وەڕینی سەگ» نەک «وەڕینی بەرخ». بەپێی کۆرپسەکە، هەندێک لە وشە هاوڕێیەکانی بنزاراوەی کۆرپسەکە بدۆزەوە.

4. **یاسای زیپف** (Zipf's law) لە ئانالیزی کۆرپسدا دەڵێ ئەگەر وشەکان لە کۆرپسێکدا بە پێی فراوانییان ڕیزبەندی بکەین، وشەی یەکەم دوو ئەوەندە وشەی دووهەم فراوانی هەیە و سێ ئەوەندەی وشەی سێهەم و هتد. کەواتە بوو، ئەو وشەیەی کە لە ڕیزی n دایە، $\frac{1}{n}$ جاری یەکەم وشە فراوانی هەیە.
دەتوانی ئەمە لە کۆرپسە سۆرانی، کورمانجی، هەورامی و زازاکەدا بسەلمێنی؟

سەرچاوە بەسوودەکان

[1] Sina Ahmadi. Building a corpus for the Zaza–gorani language family. In *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 70–78, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics (ICCL). https://aclanthology.org/2020.vardial-1.7.pdf.

[2] Sina Ahmadi. A tokenization system for the Kurdish language. In *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 114–127, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics (ICCL). https://aclanthology.org/2020.vardial-1.11.pdf.

[3] Heather Froehlich. Corpus analysis with antconc. *Programming Historian*, 2015. https://programminghistorian.org/en/lessons/corpus-analysis-with-antconc.