SimRelUz: Similarity and Relatedness scores as a Semantic Evaluation Dataset for Uzbek Language

Ulugbek Salaev[1], Elmurod Kuriyozov[2], Carlos Gómez-Rodríguez[2]

[1] Urgench State University, Uzbekistan (*ulugbek0302@gmail.com*)

[2] Universidade da Coruña, Spain (*{e.kuriyozov, carlos.gomez}@udc.es*)

Semantic Analysis

The process of drawing language-independent meaning from text;
Answers following questions:

- What is the computational meaning of individual words/phrases in context? (Lexical Semantics);
- How can we learn semantic representations from data? (Distributional Semantics).

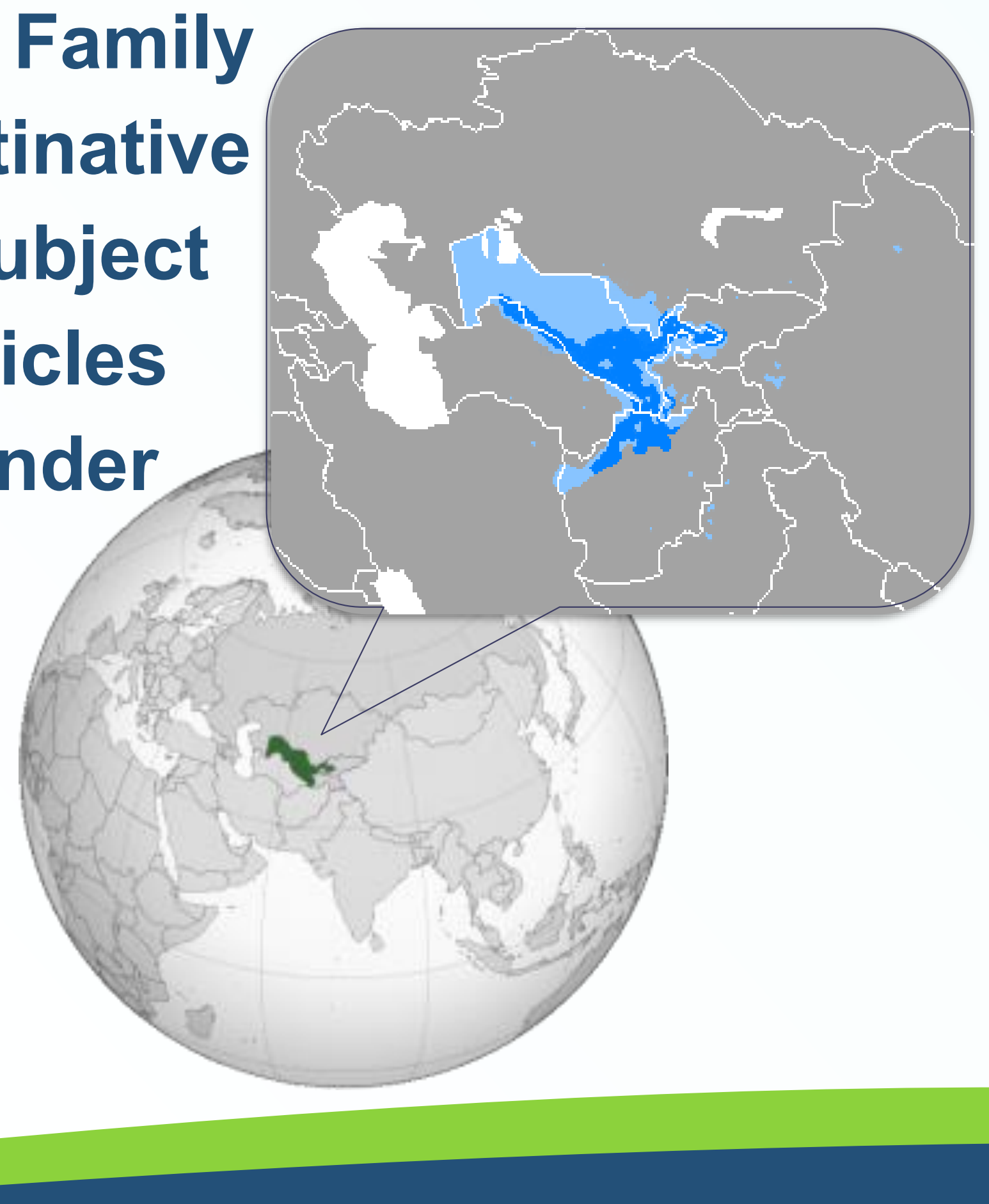
Terminology

- **Semantic Similarity:**
  - Sense of relatedness that is dependent on the amount of shared properties (degree of synonymy)
  - **Example :** Bus - Train
- **Semantic Relatedness:**
  - General sense of semantic proximity or semantic association, regardless of the causes of the connection humans can perceive
  - **Example :** Coffee – Cup

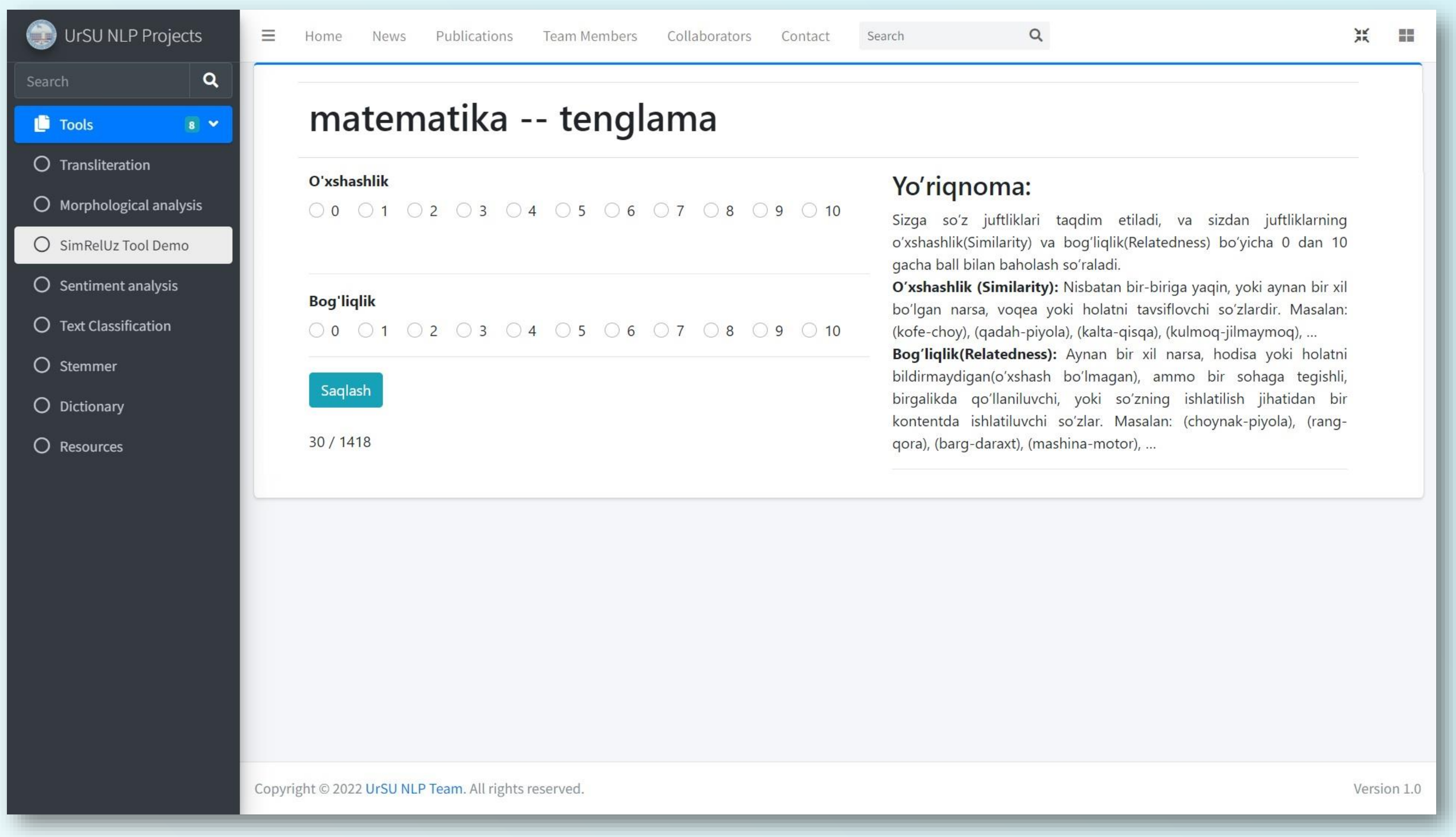
Uzbek Language

- **Official language of Uzbekistan;**
- **Native to All CA Countries, Russia, China;**
- **Spoken by more than 30 million people.**

- **Turkic Family**
- **Agglutinative**
- **Null Subject**
- **No Articles**
- **No Gender**
- **SOV**



Semantic scores Annotation Tool



matematika -- tenglama

O'xshashlik

○ 0 ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7 ○ 8 ○ 9 ○ 10

Bog'liqlik

○ 0 ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7 ○ 8 ○ 9 ○ 10

Saqlash

30 / 1418

Yo'riqnoma:

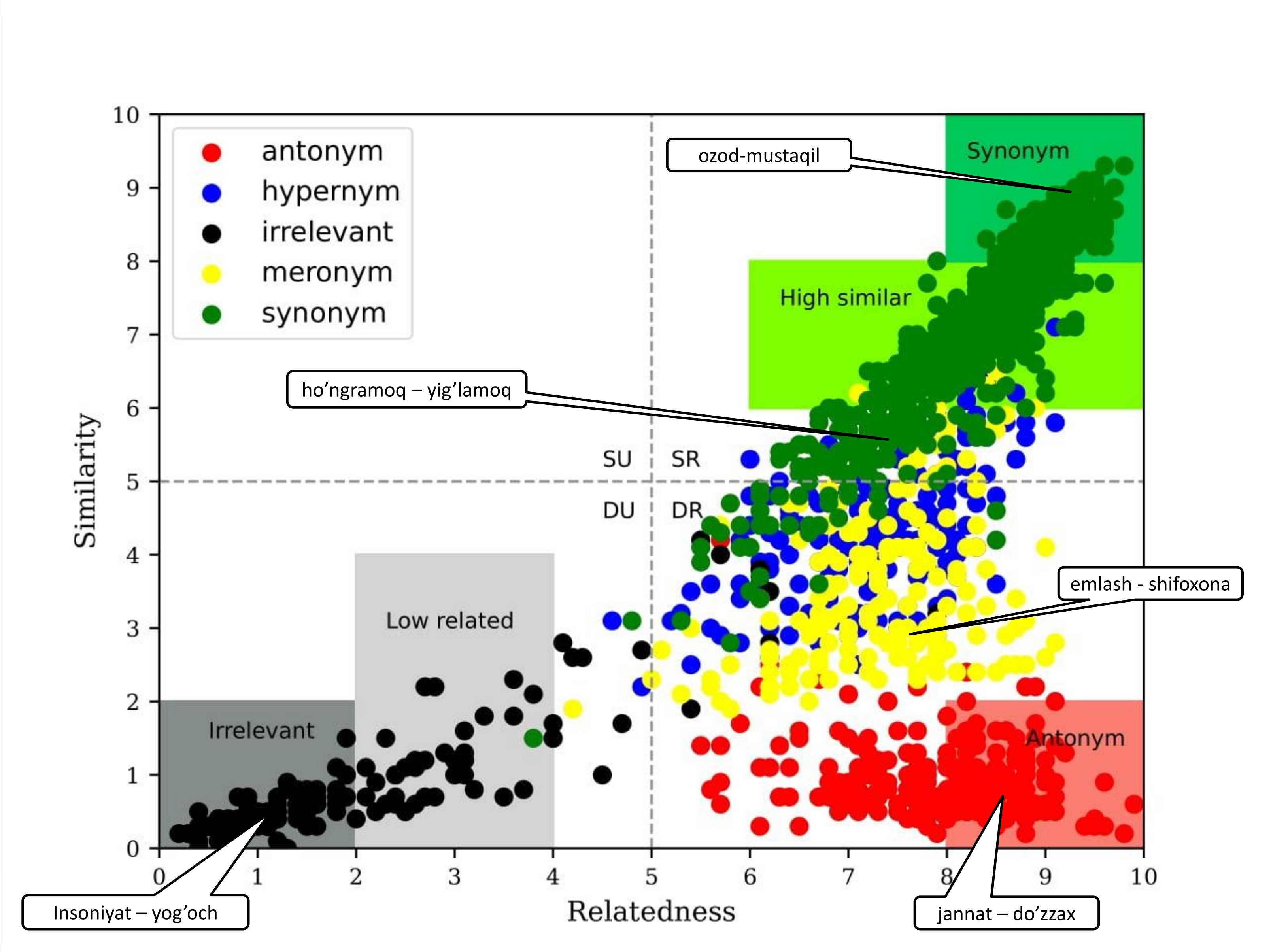
Sizga so'z juftliklari taqdim etiladi, va sizdan juftliklarning o'xshashlik(Similarity) va bog'liqlik(Relatedness) bo'yicha 0 dan 10 gacha ball bilan baholash so'raladi.
O'xshashlik (Similarity): Nisbatan bir-biriga yaqin, yoki aynan bir xil bo'lgan narsa, voqea yoki holatni tavsiflovchi so'zlardir. Masalan: (kofe-choy), (qadah-piyola), (kalta-qisqa), (kulmoq-jilmaymoq), ...
Bog'liqlik(Relatedness): Aynan bir xil narsa, hodisa yoki holatni bildirmaydigan(o'xshash bo'lmagan), ammo bir sohaga tegishli, birgalikda qo'llaniluvchi, yoki so'zning ishlatilish jihatidan bir kontentda ishlatiluvchi so'zlar. Masalan: (choynak-piyola), (rang-qora), (barg-daraxt), (mashina-motor), ...

Copyright © 2022 UrSU NLP Team. All rights reserved.

Version 1.0

- **Open-source, web-based, multi-user tool;**
- **Demo available: *https://simrel.urdu.uz***

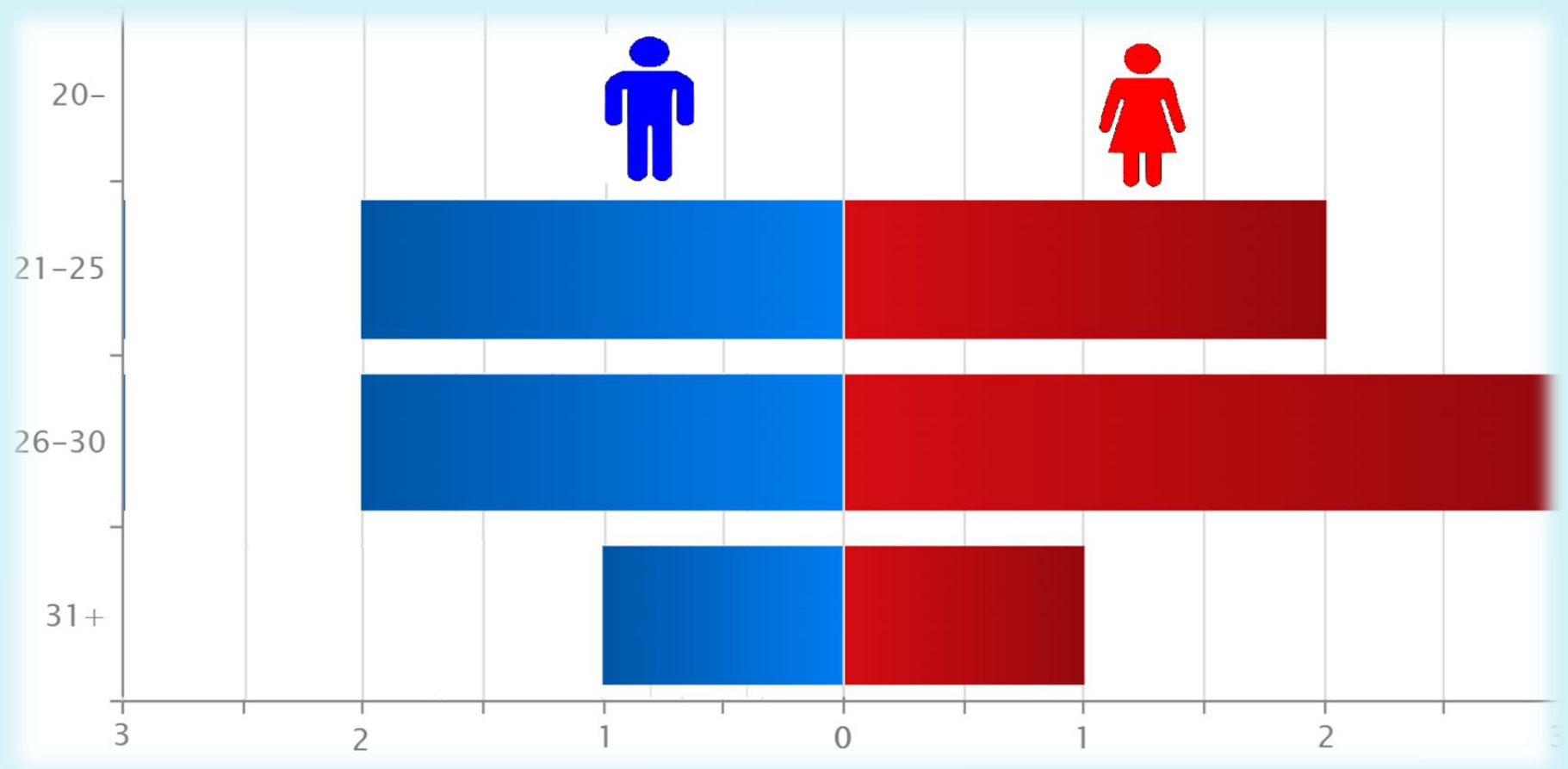
Dataset Results in a 2D Visualization



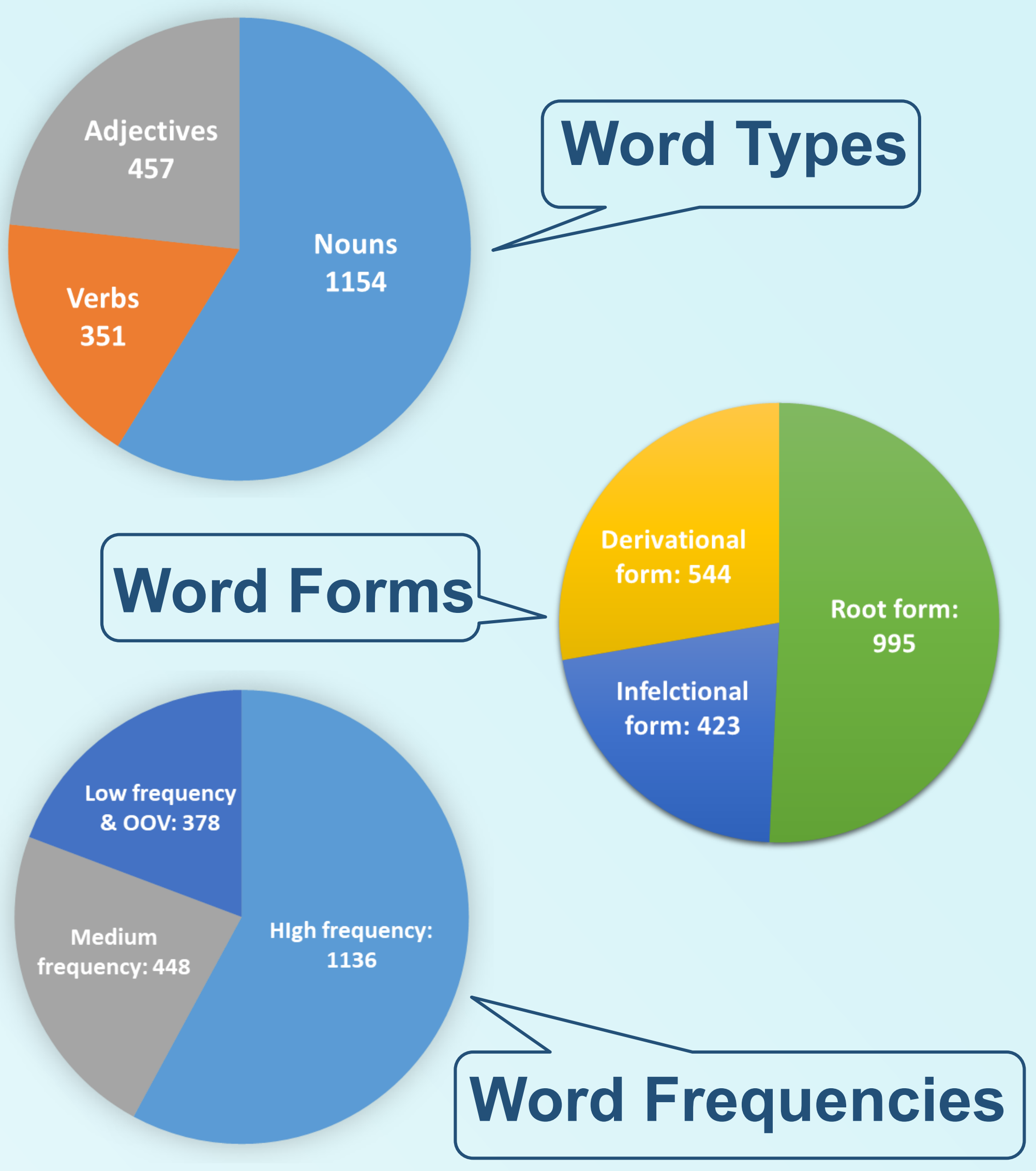
- *Each dot represents a word-pair in the dataset;*
  - *X-axis: Relatedness scores, Y-axis: Similarity scores;*
- *SU – Similar-Unrelated, SR – Similar-Related,*
- *DU – Dissimilar-Unrelated, DR – Dissimilar-Related.*

Annotation

- **Total number of annotators: 11**
  - High inter-annotator agreement score



Results



Conclusion

In this project, we presented the SimRelUz, a novel semantic evaluation dataset for the Uzbek language that consists of similarity and relatedness scores for word-pairs.

We have also presented an open-source web-based annotation tool designed for multiple-user semantic scores annotation.

Data

The Dataset and the code for the Annotation tool are available:

GOBIERNO DE ESPAÑA
MINISTERIO DE ECONOMÍA Y COMPETITIVIDAD
XUNTA DE GALICIA
LREC 2022
Marseille
13th edition
Le Palais du Pharo
June 20-25