DiaWUG: A Dataset for Diatopic Lexical Semantic Variation in Spanish
Giola Baldissin, Dominik Schlechtweg, Sabine Schulte im Walde
Institute for Natural Language Processing (IMS), University of Stuttgart
13th Language Resources and Evaluation Conference (Marseille, June 20 – 25, 2022)

Contributions

- Dataset for diatopic lexical semantic variation in Spanish
- Using DURel framework [Schlechtweg et al. 2018]
- Human annotated judgements of usage pairs
- Representation in Word Usage Graphs
- Shared WUGs

Guagua: Semasiological Perspective

1. Entre la ubicación del lugar (sin combinaciones de guaguas para llegar), el intenso verano, [...] se logró un sentido peculiar del espacio [...] ‘Among the location of the place (without bus combination to arrive there), the heavy summer, [...] a peculiar sense of space was achieved [...]’ [Cuba]
2. Tras las ventanas del tercer piso se divisas unas guaguas en sus cunas [...] ‘Behind the windows of the third floor babies in their cribs can be seen [...]’ [Argentina]

Corpora & Varieties

<table>
<thead>
<tr>
<th>Variety</th>
<th>Types</th>
<th>Tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td>Argentina (AR)</td>
<td>NN</td>
<td>381,370</td>
</tr>
<tr>
<td>Colombia (CO)</td>
<td>NN</td>
<td>346,285</td>
</tr>
<tr>
<td>Cuba (CU)</td>
<td>NN</td>
<td>243,549</td>
</tr>
<tr>
<td>Peru (PE)</td>
<td>NN</td>
<td>296,180</td>
</tr>
<tr>
<td>Spain (ES)</td>
<td>NN</td>
<td>761,875</td>
</tr>
<tr>
<td>Venezuela (VE)</td>
<td>NN</td>
<td>295,403</td>
</tr>
</tbody>
</table>

Guagua/Colectivo: Onomasiological Perspective

3. [...] los que transitamos a pie por calles y calzadas sufriendo el humo negro de camiones, guaguas y almendrones [...] ‘...those who walk through streets and roads suffering the black smoke of trucks, busses and “almendrones” [...]’ [Cuba]
4. Cuando terminaron de comer, los acompañó hasta la parada del colectivo. ‘When they finished eating, she walked them to the bus stop.’ [Argentina]

Usages per Target Word and Variety

<table>
<thead>
<tr>
<th>Target word</th>
</tr>
</thead>
</table>
| Argentina (AR)| 30
| Colombia (CO)| 30
| Peru (PE)     | 30
| Spain (ES)    | 45

Motivation: Lexical Semantic Variation in Spanish

- Task: Judge semantic relatedness of a pair of usage pairs
- 17 native speakers
- 8589 judgements
- IAA: (i) Krippendorff’s α = 0.64
- (ii) WAVG Pairwise Spearman Correlation ρ = 0.60

Word Usage Graphs (WUGs)

- Novel dataset for diatopic variation in Spanish: https://zenodo.org/record/5544553
- Semasiological and onomasiological variation
- 8589 judgements, 35 target words, 23 word combinations

Conclusions

- Reliable IAA comparable to previous studies [Erk et al. 2013, Rodina & Kultuzov 2020, Schlechtweg et al. 2018]
- Evaluation of computational modeling e.g., WIC [Arrendiz et al. 2020, Martelli et al. 2021]

References


Diawug: A Dataset for Diatopic Lexical Semantic Variation in Spanish
Motivation: Lexical Semantic Variation in Spanish

- Task: Judge semantic relatedness of a pair of usage pairs
- 17 native speakers
- 8589 judgements
- IAA: (i) Krippendorff’s α = 0.64
- (ii) WAVG Pairwise Spearman Correlation ρ = 0.60

Word Usage Graphs (WUGs)

- Novel dataset for diatopic variation in Spanish: https://zenodo.org/record/5544553
- Semasiological and onomasiological variation
- 8589 judgements, 35 target words, 23 word combinations

Conclusions

- Reliable IAA comparable to previous studies [Erk et al. 2013, Rodina & Kultuzov 2020, Schlechtweg et al. 2018]
- Evaluation of computational modeling e.g., WIC [Arrendiz et al. 2020, Martelli et al. 2021]

References


