DiaWUG: A Dataset for Diatopic Lexical Semantic Variation in Spanish

13th Language Resources and Evaluation Conference
Marseille, June 20 – 25, 2022

Gioia Baldissin, Dominik Schlechtweg, Sabine Schulte im Walde
Institute for Natural Language Processing (IMS), University of Stuttgart, Germany
Contributions

• Dataset for **diatopic lexical semantic variation in Spanish**
  • Using DURel framework (Schlechtweg, Schulte im Walde, & Eckmann, 2018)
    ▶ human annotated judgements of usage pairs
    ▶ representation in Word Usage Graphs

• **semasiological perspective**
  e.g., *boot*: “type of shoe”; “storage space at the back of the car” [UK]

• **onomasiological perspective**
  e.g., “storage space at the back of the car”: *boot* [UK]; *trunk* [US]

• Gold Standard for sense-related NLP tasks
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 divisan unas guaguas en sus cunas [...]
‘Behind the windows of the third floor babies in their cribs can be seen [...]’ [Argentina]

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]
## Corpora & Varieties

| Variety         | Types  | Tokens          |
|-----------------|--------|-----------------|
| Argentina (AR)  | 381,370| 97,117,561      |
| Colombia (CO)   | 346,285| 91,141,040      |
| Cuba (CU)       | 243,549| 32,938,685      |
| Peru (PE)       | 296,180| 60,324,754      |
| Spain (ES)      | 761,875| 240,488,211     |
| Venezuela (VE)  | 259,403| 52,277,543      |

1. *Corpus del Español: Web/Dialects* (Davies, 2016)
### Usages per Target Word and Variety

| Target words                      | |U| |
|----------------------------------|--|---|
| amarrar\textsubscript{VB} (ES, VE), atar\textsubscript{VB} (ES) | 45 |
| argolla\textsubscript{NN} (ES, PE) | 30 |
| banco\textsubscript{NN} (AR, PE) | 30 |
| baúl\textsubscript{NN} (AR, ES), maletero\textsubscript{NN} (ES) | 45 |
| bolo\textsubscript{NN} (AR, CU) | 30 |
| botar\textsubscript{VB} (ES, VE) | 30 |
| cartera\textsubscript{NN} (CU, ES), bolso\textsubscript{NN} (ES) | 45 |
| chamaco\textsubscript{NN} (CU), pibe\textsubscript{NN} (AR), chico\textsubscript{NN} (ES) | 45 |
| churro\textsubscript{NN} (CO, ES) | 30 |
| coche\textsubscript{NN} (ES), carro\textsubscript{NN} (CU) | 30 |
| flete\textsubscript{NN} (CO, ES) | 30 |
| franela\textsubscript{NN} (CO, ES) | 30 |

| Target words                      | |U| |
|----------------------------------|--|---|
| gato\textsubscript{NN} (AR, ES) | 30 |
| guagua\textsubscript{NN} (AR, CU, PE), colectivo\textsubscript{NN} (AR, ES) | 74 |
| plomero\textsubscript{NN} (ES, VE), fontanero\textsubscript{NN} (ES) | 42 |
| pollera\textsubscript{NN} (AR), falda\textsubscript{NN} (ES) | 30 |
| saco\textsubscript{NN} (ES, PE) | 30 |
| sindicar\textsubscript{VB} (CO, ES), acusar\textsubscript{VB} (ES) | 45 |
| tinto\textsubscript{NN} (CO, ES) | 30 |
| vaina\textsubscript{NN} (ES, VE) | 30 |
| vereda\textsubscript{NN} (ES, PE) | 30 |
| vidriera\textsubscript{NN} (CU, ES), escaparate\textsubscript{NN} (ES, VE) | 60 |
| volante\textsubscript{NN} (ES), timón\textsubscript{NN} (CU, ES) | 45 |

- POS: \textit{VB}: verb, \textit{NN}: noun
- |\textit{U}|: number of usages
- Total |\textit{U}|: 866
- \textit{guagua}: “baby” (AR), “bus” (CU), “bread with child shape” (PE)
- \textit{colectivo}: “bus” (AR), “group, union, corporation” (ES)
Usages per Target Word and Variety

| Target words | U |
|--------------|---|
| **amarrar**<sub>VB</sub> (ES, VE), **atar**<sub>VB</sub> (ES) | 45 |
| **argolla**<sub>NN</sub> (ES, PE) | 30 |
| **bano**<sub>NN</sub> (AR, PE) | 30 |
| **baul**<sub>NN</sub> (AR, ES), **maletero**<sub>NN</sub> (ES) | 45 |
| **bolo**<sub>NN</sub> (AR, CU) | 30 |
| **botar**<sub>VB</sub> (ES, VE) | 30 |
| **cartera**<sub>NN</sub> (CU, ES), **bolso**<sub>NN</sub> (ES) | 45 |
| **chamaco**<sub>NN</sub> (CU), **pibe**<sub>NN</sub> (AR), **chico**<sub>NN</sub> (ES) | 45 |
| **churrco**<sub>NN</sub> (CO, ES) | 30 |
| **coche**<sub>NN</sub> (ES), **carro**<sub>NN</sub> (CU) | 30 |
| **flete**<sub>NN</sub> (CO, ES) | 30 |
| **franela**<sub>NN</sub> (CO, ES) | 30 |

| Target words | U |
|--------------|---|
| **gato**<sub>NN</sub> (AR, ES) | 30 |
| **guagua**<sub>NN</sub> (AR, CU, PE), **colectivo**<sub>NN</sub> (AR, ES) | 74 |
| **plomero**<sub>NN</sub> (ES, VE), **fontanero**<sub>NN</sub> (ES) | 42 |
| **pollera**<sub>NN</sub> (AR), **falda**<sub>NN</sub> (ES) | 30 |
| **saco**<sub>NN</sub> (ES, PE) | 30 |
| **sindicar**<sub>VB</sub> (CO, ES), **acusar**<sub>VB</sub> (ES) | 45 |
| **tinto**<sub>NN</sub> (CO, ES) | 30 |
| **vaina**<sub>NN</sub> (ES, VE) | 30 |
| **vereda**<sub>NN</sub> (ES, PE) | 30 |
| **vidriera**<sub>NN</sub> (CU, ES), **escaparate**<sub>NN</sub> (ES, VE) | 60 |
| **volante**<sub>NN</sub> (ES), **timon**<sub>NN</sub> (CU, ES) | 45 |

- **POS**: **VB**: verb, **NN**: noun
- **|U|**: number of usages
- **Total |U|**: 866
- **pollera (AR), falda (ES)**: “skirt”
### Usages per Target Word and Variety

| Target words                      | \(|U|\) |
|-----------------------------------|---------|
| `amarrar_{VB}` (ES, VE), `atar_{VB}` (ES) | 45      |
| `argolla_{NN}` (ES, PE)           | 30      |
| `banco_{NN}` (AR, PE)             | 30      |
| `baúl_{NN}` (AR, ES), `maletero_{NN}` (ES) | 45  |
| `bolo_{NN}` (AR, CU)              | 30      |
| `botar_{VB}` (ES, VE)             | 30      |
| `cartera_{NN}` (CU, ES), `bolso_{NN}` (ES) | 45  |
| `chamaco_{NN}` (CU), `pibe_{NN}` (AR), `chico_{NN}` (ES) | 45  |
| `churro_{NN}` (CO, ES)            | 30      |
| `coche_{NN}` (ES), `carro_{NN}` (CU) | 30  |
| `flete_{NN}` (CO, ES)             | 30      |
| `franela_{NN}` (CO, ES)           | 30      |

| Target words                      | \(|U|\) |
|-----------------------------------|---------|
| `gato_{NN}` (AR, ES)              | 30      |
| `guagua_{NN}` (AR, CU, PE), `colectivo_{NN}` (AR, ES) | 74  |
| `plomero_{NN}` (ES, VE), `fontanero_{NN}` (ES) | 42  |
| `pollera_{NN}` (AR), `falda_{NN}` (ES) | 30  |
| `saco_{NN}` (ES, PE)              | 30      |
| `sindicar_{VB}` (CO, ES), `acusar_{VB}` (ES) | 45  |
| `tinto_{NN} (CO, ES)`             | 30      |
| `vaina_{NN}` (ES, VE)             | 30      |
| `vereda_{NN}` (ES, PE)            | 30      |
| `vidriera_{NN}` (CU, ES), `escaparate_{NN}` (ES, VE) | 60  |
| `volante_{NN}` (ES), `timón_{NN}` (CU, ES) | 45  |

- **POS**: `VB`: verb, `NN`: noun
- **|U|**: number of usages
- **Total |U|**: 866
- **tinto**: “(black) coffee” (CO), “(red) wine” (ES)
Annotation scale (Schlechtweg et al., 2018)

4: Identical
3: Closely Related
2: Distantly Related
1: Unrelated
0: Cannot decide
Judging the Semantic Relatedness of Usage Pairs

1: Unrelated

(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 divisan unas guaguas en sus cunas [...] ‘Behind the windows of the third floor babies in their cribs can be seen [...]’ [Argentina]

4: Identical

(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]
Annotation

- Task: Judge semantic relatedness of a pair of usages
- 17 native speakers
- 8589 judgements
- Inter-annotator agreement:
  - Krippendorff’s $\alpha = 0.64$
  - Weighted Average Pairwise Spearman Correlation $\rho = 0.60$

4: Identical
3: Closely Related
2: Distantly Related
1: Unrelated
0: Cannot decide
Annotation → Word Usage Graph

Usage 1

[...] los que transitamos a pie por calles y calzadas sufriendo el humo negro de camiones, guaguas y almendrones [...] 

Usage 2

Cuando terminaron de comer, los acompañó hasta la parada del colectivo.

When they finished eating, she walked them to the bus stop.
Word Usage Graphs (WUGs)

- guagua/colectivo
- pollera/falda
- tinto
WUGs: guagua/colectivo (full graph)

(a) Complete WUG

(b) Cuba

(c) Argentina

Labels: "g": guagua; "c": colectivo
WUGs: guagua/colectivo (Cuba’s subgraph)

Labels: "g": guagua; "c": colectivo
WUGs: guagua/colectivo (Argentina’s subgraph)

(a) Complete WUG

(b) Cuba

(c) Argentina

Labels: "g": guagua; "c": colectivo
Conclusions

• Novel dataset for diatopic variation in Spanish
• Semasiological and **onomasiological** variation
• **8589** judgements, **35** target words, **23** word combinations
• Reliable
  IAA comparable to previous related studies such as Erk, McCarthy, and Gaylord (2013); Rodina and Kutuzov (2020); Schlechtweg et al. (2018)

• Evaluation of computational modeling
e.g., WiC (Armendariz et al., 2020; Martelli, Kalach, Tola, & Navigli, 2021)

• Further steps to improve its representativeness:
  • + varieties/usages/annotators/annotations
Armendariz, C. S., Purver, M., Pollak, S., Ljubešić, N., Ulčar, M., Vulić, I., & Pilehvar, M. T. (2020). SemEval-2020 Task 3: Graded Word Similarity in Context. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation* (pp. 36–49). Barcelona (online): International Committee for Computational Linguistics.

Davies, M. (2016). *Corpus del español: Two billion words, 21 countries (Web/Dialects)*. Brigham Young University.

Erk, K., McCarthy, D., & Gaylord, N. (2013). Measuring Word Meaning in Context. *Computational Linguistics, 39*(3), 511–554.

Martelli, F., Kalach, N., Tola, G., & Navigli, R. (2021). SemEval-2021 Task 2: Multilingual and Cross-lingual Word-in-Context Disambiguation (MCL-WiC). In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)* (pp. 24–36). Online: Association for Computational Linguistics.

Rodina, J., & Kutuzov, A. (2020). RuSemShift: A Dataset of Historical Lexical Semantic Change in Russian. In *Proceedings of the 28th International Conference on Computational Linguistics (COLING 2020)*. Association for Computational Linguistics.
Schlechtweg, D., Schulte im Walde, S., & Eckmann, S. (2018). Diachronic Usage Relatedness (DURel): A Framework for the Annotation of Lexical Semantic Change. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers) (pp. 169–174). New Orleans, Louisiana: Association for Computational Linguistics.