



**University of Stuttgart**  
Germany



# 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

# Motivation: Lexical Semantic Variation in Spanish

## 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<sup>1</sup> & Varieties

<b>Variety</b>	<b>Types</b>	<b>Tokens</b>
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

---

<sup>1</sup>*Corpus del Español: Web/Dialects* (Davies, 2016)

# Usages per Target Word and Variety

Target words	U	Target words	U
<i>amarrar</i> <sub>VB</sub> (ES, VE), <i>atar</i> <sub>VB</sub> (ES)	45	<i>gato</i> <sub>NN</sub> (AR, ES)	30
<i>argolla</i> <sub>NN</sub> (ES, PE)	30	<b><i>guagua</i><sub>NN</sub> (AR, CU, PE), <i>colectivo</i><sub>NN</sub> (AR, ES)</b>	<b>74</b>
<i>banco</i> <sub>NN</sub> (AR, PE)	30	<i>plomero</i> <sub>NN</sub> (ES, VE), <i>fontanero</i> <sub>NN</sub> (ES)	42
<i>baúl</i> <sub>NN</sub> (AR, ES), <i>maletero</i> <sub>NN</sub> (ES)	45	<i>pollera</i> <sub>NN</sub> (AR), <i>falda</i> <sub>NN</sub> (ES)	30
<i>bolo</i> <sub>NN</sub> (AR, CU)	30	<i>saco</i> <sub>NN</sub> (ES, PE)	30
<i>botar</i> <sub>VB</sub> (ES, VE)	30	<i>sindicar</i> <sub>VB</sub> (CO, ES), <i>acusar</i> <sub>VB</sub> (ES)	45
<i>cartera</i> <sub>NN</sub> (CU, ES), <i>bolso</i> <sub>NN</sub> (ES)	45	<i>tinto</i> <sub>NN</sub> (CO, ES)	30
<i>chamaco</i> <sub>NN</sub> (CU), <i>pibe</i> <sub>NN</sub> (AR), <i>chico</i> <sub>NN</sub> (ES)	45	<i>vaina</i> <sub>NN</sub> (ES, VE)	30
<i>churro</i> <sub>NN</sub> (CO, ES)	30	<i>vereda</i> <sub>NN</sub> (ES, PE)	30
<i>coche</i> <sub>NN</sub> (ES), <i>carro</i> <sub>NN</sub> (CU)	30	<i>vidriera</i> <sub>NN</sub> (CU, ES), <i>escaparate</i> <sub>NN</sub> (ES, VE)	60
<i>flete</i> <sub>NN</sub> (CO, ES)	30	<i>volante</i> <sub>NN</sub> (ES), <i>timón</i> <sub>NN</sub> (CU, ES)	45
<i>franela</i> <sub>NN</sub> (CO, ES)	30		

- POS: *VB*: verb, *NN*: noun
- |U|: number of usages
- Total |U|: 866
- *guagua*: "baby" (AR), "bus" (CU), "bread with child shape" (PE)
- *colectivo*: "bus" (AR), "group, union, corporation" (ES)

# Usages per Target Word and Variety

Target words	U	Target words	U
<i>amarrar</i> <sub>VB</sub> (ES, VE), <i>atar</i> <sub>VB</sub> (ES)	45	<i>gato</i> <sub>NN</sub> (AR, ES)	30
<i>argolla</i> <sub>NN</sub> (ES, PE)	30	<i>guagua</i> <sub>NN</sub> (AR, CU, PE), <i>colectivo</i> <sub>NN</sub> (AR, ES)	74
<i>banco</i> <sub>NN</sub> (AR, PE)	30	<i>plomero</i> <sub>NN</sub> (ES, VE), <i>fontanero</i> <sub>NN</sub> (ES)	42
<i>baúl</i> <sub>NN</sub> (AR, ES), <i>maletero</i> <sub>NN</sub> (ES)	45	<b><i>pollera</i><sub>NN</sub> (AR), <i>falda</i><sub>NN</sub> (ES)</b>	<b>30</b>
<i>bolo</i> <sub>NN</sub> (AR, CU)	30	<i>saco</i> <sub>NN</sub> (ES, PE)	30
<i>botar</i> <sub>VB</sub> (ES, VE)	30	<i>sindicar</i> <sub>VB</sub> (CO, ES), <i>acusar</i> <sub>VB</sub> (ES)	45
<i>cartera</i> <sub>NN</sub> (CU, ES), <i>bolso</i> <sub>NN</sub> (ES)	45	<i>tinto</i> <sub>NN</sub> (CO, ES)	30
<i>chamaco</i> <sub>NN</sub> (CU), <i>pibe</i> <sub>NN</sub> (AR), <i>chico</i> <sub>NN</sub> (ES)	45	<i>vaina</i> <sub>NN</sub> (ES, VE)	30
<i>churro</i> <sub>NN</sub> (CO, ES)	30	<i>vereda</i> <sub>NN</sub> (ES, PE)	30
<i>coche</i> <sub>NN</sub> (ES), <i>carro</i> <sub>NN</sub> (CU)	30	<i>vidriera</i> <sub>NN</sub> (CU, ES), <i>escaparate</i> <sub>NN</sub> (ES, VE)	60
<i>flete</i> <sub>NN</sub> (CO, ES)	30	<i>volante</i> <sub>NN</sub> (ES), <i>timón</i> <sub>NN</sub> (CU, ES)	45
<i>franela</i> <sub>NN</sub> (CO, ES)	30		


- 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	Target words	U
<i>amarrar</i> <sub>VB</sub> (ES, VE), <i>atar</i> <sub>VB</sub> (ES)	45	<i>gato</i> <sub>NN</sub> (AR, ES)	30
<i>argolla</i> <sub>NN</sub> (ES, PE)	30	<i>guagua</i> <sub>NN</sub> (AR, CU, PE), <i>colectivo</i> <sub>NN</sub> (AR, ES)	74
<i>banco</i> <sub>NN</sub> (AR, PE)	30	<i>plomero</i> <sub>NN</sub> (ES, VE), <i>fontanero</i> <sub>NN</sub> (ES)	42
<i>baúl</i> <sub>NN</sub> (AR, ES), <i>maletero</i> <sub>NN</sub> (ES)	45	<i>pollera</i> <sub>NN</sub> (AR), <i>falda</i> <sub>NN</sub> (ES)	30
<i>bolo</i> <sub>NN</sub> (AR, CU)	30	<i>saco</i> <sub>NN</sub> (ES, PE)	30
<i>botar</i> <sub>VB</sub> (ES, VE)	30	<i>sindicar</i> <sub>VB</sub> (CO, ES), <i>acusar</i> <sub>VB</sub> (ES)	45
<i>cartera</i> <sub>NN</sub> (CU, ES), <i>bolso</i> <sub>NN</sub> (ES)	45	<b><i>tinto</i><sub>NN</sub> (CO, ES)</b>	<b>30</b>
<i>chamaco</i> <sub>NN</sub> (CU), <i>pibe</i> <sub>NN</sub> (AR), <i>chico</i> <sub>NN</sub> (ES)	45	<i>vaina</i> <sub>NN</sub> (ES, VE)	30
<i>churro</i> <sub>NN</sub> (CO, ES)	30	<i>vereda</i> <sub>NN</sub> (ES, PE)	30
<i>coche</i> <sub>NN</sub> (ES), <i>carro</i> <sub>NN</sub> (CU)	30	<i>vidriera</i> <sub>NN</sub> (CU, ES), <i>escaparate</i> <sub>NN</sub> (ES, VE)	60
<i>flete</i> <sub>NN</sub> (CO, ES)	30	<i>volante</i> <sub>NN</sub> (ES), <i>timón</i> <sub>NN</sub> (CU, ES)	45
<i>franela</i> <sub>NN</sub> (CO, ES)	30		

- 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]

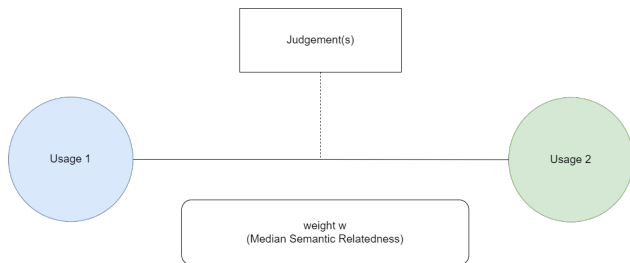
# Annotation

- 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 transilamos a pie por calles y calzadas sufriendo el humo negro de camiones, **guaguas** y alimendrones [...]

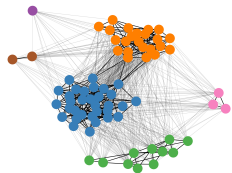
[...] those who walk through streets and roads suffering the black smoke of trucks, **busses** and "alimendrones" [...]

Usage 2

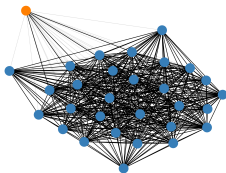
Cuando terminaron de comer, los acompañó hasta la parada del **colectivo**.

'When they finished eating, she walked them to the **bus stop**.'

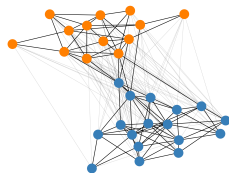
# Word Usage Graphs (WUGs)



*guagua/colectivo*

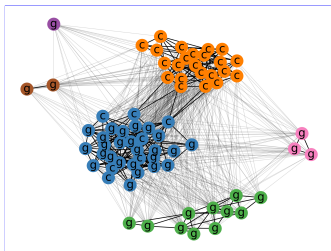


*pollera/falda*



*tinto*

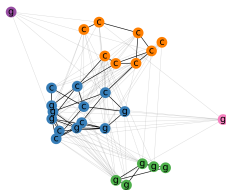
# WUGs: *guagua/colectivo* (full graph)



(a) Complete WUG



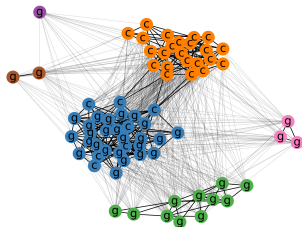
(b) Cuba



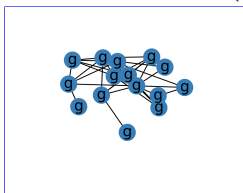
(c) Argentina

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

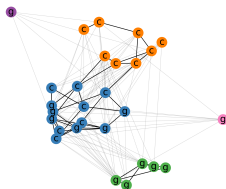
# WUGs: *guagua/colectivo* (Cuba's subgraph)



(a) Complete WUG



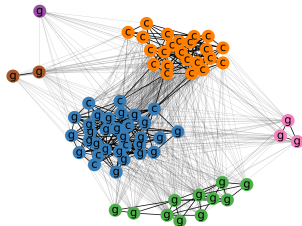
(b) Cuba



(c) Argentina

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

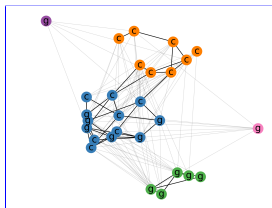
# WUGs: *guagua/colectivo* (Argentina's subgraph)



(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



# Bibliography I

- 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.

## Bibliography II

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.