University of Stuttgart
Germany

DiaWUG: A Dataset for Diatopic Lexical Semantic Variation in Spanish

Gioia 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
- **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 & 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

Corpora and varieties used in the study [Davies 2016].

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

Guagua: “baby” (AR, PE), “bus” (CU), “bread with child shape” (PE); ***colectivo***: “bus” (AR), “group, union, corporation” (ES). ***Pollera*** (AR), ***falda*** (ES): “skirt”. ***Tinto***: “(black) coffee” (CO), “(red) wine” (ES).

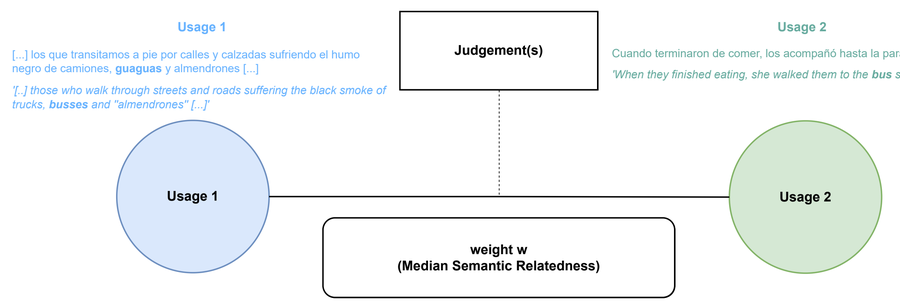
Annotation

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

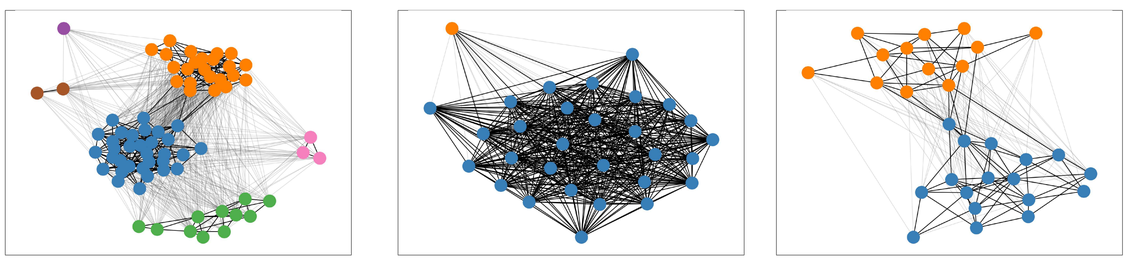
↑
4: Identical
3: Closely Related
2: Distantly Related
1: Unrelated

DURel relatedness scale.

Annotation → Word Usage Graph

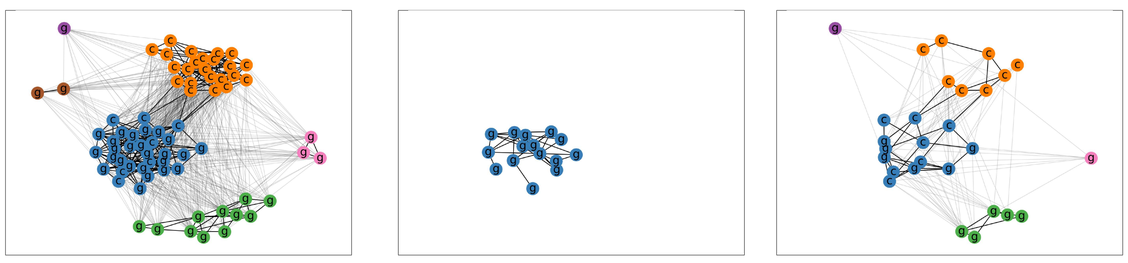


Word Usage Graphs (WUGs)



guagua/colectivo | *pollera/falda* | *tinto*

WUGs: *guagua/colectivo*



Complete WUG | [CU]: ”g”: *guagua*, ”c”: *colectivo* | [AR]: ”g”: *guagua*, ”c”: *colectivo*

Conclusions

- Novel dataset for diatopic variation in Spanish:
  `https://zenodo.org/record/5544553`
- Semasiological and **onomasiological** variation
- **8589** judgements, **35** target words, **23** word combinations
- Reliable
  IAA comparable to previous studies [Erk et al. 2013, Rodina & Kutuzov 2020, Schlechtweg et al. 2018]
- Evaluation of computational modeling
  e.g., WiC [Armendariz et al. 2020, Martelli et al. 2021]

References

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 *Proc. 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. Retrieved from `http://www.corpusdelespanol.org/web-dial/`

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 *Proc. 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 *Proc. of the 28th International CONF 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 *Proc. of the 2018 CONF 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.