Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 2601–2609
Marseille, 20-25 June 2022

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
Pfaffenwaldring 5b, 70569 Stuttgart, Germany
{baldisga,schlecdk,schulte}@ims.uni-stuttgart.de

Abstract

We provide a novel dataset – *DiaWUG* – with judgements on diatopic lexical semantic variation for six Spanish variants in Europe and Latin America. In contrast to most previous meaning-based resources and studies on semantic diatopic variation, we collect annotations on semantic relatedness for Spanish target words in their contexts from both a *semasiological perspective* (i.e., exploring the meanings of a word given its form, thus including polysemy) and an *onomasiological perspective* (i.e., exploring identical meanings of words with different forms, thus including synonymy). In addition, our novel dataset exploits and extends the existing framework DURel for annotating word senses in context (Erk et al., 2013; Schlechtweg et al., 2018) and the framework-embedded Word Usage Graphs (WUGs) – which up to now have mainly be used for semasiological tasks and resources – in order to distinguish, visualize and interpret lexical semantic variation of contextualized words in Spanish from these two perspectives, i.e., semasiological and onomasiological language variation.

Keywords: diatopic variation, semasiology, onomasiology, semantic relatedness, word usage graphs

1. Introduction

Theoretical and empirical investigations of language variation in dialectology have a long history and draw on a rich variety of linguistic features on the phonetic, phonological, morphological, semantic and syntactic levels (Boberg et al., 2018). Computational approaches to dialectology, however, have been more limited and rarely address language variation regarding lexical meaning, as in Franco et al. (2019). In contrast, lexical semantics has been a major focus of computational research in other types of language variation, such as diachronic (Kutuzov et al., 2018; Bizzoni et al., 2020; Schlechtweg et al., 2020; Zamora-Reina et al., 2022), diastratic (Frassinelli et al., 2021) and domain-specific sense variation (Pérez, 2016; Hätty et al., 2019). As to our knowledge, only one approach up to date has worked on the intersection of the above, i.e., sense variations in diatopic language varieties (Franco et al., 2019). Diatopic variation is not only relevant for mere sociolinguistic, lexicographic and lexicological purposes, but may be extended to the more general case of word sense disambiguation studies. In the current paper we provide a novel dataset for diatopic sense variation in Spanish, making use of an actual sample of the language to explore synchronic variation, rather than relying on static manual resources such as dialect dictionaries (Franco et al., 2019). For this, we activate both sense-related perspectives, the semasiological perspective (i.e., exploring the meanings of a word given its form, thus including polysemy) and the onomasiological perspective (i.e., exploring related meanings of words with different forms, thus including synonymy). This resource is intended to function as a gold standard for computational models used for sense-related NLP tasks.

Spanish is one of the most widely spoken languages in the world, and differences in its lexicon manifest themselves between and amongst European and American Spanish variants (Lipski, 1994; Sánchez Lobato, 1994; Haensch, 2002; Frago García and Franco Figueroa, 2003; Enguita Utrilla, 2010). Important aspects to contemplate when researching on synchronic varieties of American Spanish are, in particular, the origin of the settlers and the viceroyalty distribution, the substratum of the native inhabitants and, to some extent, the African component (Lipski, 1994; Lipski, 2014). For our research focus on semasiological and onomasiological sense variation, we rely on characterizations in standard monographs and dictionaries to identify interesting target words, and on the *Corpus del Español* (Davies, 2016) to extract those target words in their contexts. For example, *guagua* means "bus" in the Antilles, Equatorial Guinea and the Canary Islands, while it refers to "baby" in Argentina, Bolivia, Chile, Colombia, Ecuador and Peru, and in Peru it has an additional reading "sweet bread with a child shape" (semasiological perspective); in contrast, the concept "vehicle for public transportation" can be expressed also by *autobús* and *colectivo*, next to *guagua* (onomasiological perspective).

Our main contributions are two-fold: (1) We tackle a theoretically well-defined task (i.e., diatopic lexical semantic variation in Spanish) from an empirical perspective and provide a novel dataset as basis for computational modeling. (2) We extend and exploit the existing framework DURel (Erk et al., 2013; Schlechtweg et al., 2018) – which up to now has mainly been used for annotating and detecting semasiological meaning variation – to distinguish and annotate meaning variation also from an onomasiological perspective.

2. Data

2.1. Corpus

We extracted word uses (i.e., words in their context) for a selection of target words (see details below) from the *Corpus del Español: Web/Dialects*, a tagged and lem-

matized corpus containing about two billion words from web pages, categorized into blog and general (Davies, 2016). The corpus is divided into 21 sub-corpora in accordance with the 21 Spanish-speaking countries across Europe and America, including the United States. The identification of each variety was based on Google search and validated with Lipski (1994). We took six varieties into account for our dataset: Argentina, Colombia, Cuba, Peru, Spain, and Venezuela, see Table 1.

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

Table 1: Sizes of the subcorpora used in the study.

2.2. Target and Context Selection

For target selection we consulted standard literature on Spanish language variation (Lipski, 1994; Haensch, 2002; Enguita Utrilla, 2010) for a first choice, which was then validated with the two Spanish dictionaries *Diccionario de la Lengua Española* (RAE, 2020) and *Diccionario de Americanismos* (ASALE, 2010), where all entries specify the respective varieties.

Given these candidate target words and their meanings, the 1–3 most representative meanings across varieties as well as the varieties sharing these meaning(s) were chosen in order to automatically extract all instances (i.e., usages) from the respective subcorpora. Each contextual usage relies on a window size of $\pm$70 words and is split into target sentence (i.e., the sentence containing the target word), and preceding and following sentences. We then randomly sampled 15 usages per target word and language variety. In total, we retrieved 30–75 usages per target word, depending on the number of varieties and meanings (see Table 2). Some target words were disregarded if we did not find a sufficient number of usages across varieties, or if we observed a comparably high proportion of wrong part-of-speech tags.

3. Annotation and Representation

For the annotation and visualization, we relied on the openly available DURel interface.[1] Annotators were asked to judge the semantic relatedness of pairs of word usages, such as examples (1) and (2) for *guagua* in Table 3, on the relatedness scale in Table 4. This is a case of *semasiological variation*: (un)related meanings of the same word. We extended the DURel guidelines in order to cover also *onomasiological variation* (different words with related meanings), such as *guagua* and *colectivo* in examples (3) and (4).[2]

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
$gato_{NN}$ (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
Total	**866**

Table 2: Semasiological and onomasiological target word combinations: Each line provides target words either representing the same concept expressed by different lemmas (onomasiological perspective), exhibiting several senses (semasiological perspective), or a combination of both across varieties. Subscripts indicate the POS: *VB*: verb, *NN*: noun. The column $|U|$ refers to the number of usages.

The DURel tool presents usage pairs to annotators in a random order by randomly sampling pairs from the full set of usage pairs of the target word until a stopping criterion is reached (see below).

3.1. Annotation

17 native speakers from different Spanish-speaking countries participated voluntarily in the annotation process. The distribution of annotators per variety is as follows: 9 Spanish (8 from peninsular Spain, 1 from Canary Islands), 3 Cuban, 2 Colombian, 1 Peruvian, 1 Mexican and 1 Costa Rican. Most of them are 30–45 years old. Except for two annotators, all of them were undergraduate students or had obtained a high educational degree (almost $\frac{1}{3}$ in Spanish Philology, Linguistics and Humanities-related disciplines; two are Spanish

[1] https://www.ims.uni-stuttgart.de/data/durel-tool.

[2] The guidelines are available at https://zenodo.org/record/5544553.

(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 [...]*'
(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 [...]*'
(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" [...]*'
(4)	Cuando terminaron de comer, los acompañó hasta la parada del **colectivo**. '*When they finished eating, she walked them to the **bus** stop.*'

Table 3: A semasiological usage pair (above) and an onomasiological usage pair (below), illustrating the difference between polysemy and synonymy in lexical semantic language variation.

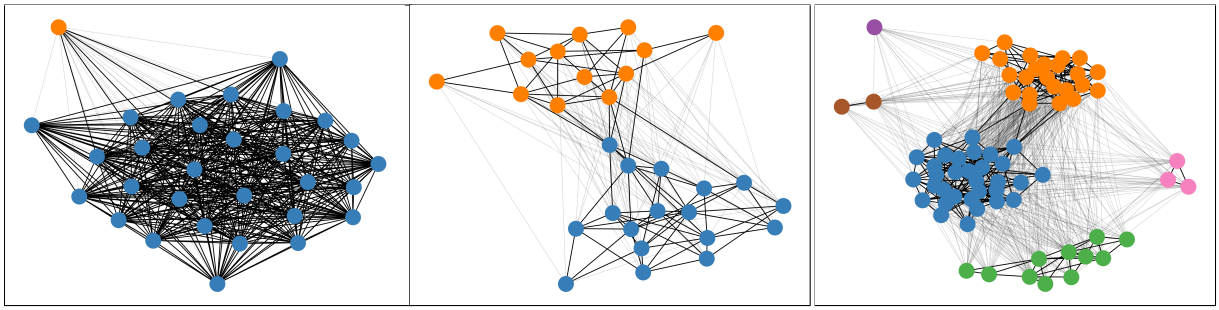


Figure 1: Word Usage Graphs of *pollera* (left), *tinto* (middle) and *guagua* (right). Nodes represent usages of the respective target word. Edge weights represent the median of relatedness judgements between usages (**black**/gray lines for **high**/low edge weights, i.e., weights ≥ 2.5/weights < 2.5).

teachers).

When assigning the target words to the annotators, we took their native variety, their knowledge of other Spanish varieties and the meanings of the target words in their specific varieties into account. For instance, *tinto* was exclusively annotated by Colombian native speakers, since it means "(glass of) red wine" across varieties, while in Colombian Spanish it also means "(cup of) black coffee" (Porto Dapena, 2018).

During the annotation process we collected a total of 8,632 judgements. The agreement at this point was approx. 0.5 for Krippendorff's α. However, after a manual inspection, we excluded judgements from some annotators for specific target words for the following reasons: one annotator applied the DURel scale in the inverse direction; one annotator provided a disproportionally high number of 0-judgements; one annotator gave identical scores for all annotation pairs; the same three annotators annotated target words in addition to the ones assigned to them. We further automatically removed usages from the data where more than half of the judgements involving the usage were 0 (cannot decide) on the DURel scale, see Table 4.

After removing the above-mentioned annotations from the dataset, the agreement increased to Krippendorff's $\alpha = 0.64$ and weighted average pairwise Spearman correlation $\rho = 0.60$, which is comparable to previous work (Erk et al., 2013; Schlechtweg et al., 2018; Rodina and Kutuzov, 2020). The final dataset comprises 8,589 judgements for 35 target words and 23 word combinations (see Table 2) and is publicly available together with the interactive WUGs (see next section).[3]

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

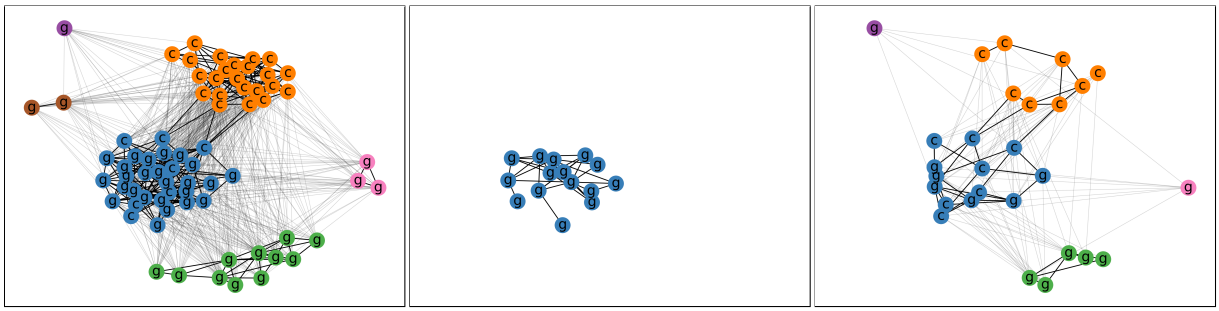
0: Cannot decide

Table 4: DURel scale.

3.2. Representation

We represented the annotated data of a word in a Word Usage Graph (WUG), where vertices represent word usages, edges represent the existence of at least a judgement between a pair of usages, and the weights represent the (median) semantic relatedness of the judgement(s). The WUGs are then clustered with a variation of correlation clustering (Schlechtweg et al., 2020; Schlechtweg et al., 2021; Schlechtweg, 2022), and clusters are interpreted as word senses, see Figure 1. Additionally, we split the graph for a target word into multiple subgraphs representing nodes from different varieties, see Figure 2 (middle and right). Similar to previous work (Schlechtweg et al., 2020; Kurtyigit et

[3]The dataset is publicly available at: `https://zenodo.org/record/5544553`. Find more data sets at: `https://www.ims.uni-stuttgart.de/data/wugs`.



Complete WUG | Cuba | Argentina

Figure 2: WUGs for *guagua/colectivo* (left), subgraphs for Cuba (middle) and Argentina (right). Zooming into nodes shows the labels of the respective target words, where "g" refers to *guagua* and "c" to *colectivo*.

al., 2021; Zamora-Reina et al., 2022), usages of a target word were annotated until all clusters with more than one usage (i.e., multi-clusters) were connected ensuring the robustness of the clustering algorithm.

4. Analysis

We demonstrate the interaction and representation of the semasiological and the onomasiological perspectives in our dataset as given in Figures 1 and 2.

The target word *pollera* (Argentina) accounts for the onomasiological perspective (see Figure 1, left). Its usages were compared with those of the more extended Spanish sense variant *falda* because, depending on the context, both denote the same meaning, namely "skirt" (as in examples (1)-(2) below): all the nodes of the graph but one belong to this sense cluster (blue). The outlier constitutes the orange cluster meaning "[hill] side, skirt", which is not specific of any Spanish variety (see example (3)).

(1) Además de esas 2 camisas de seda, me compré una **pollera** corta de cuero y un chaleco de gamuza, este último a 30.
'*[...] Besides these 2 silk blouses, I bought a short leather* ***skirt*** *and a suede vest, this one for 30.*'

(2) Frente a opciones más clásicas como los vestidos o las **faldas**, este verano hemos dado la bienvenida a una nueva prenda: la skort.
'*As opposed to more classic options such as dresses and* ***skirts****, this summer we have welcomed a new garment: the skort.*'

(3) Rayuela (capítulo 155) El nombre de esta calle ubicada en las **faldas** del Sagrado Corazón de Montmartre proviene de los abbesses, abadeses, presentes en la abadía de Montmartre fundada en el siglo XII.
'*Rayuela (chapter 155) The name of this street located on the* ***hill skirt*** *of [the Basilica of] the Sacred Heart of Montmartre comes from the abbesses, abbots, present in the Montmartre Abbey founded in the 12th century.*'

Tinto exemplifies the semasiological perspective (see Figure 1, middle). The annotated judgements output a graph with two well-defined clusters, i.e., "black coffee" (orange), almost exclusively composed of Colombian usages, and "red wine" (blue), predominantly containing usages from European Spanish. This corresponds to our expectations (cf. Section 3.1), since the "red wine" sense is spread across varieties whereas the "black coffee" sense is restricted to a small subset of varieties (excluding European Spanish). In the following we show two Colombian usages, where *tinto* denotes "black coffee" (4) and "red wine" (5):

(4) Los soldados quieren un autógrafo pero el Patrón no tiene tiempo, además está cansado, lleno de bostezos y necesita dos **tintos**.
'*The soldiers want an autograph but the Patrón does not have time, besides he is tired, yawning and needs two* ***coffees****.*'

(5) Aunque prácticamente puede utilizar se cualquier vino el más adecuado es un **tinto** joven afrutado.
'*Even though any wine can be practically used, the most appropriate one is a young, fruity* ***red wine****.*'

The pair *guagua/colectivo* illustrates the interaction of the semasiological and onomasiological perspectives across varieties (see Figure 1, right, Figure 2, and also Table 3). Figure 2 (left) depicts two main sense clusters: The orange nodes refer to "group, corporate, union" (only expressed by *colectivo*), while the blue ones refer to "bus" (expressed by both *guagua* (Cuba; middle) and *colectivo* (Argentina; right)). In Argentinian, *colectivo* expresses "group, corporate, union", while *guagua* acquires the meaning "baby" (green cluster). However, instances of the target word *guagua* meaning "bus" were unexpectedly sampled from the Argentinian subcorpus, as in (6). Such usage is proper of the Cuban variant, as inferable from the information provided by the context, and confirmed by Cuban native speakers:

(6) No solo artistas y escritores, el chofer de **guagua**, el mesero, el neurocirujano, nadie es ajeno o inmune a esa realidad. Que se puede esperar,

cresemos en un entorno donde desde el kínder se nos ínsita a que digamos al unisonó: Pioneros por el comunismo, seremos como el Che.
'*Not only artists and writers, the* ***bus*** *driver, the waiter, the neurosurgeon, nobody is foreign to that reality. What can be expected, we grow up in an environment in which since the kinder we are incited to say in unison: Pioneros por el comunismo, seremos como el Che [Pioneers for Communism, we will be like Che Guevara* [4] *].'*

Given the small size of the usages sample, this cannot be attributed with certainty to a noise effect. It might reflect a potential ongoing process of borrowing, a phenomenon intrinsic in language change and variation, and expected to be detected when observing a synchronic sample of the language use.[5]

5. Conclusion

We presented a new dataset for diatopic lexical semantic variation in six Spanish language varieties. Most importantly, we incorporated two sense-related perspectives (semasiological and onomasiological) into the annotation framework and illustrated the resulting meaning representations and distinctions between forms and concepts.

Although this work has a limited scope, we provide a resource as a starting point for further investigations, which may be extended by (1) covering other varieties, (2) sampling significantly more usages in order to be more representative, (3) relying on more annotators (per target variety), thus (4) considerably more annotations to near a full graph representation. This goal can only be achieved on a bigger scale research, if we consider that the full graph requires $\frac{n(n-1)}{2}$ annotations even for a relatively small set of n usages. Furthermore, this approach can also be exploited to encompass diatopic variation on a domain-specific dimension.

On the other hand, with a bigger amount of data it is possible to address a broad spectrum of questions on diatopic lexical-semantic divergence or correlation across Spanish varieties as consequence of linguistic or extralinguistic aspects like geographical proximity, historical factors (i.e., colonization stages, viceroyalty distribution, migratory waves on both directions), substrata, as well as on potential neologisms.

6. Acknowledgements

We thank all our annotators who contributed to the creation of this dataset. Dominik Schlechtweg has been funded by the project 'Towards Computational Lexical Semantic Change Detection' supported by the Swedish Research Council (2019–2022; contract 2018-01184) and by the research program 'Change is Key!' supported by Riksbankens Jubileumsfond (under reference number M21-0021).

[4] This is a slogan repeated by children and teenagers in Cuban schools.

[5] The same word *guagua* had probably been incorporated into the Canarian lexicon after the second decade of the past century due to bidirectional migratory waves between Cuba and the Canary Islands: `https://www.academiacanarialengua.org/consultas/58/`, `https://www.rae.es/tdhle/guagua`.

7. Bibliographical References

Bizzoni, Y., Degaetano-Ortlieb, S., Fankhauser, P., and Teich, E. (2020). Linguistic variation and change in 250 years of English scientific writing: A data-driven approach. *Frontiers in Artificial Intelligence*, 3(73).

Charles Boberg, et al., editors. (2018). *The Handbook of Dialectology*. John Wiley and Sons, Oxford.

Enguita Utrilla, J. M. (2010). Capítulo 6: Léxico y formación de palabras. In Milagros Aleza Izquierdo et al., editors, *La lengua española en América: Normas y usos actuales*, pages 185–216. Universitat de València, Valencia.

Frago García, J. A. and Franco Figueroa, M. (2003). Sobre la formación del español de América. In Juan Antonio Frago García et al., editors, *El español de América*, chapter 1, pages 11–36. Servicio de Publicaciones de la Universidad, Cádiz.

Franco, K., Geeraerts, D., Speelman, D., and Van Hout, R. (2019). Concept characteristics and variation in lexical diversity in two Dutch dialect areas. *Cognitive Linguistics*, 30(1):205–242.

Frassinelli, D., Lapesa, G., Alatrash, R., Schlechtweg, D., and Schulte im Walde, S. (2021). Regression analysis of lexical and morpho-syntactic properties of kiezdeutsch. In *Proceedings of the Eighth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 21–27, Kyiv, Ukraine. Association for Computational Linguistics.

Haensch, G. (2002). Español de América y Español de Europa. *Panace@*, Vol. III(7):37–64, 03.

Hätty, A., Schlechtweg, D., and Schulte im Walde, S. (2019). SURel: A gold standard for incorporating meaning shifts into term extraction. In *Proceedings of the 8th Joint Conference on Lexical and Computational Semantics*, pages 1–8, Minneapolis, MN, USA.

Kurtyigit, S., Park, M., Schlechtweg, D., Kuhn, J., and Schulte im Walde, S. (2021). Lexical semantic change discovery. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Online, aug. Association for Computational Linguistics.

Kutuzov, A., Ovrelid, L., Szymanski, T., and Velldal, E. (2018). Diachronic word embeddings and semantic shifts: A survey. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1384–1397, Santa Fe, New Mexico, USA.

Lipski, J. M. (1994). *Latin American Spanish*. Linguistics Library. Longman.

Lipski, J. M. (2014). The many facets of Spanish dialect diversification in Latin America. In S.S. Mufwene, editor, *Iberian Imperialism and Language Evolution in Latin America*, chapter 2. University of Chicago Press.

Pérez, M. J. M. (2016). Measuring the degree of specialisation of sub-technical legal terms through corpus comparison. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, 22(1):80–102.

Porto Dapena, J. Á. (2018). Sobre ambigüedad y vaguedad en los diccionarios. *Revista de Filología*, 36:329–365, 03.

Sánchez Lobato, J. (1994). El Español en América. In *ASELE. Actas IV*, pages 553–570. Centro Virtual Cervantes.

Schlechtweg, D., McGillivray, B., Hengchen, S., Dubossarsky, H., and Tahmasebi, N. (2020). SemEval-2020 Task 1: Unsupervised Lexical Semantic Change Detection. In *Proceedings of the 14th International Workshop on Semantic Evaluation*, Barcelona, Spain. Association for Computational Linguistics.

Schlechtweg, D., Tahmasebi, N., Hengchen, S., Dubossarsky, H., and McGillivray, B. (2021). DWUG: A large resource of diachronic word usage graphs in four languages. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7079–7091, Online and Punta Cana, Dominican Republic, nov. Association for Computational Linguistics.

Schlechtweg, D. (2022). *Human and Computational Measurement of Lexical Semantic Change*. Ph.D. thesis, University of Stuttgart, Stuttgart, Germany.

Zamora-Reina, F. D., Bravo-Marquez, F., and Schlechtweg, D. (2022). LSCDiscovery: A shared task on semantic change discovery and detection in Spanish. In *Proceedings of the 3rd International Workshop on Computational Approaches to Historical Language Change*, Dublin, Ireland. Association for Computational Linguistics.

8. Language Resource References

ASALE. (2010). *Diccionario de Americanismos*. Santillana.

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

Katrin Erk and Diana McCarthy and Nicholas Gaylord. (2013). *Measuring Word Meaning in Context*. Computational Linguistics, 39(3):511-554.

RAE. (2020). *Diccionario de la Lengua Española, 23.ª ed.* Real Academia Española.

Julia Rodina and Andrey Kutuzov. (2020). *RuSemShift: a dataset of historical lexical semantic change in Russian*. Proceedings of the 28th International Conference on Computational Linguistics (COLING 2020), Association for Computational Linguistics.

Schlechtweg, Dominik and Schulte im Walde, Sabine and Eckmann, Stefanie. (2018). *Diachronic Usage Relatedness (DURel): A Framework for the Annotation of Lexical Semantic Change*. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 169-174, New Orleans, Louisiana.

A. Appendix

Data set	LGS	t	n	N/V/A	|U|	AN	JUD	AV	SPR	KRI	UNC	LOSS	Sample
DiaWUG	ES	6	23	20/3/0	37	17	8k	1	.60	.64	0	.28	random

Table 5: Overview usage graphs. t = no. of varieties, n = no. of graphs, N/V/A = no. of nouns/verbs/adjectives, $|U|$ = avg. no. usages per word, AN = no. of annotators, JUD = total no. of judged usage pairs, AV = avg. no. of judgements per usage pair, SPR = weighted mean of pairwise Spearman, KRI = Krippendorff's alpha, UNC = avg. no. of uncompared multi-cluster combinations, LOSS = avg. of normalized clustering loss * 10, Sample = sampling strategy.

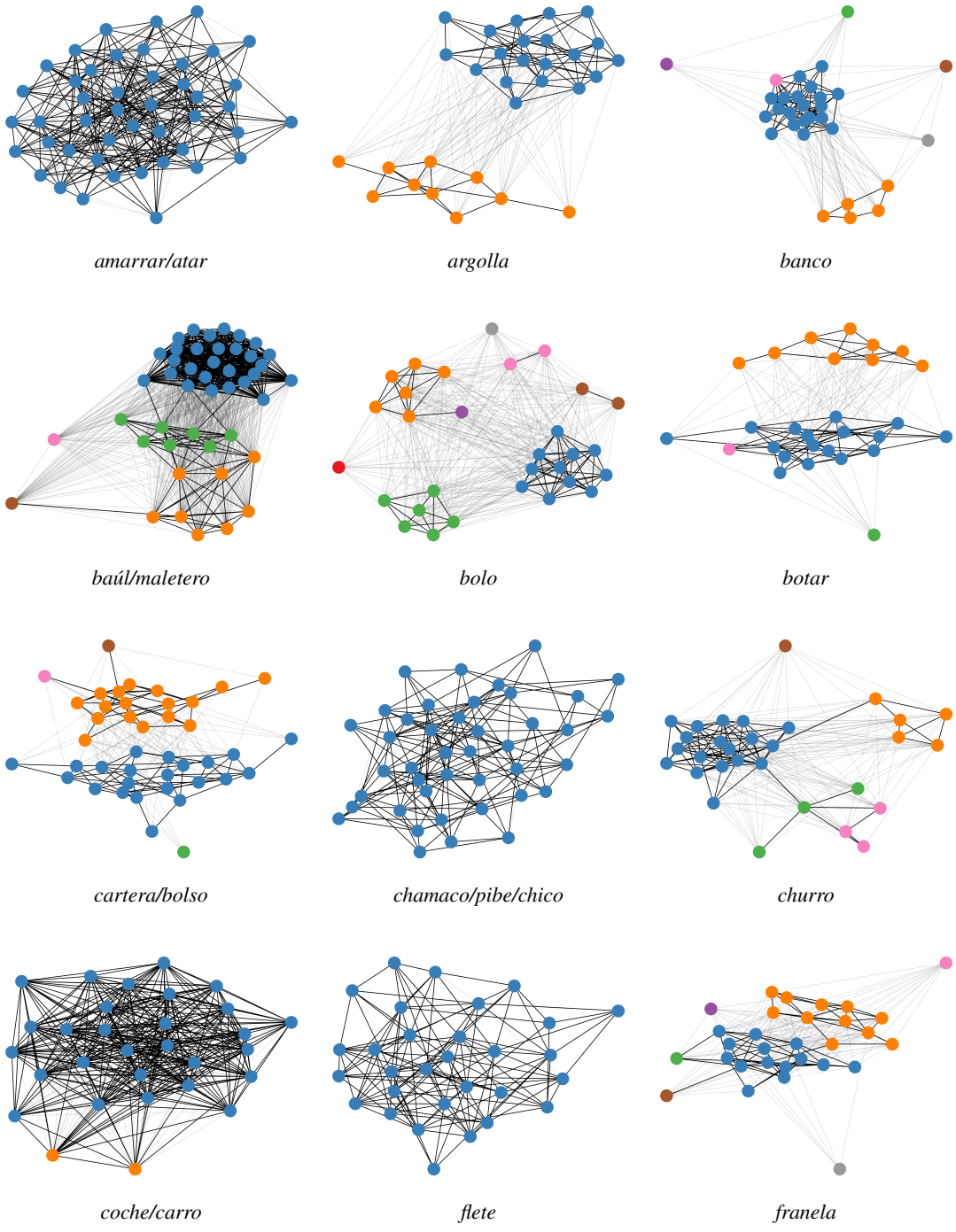


Figure 3: Examples from the DiaWUG data set.

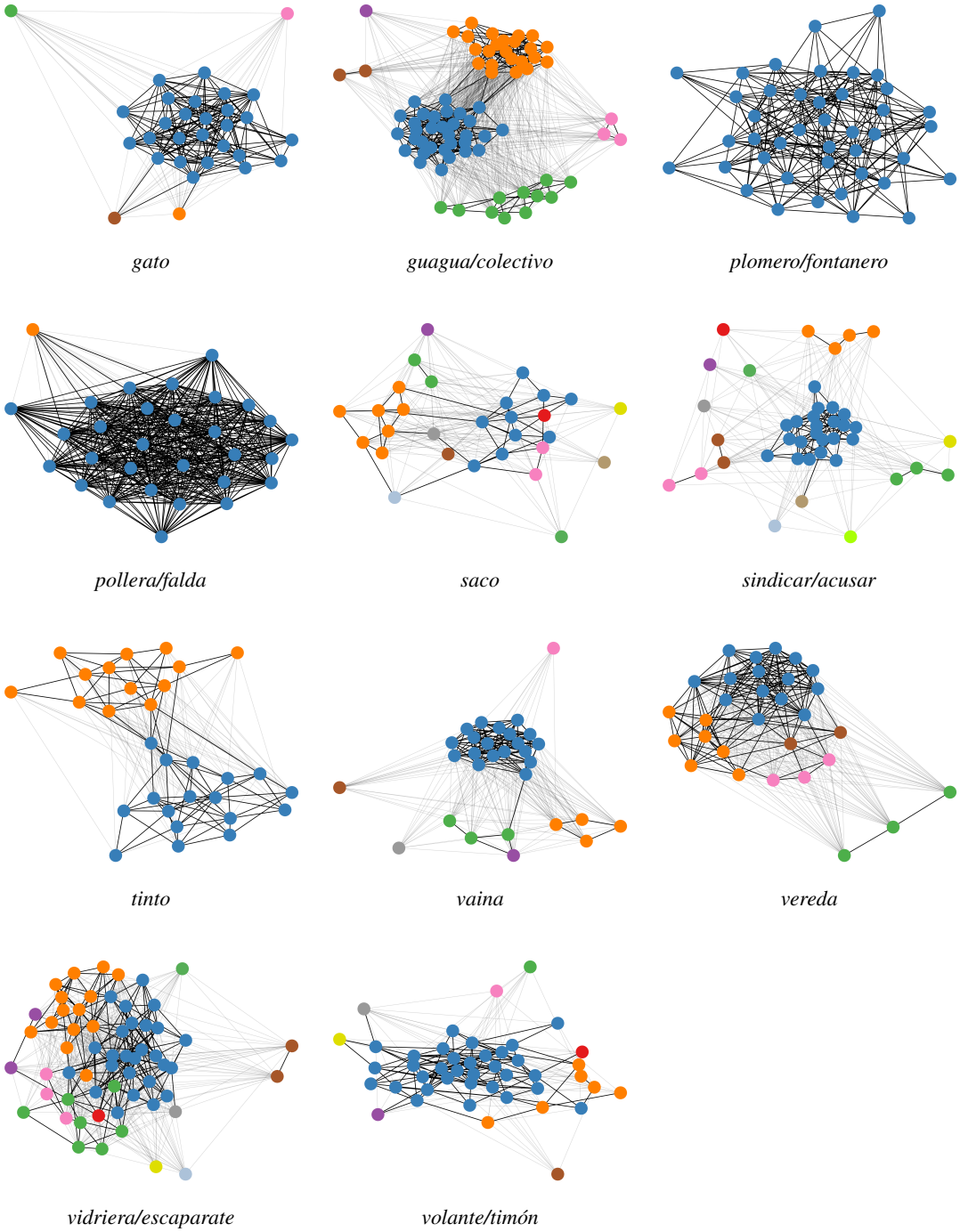


Figure 4: Examples from the DiaWUG data set.