RESCIENCE C

Replication / ML Reproducibility Challenge 2022

[Re] Exploring the Representation of Word Meanings in Context

Matteo Brivio[1] and Çağrı Çöltekin[1]
[1]University of Tübingen, Tübingen, Germany

Edited by
Koustuv Sinha,
Maurits Bleeker,
Samarth Bhargav

Received
04 February 2023

Published
20 July 2023

DOI
10.5281/zenodo.8173658

Reproducibility Summary

This report summarizes our reproduction of the ACL2021 paper Exploring the Representation of Word Meanings in Context: A Case Study on Homonymy and Synonymy by Garcia[1].

Scope of Reproducibility — The original author looks at both static and contextualized word embeddings to assess their ability to adequately represent different lexical-semantic relations, such as homonymy and synonymy. While the author describes experiments with a number of contextualized and static models, we limit our reproducibility attempt to the results reported for BERT and fastText. We also extend the original experiment by compiling a new Italian dataset and report our findings for this additional resource.

Methodology — We rely on the existing code-base, modifying it where necessary and integrating it with a few additional scripts for data preparation and statistics computation. Our code is available at https://github.com/matteobrv/repro-homonymy-acl21.

Results — We only manage to partially reproduce the original scores. Nonetheless, the hypothesis formulated in the original paper are still corroborated.

What was easy — Overall, the paper is clear and provides a good overview of the experiments. It outlines the structure of the data-sets and how they were compiled. The code and the data are available at https://github.com/marcospln/homonymy_acl21.

What was difficult — An amended version of the original paper with additional details about the experiments is available on arXiv [2]. We initially relied on the ACL [1] version which led to some minor issues during the reproducibility attempt. The code-base does not include the script used to compute the reported statistics, but upon request the author provided a preliminary version which we re-implemented. Lastly, due to some minor bugs and lack of information about the version of the libraries being used, some minor changes to the original code-base were necessary.

Communication with original authors — We exchanged a number of emails with the author to discuss implementation details and discrepancies in the reproducibility results. We received prompt and helpful responses to all of our questions.

Copyright © 2023 M. Brivio and Ç. Çöltekin, released under a Creative Commons Attribution 4.0 International license.
Correspondence should be addressed to Çağrı Çöltekin (ccoltekin@sfs.uni-tuebingen.de)
The authors have declared that no competing interests exist.
Code is available at https://github.com/matteobrv/repro-homonymy-acl21 – DOI 10.5281/zenodo.7886291. – SWH swh:1:dir:c0cbb81bf15ace80d98da9d95fed3440046539f9.
Open peer review is available at https://openreview.net/forum?id=Od5dD58libt.

1 Introduction

The distributional hypothesis, the idea that the statistical distribution of linguistic items in context plays a key role in characterizing their semantic behavior [3, 4], lies at the heart of modern word representation models. When trained on large enough data-sets, these models allow to project linguistic items into a unified vector space [5], producing distributional vectors commonly referred to as word embeddings.

A major distinction can be made between static and contextualized word embeddings. Static word vector representations obtained through models such as Word2Vec [6] and fastText [7] are context independent, that is to say a single word is always mapped to the same vector, irrespective of its context. On the other hand, in contextualized representations, produced by models such as BERT [8], each word vector is dependent on and embeds at least some information about its particular context of occurrence [9].

The paper we examine [1] investigates whether and to what extent both static and contextualized embeddings are able to adequately represent different lexical-semantic relations, such as homonymy and synonymy. In other words, it evaluates whether such models can discriminate between unrelated meanings represented by the same word form (homonymy) and identify the same sense conveyed by different words (synonymy). To this end the author compiles data-sets for four language varieties – English, Spanish, Portuguese and Galician – and carries out four experiments (see section 2). We manage to partially reproduce the original results and further test the author's hypotheses on a newly-compiled Italian data-set. Despite some discrepancies, our observations validate the claims of the original author.

2 Scope of reproducibility

For each language variety Garcia[1] carries out four experiments and reports the results in Table 4 of the original paper. We try to reproduce these results and in doing so to verify the hypothesis formulated for each of the four experiments.

- **Experiment 1**: given a target-word and three identical words in different contexts, two of them synonyms of the target, we test whether the model correctly discriminates the outlier. As an example, consider rows 1.1 and 1.2 in Table 1. **Hypothesis**: static embeddings are expected to fail, producing three identical vectors, while contextualized models should correctly identify the outlier.

- **Experiment 2**: given a target-word and three different words in different contexts, two of them synonyms of the target, we test whether the model is biased towards one specific sense. For example, consider rows 2.1 and 2.2 in Table 1. The target-word *coach* can either mean *bus* or *trainer*. If the model were biased towards the latter sense, it would always represent *coach* closer to *trainer* even in sentences where this would not make sense. **Hypothesis**: biased models will not identify the outlier sense. However, incorporating contextual information might be helpful.

- **Experiment 3**: given a target-word and three words in different contexts, two of them synonyms of the target, we test whether the model correctly discriminates the outlier. In this experiment the outlier is an homonym of one of the two synonyms. For example, consider rows 3.1 and 3.2 in Table 1. **Hypothesis**: static embeddings are likely to incorrectly represent homonyms closer than synonyms, while contextualized embeddings should model the three words correctly.

- **Experiment 4**: given a target-word and three different words, two of them synonyms of the target, we test whether the model correctly discriminates the outlier. In this experiment the outlier has to be in the same context as at least one of the two synonyms. As an example, consider rows 4.1 and 4.2 in Table 1. **Hypothesis**:

static embeddings may pass the test as they tend to represent type-level synonyms closely. Contextualized models, on the other hand, might be puzzled at targets with different meanings occurring in the same context.

3 Methodology

3.1 Datasets

We work with the four data-sets provided by Garcia[1] and with an additional Italian data-set that we compiled ourselves. Each of the five data-sets is composed of triples, where each triple is characterized by a specific target-word and three sentences. Two of the sentences contain synonyms of the target-word or the target itself. The remaining sentence might contain either an homonym of the target, an homonym of one of the synonyms or an unrelated word. For each triple three features are given: `POS`, `Context` and `Overlap`.

The three features are obtained through a comparison between `Sent.1` and `Sent.2`, `Sent.1` and `Sent.3` as well as `Sent.2` and `Sent.3`. This allows to control for context, word and POS-tag overlap. As an example, consider row 4.2 in Table 1. For this triple the value of `Context` is `false|true|false`, as the first and last word occur in the same context, thus making the second comparison (`Sent.1` vs `Sent.3`) true. As another example, consider row 3.2. In this triple the value of `Overlap` is `false|false|true`, as the words in `Sent.2` and `Sent.3` share the same form. Lastly, looking at the same triple, the value of `POS` is `same|same|same`, as the three words are all nouns.

	Target	Sent. 1	Sent. 2	Sent. 3
1.1	Coach	He was appointed as the new **coach**.	She joined the team as **coach**.	We go to the airport by **coach**.
1.2	Match	He watched the football **match**.	Chelsea have a **match** with United.	He lit the **match** on his shoe.
2.1	Coach	We go to the airport by **coach**.	They traveled by **bus**.	She was appointed as the new **trainer**.
2.2	Bank	I used to work in a **bank**.	Banks are **financial institutions**.	They camped on the **shores** of the lake.
3.1	Spring	We planted flowers in **spring**.	Cherry trees bloom in the **springtime**.	The **spring** in the fuel pump is broken.
3.2	Lead	They have an **advantage** of 28 points.	Before his goal, the team had the **lead**.	The detectives are chasing a new **lead**.
4.1	Drop	Temperatures **drop** to freezing at night.	Temperatures **fall** to freezing at night.	Temperatures **rise** to freezing at night.
4.2	Duck	I **duck** to dodge the ball.	She **lowers** her gaze.	I **raise** to dodge the ball.

Table 1. Triples examples, two for each experiment. For each triple, two sentences contain words with the same meaning as the target, while the remaining one does not.

The English, Spanish, Portuguese and Galician data-sets consist of 709, 645, 358 and 1365 triples, respectively. The Italian data-set is smaller and comprises 243 samples. Following Garcia[1], we only consider triples containing words with the same POS-tag. We report the total number of such triples and their distribution per experiment in Table 2.

3.2 Model descriptions

We work with two types of models, BERT monolingual and fastText. BERT models are loaded through the Transformers library [10]. For English, we rely on the base-uncased

Language	E1	E2	E3	E4	Total
Galician	105	149	229	135	618
English	52	58	91	68	269
Portuguese	41	37	74	41	193
Spanish	49	71	110	59	289
Italian	59	80	54	50	243

Table 2. Total number of triples in which the `POS` feature is `same` | `same` | `same` together with their distribution per experiment.

BERT model by Devlin et al.[8]. For Portuguese and Spanish we use the base-cased models released by Souza, Nogueira, and Lotufo[11] and Cañete et al.[12], respectively. For Galician, we rely on the small-cased model trained by the original author while for Italian we use the base-cased model originally provided by Schweter[13].
For fastText we use models of 300 dimensions and experiment with three architectures: skip-gram [7], CBOW [14] and MCBOW, a variation of the latter [15]. Specifically, for English we work both with the official MCBOW[1] and skip-gram[2] versions. For Spanish we use both the official skip-gram[3] and CBOW[4] models. For Portuguese we rely on the skip-gram[5] version by Hartmann et al.[16]. For Galician we use the skip-gram[6] model trained by Garcia[1] and for Italian the official skip-gram[7] version.

3.3 Experimental setup and code

The four experiments described in section 2 rely on BERT and fastText embeddings. For both architectures a number embedding creation strategies are explored. We only describe those for which a result is reported in Table 4 of the original paper. Note that whenever a word consists of more than one token (e.g. *financial institutions*) the average of the token embeddings is considered.
For BERT models the following three approaches are tested:

- **Sentence vector (*Sent*)**: an embedding obtained by averaging the representations of all the words in a given sentence, but the `[CLS]` and `[SEP]` tokens. Each representation is obtained by concatenating the vectors produced in the last four layers of the model.
- **Word vector sum (*Add*)**: an embedding of a specific word obtained by summing its representations across the last four layers of the model.
- **Word vector concatenation (*Cat*)**: an embedding of a specific word obtained by concatenating its representations across the last four layers of the model.

For fastText models the following three approaches are tested:

- **Word vector (*WV*)**: an embedding of a specific word.
- **Sentence vector (*Sent*)**: an embedding obtained by averaging the representations of all the words in a given sentence.

[1] English MCBOW fastText: https://dl.fbaipublicfiles.com/fasttext/vectors-english/wiki-news-300d-1M.vec.zip
[2] English skip-gram fastText: https://dl.fbaipublicfiles.com/fasttext/vectors-wiki/wiki.en.vec
[3] Spanish skip-gram fastText: https://dl.fbaipublicfiles.com/fasttext/vectors-wiki/wiki.es.vec
[4] Spanish CBOW fastText: https://dl.fbaipublicfiles.com/fasttext/vectors-crawl/cc.es.300.vec.gz
[5] Portuguese skip-gram fastText: http://143.107.183.175:22980/download.php?file=embeddings/fasttext/skip_s300.zip
[6] Galician skip-gram fastText: https://zenodo.org/record/4481614/files/fasttext_sg_300d_w5.zip?download=1
[7] Italian skip-gram fastText: https://dl.fbaipublicfiles.com/fasttext/vectors-wiki/wiki.it.vec

- **Syntax (*Syn3*)**: an embedding of a specific word obtained by adding the word representation to those of its syntactic head and dependents. The underlying assumption is that the syntactic context of a word would characterize its meaning. Generating this embedding requires converting each sentence in each triple to a CoNLL-U format.

The output of each model is evaluated as follows: for each triple, where two words (*a* and *b*) are synonyms and a third one (*c*) has a different meaning, three cosine similarities between their embeddings are computed: *sim1* = `cos(a, b)`, *sim2* = `cos(a, c)` and *sim3* = `cos(b, c)`. Ultimately, the model should be able to discriminate the outlier *c* i.e. the output for a specific triple is correct only if *sim1* > *sim2* and *sim1* > *sim3*.

For each model and embedding creation strategy we compute four accuracy scores, one per experiment, and provide macro- and micro-average results across them. Lastly, for each embedding creation strategy, we also report the micro-average accuracy score across the entire set of generated embeddings. It is worth noting that for each language variety the corresponding data-sets also contain triples of sentences that are not part of any of the four experiments described in Section 2.

We only report results for triples in which the synonyms and homonyms belong to the same word category (`POS` = `same`|`same`|`same`). This allows to focus on the semantic knowledge encoded in their embeddings rather than on the morpho-syntactic information.

We rely on the code-base released by the original author, modifying it where necessary and integrating it with a few additional scripts for statistics computation and data preparation.

3.4 Computational requirements

We work with pre-trained BERT and fastText models, relying on a 3.10GHz Intel Core i9-7940X CPU. The total running-time for the experiment is about 30 minutes.

4 Results

4.1 Results reproducing the original paper

Our results partially match those reported in the original study. In particular, we successfully reproduce the accuracy scores for the Galician BERT and fastText models, as well as for the English BERT and Portuguese fastText ones. However, we also observe a number of discrepancies across English, Spanish and Portuguese. We plot the difference between our results and the original ones in Figure 1 and report the exact values in Table 4 in the Appendix. Galician results are not included as they match the original ones but we provide them in our GitHub repository.

Looking at BERT, the scores we obtain for Portuguese and Spanish are mostly higher than those originally indicated, with only two exceptions for Spanish in experiment 2 (BERT sent) and experiment 4 (BERT add). Turning to fastText, the results for the English and Spanish skip-gram models also deviate from the original ones. Nonetheless, relying on the MCBOW [15] and CBOW [14] implementations we manage to reproduce the original English and Spanish scores, respectively (see Table 4).

Overall, BERT models perform consistently better than fastText in the first three experiments across all languages, while fastText models score higher accuracy values in the fourth one.

4.2 Results beyond the original paper

We summarize our findings for the Italian data-set in Table 3. Overall, the trend across the four experiments is consistent with the one we observe for the original language

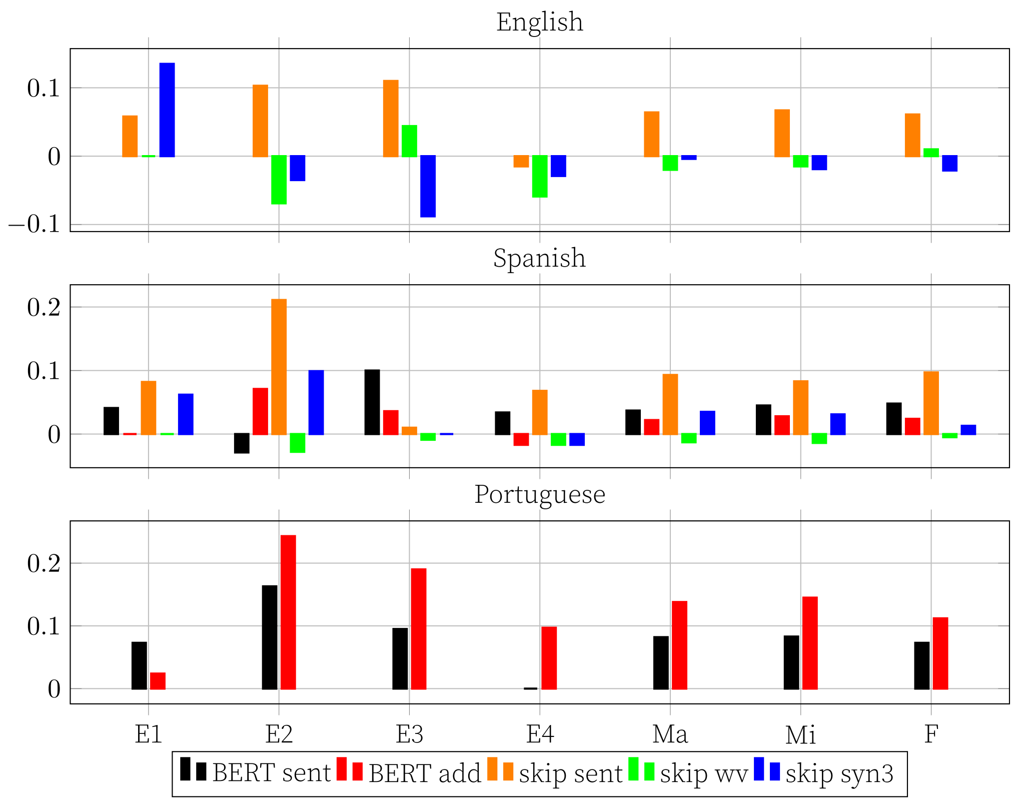

Figure 1. Difference between our accuracy scores and those reported in the original paper computed across the four experiments, as well as for the macro (Ma), micro (Mi) and full (F) accuracy. For each language, only the models for which a discrepancy is observed are considered.

varieties, with BERT achieving the highest accuracy in the first three experiments for at least one embedding type (Vec) and fastText being the top performer in the last experiment. It is worth noting that the micro-average results across the four experiments (Mi) and the micro-average values on the whole data-set (F) are the same for each embedding type.

Model	Vec	E1	E2	E3	E4	Ma	Mi	F
	Sent	0.763	0.8	0.741	0.14	0.611	0.642	0.642
BERT	Add	0.831	0.888	0.463	0.3	0.620	0.658	0.658
	Cat	0.831	0.875	0.444	0.24	0.597	0.638	0.638
	Sent skip	0.763	0.863	0.704	0.16	0.622	0.658	0.658
fastText	WV skip	0	0.462	0	0.62	0.271	0.28	0.28
	Syn 3 skip	0.576	0.75	0.389	0.22	0.484	0.519	0.519

Table 3. Summary of the BERT and fastText results for Italian. For each embedding type (Vec) we report the accuracy scores of the four experiments, together with their macro- and micro-average results. Lastly, we indicate the micro-average on the whole data-set (F).

5 Discussion

For all five language varieties, our results support the hypotheses formulated in Section 2. Specifically, across all languages, BERT achieves higher accuracy scores in the first three experiments, confirming that contextualized models are better at discriminating an outlier sense when dealing with different sentences. At the same time, these models rely heavily on the surrounding context and struggle to tell apart target-words with different meanings that occur in similar sentences. This is confirmed by the poor performance of BERT models in the fourth experiment which also corroborates the last hypothesis.

Despite confirming the original claims, our results partially deviate from those originally reported. With respect to BERT, the accuracy scores we observe for Spanish and Portuguese are overall higher than the original ones. In this regard, it is worth mentioning the result we obtain for Portuguese in the second experiment. While the original study indicates a highest score of 0.541 (Add(o)), which is too low to support the second hypothesis, we register a score of 0.784 (Add(r)) which is better than the fastText ones. We try to investigate the cause of these discrepancies, but the respective models show no recent updates that could motivate them. However, it should be noted that the original contribution does not specify the version of the Transformers library [10] being used and that we likely rely on a more recent version to carry out our experiment. Turning to fastText, our results for English and Spanish are also inconsistent with those originally reported. For these languages the original author claims to rely on 300-dimensional vectors obtained through the skip-gram model described in Bojanowski et al.[7]. However, we only manage to reproduce the English and Spanish scores using vectors obtained through the CBOW [14] and MCBOW [14] implementations, respectively.

Coming to the Italian results, we observe that the micro-averages across the four experiments (Mi) and on the whole data-set (F) are the same for each embedding type (Vec). This is not the case for the language varieties considered in the original study. The reason for this is that the original data-sets also contain triples of sentences that are not contemplated in the four experiments, for example triples in which the synonyms and homonyms belong to different word categories.

In summary, despite some discrepancies, our results are consistent with the findings of the original paper and support the initial hypotheses.

References

1. M. Garcia. "Exploring the Representation of Word Meanings in Context: A Case Study on Homonymy and Synonymy." 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: Association for Computational Linguistics, Aug. 2021, pp. 3625–3640. DOI: 10.18653/v1/2021.acl-long.281. URL: https://aclanthology.org/2021.acl-long.281.
2. M. Garcia. "Exploring the Representation of Word Meanings in Context: A Case Study on Homonymy and Synonymy." In: **CoRR** (2021). DOI: 10.48550/ARXIV.2106.13553. URL: https://arxiv.org/abs/2106.13553.
3. Z. S. Harris. "Distributional Structure." In: **Word** 10.2–3 (1954), pp. 146–162. DOI: 10.1080/00437956.1954.11659520. eprint: https://doi.org/10.1080/00437956.1954.11659520. URL: https://doi.org/10.1080/00437956.1954.11659520.
4. A. Lenci. "Distributional Models of Word Meaning." In: **Annual Review of Linguistics** 4.1 (2018), pp. 151–171. DOI: 10.1146/annurev-linguistics-030514-125254. URL: https://doi.org/10.1146/annurev-linguistics-030514-125254.
5. Z. Liu, Y. Lin, and M. Sun. "Representation Learning and NLP." In: **Representation Learning for Natural Language Processing**. Springer Singapore, 2020, pp. 1–11. DOI: 10.1007/978-981-15-5573-2_1. URL: https://doi.org/10.1007/978-981-15-5573-2_1.
6. T. Mikolov, K. Chen, G. Corrado, and J. Dean. "Distributed representations of words and phrases and their compositionality." In: **Advances in neural information processing systems** 26 (2013).
7. P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov. "Enriching Word Vectors with Subword Information." In: **Transactions of the Association for Computational Linguistics** 5 (2017), pp. 135–146.
8. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." In: **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)**. Minneapolis, Minnesota: Association for Computational Linguistics, 2019, pp. 4171–4186. DOI: 10.18653/v1/N19-1423. URL: https://aclanthology.org/N19-1423.

9. A. Rogers, O. Kovaleva, and A. Rumshisky. "A Primer in BERTology: What We Know About How BERT Works." In: **Transactions of the Association for Computational Linguistics** 8 (2021), pp. 842–866. DOI: 10.1162/tacl_a_00349. eprint: https://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00349/1923281/tacl_a_00349.pdf. URL: https://doi.org/10.1162/tacl%5C_a%5C_00349.
10. T. Wolf et al. "Transformers: State-of-the-Art Natural Language Processing." In: **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations**. Online: Association for Computational Linguistics, 2020, pp. 38–45. URL: https://www.aclweb.org/anthology/2020.emnlp-demos.6.
11. F. Souza, R. Nogueira, and R. Lotufo. "BERTimbau: pretrained BERT models for Brazilian Portuguese." In: **9th Brazilian Conference on Intelligent Systems, BRACIS, Rio Grande do Sul, Brazil, October 20-23 (to appear)**. 2020.
12. J. Cañete, G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, and J. Pérez. "Spanish Pre-Trained BERT Model and Evaluation Data." In: **PML4DC at ICLR 2020**. 2020.
13. S. Schweter. **Italian BERT and ELECTRA models**. Version 1.0.1. 2020. DOI: 10.5281/zenodo.4263142. URL: https://doi.org/10.5281/zenodo.4263142.
14. E. Grave, P. Bojanowski, P. Gupta, A. Joulin, and T. Mikolov. "Learning Word Vectors for 157 Languages." In: **Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)**. 2018.
15. T. Mikolov, E. Grave, P. Bojanowski, C. Puhrsch, and A. Joulin. "Advances in Pre-Training Distributed Word Representations." In: **Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)**. 2018.
16. N. Hartmann, E. Fonseca, C. Shulby, M. Treviso, J. Silva, and S. Aluísio. "Portuguese Word Embeddings: Evaluating on Word Analogies and Natural Language Tasks." In: **Proceedings of the 11th Brazilian Symposium in Information and Human Language Technology**. Uberlândia, Brazil: Sociedade Brasileira de Computação, Oct. 2017, pp. 122–131. URL: https://aclanthology.org/W17-6615.

Model	Vec.	E1	E2	E3	E4	Ma	Mi	F	E1	E2	E3	E4	Ma	Mi	F	E1	E2	E3	E4	Ma	Mi	F
		Portuguese							Spanish							English						
BERT	Sent (o)	0.683	0.432	0.635	0.22	0.493	0.518	0.564	0.755	0.592	0.536	0.186	0.517	0.516	0.595	0.788	0.655	0.736	0.221	0.6	0.599	0.7
	Sent (r)	0.756	0.595	0.73	0.22	0.575	0.601	0.637	0.796	0.563	0.636	0.220	0.554	0.561	0.643	0.788	0.655	0.736	0.221	0.6	0.599	0.7
	Cat (o)	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	Cat (r)	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	Add (o)	0.854	0.541	0.378	0.366	0.535	0.508	0.67	0.857	0.704	0.409	0.441	0.603	0.564	0.74	0.981	0.81	0.758	0.441	0.748	0.732	0.839
	Add (r)	0.878	0.784	0.568	0.463	0.673	0.653	0.782	0.857	0.775	0.445	0.424	0.625	0.592	0.764	0.981	0.81	0.758	0.441	0.748	0.732	0.839
fastText	Sent skip (o)	0.61	0.622	0.527	0.171	0.482	0.487	0.55	0.449	0.338	0.445	0.085	0.329	0.346	0.429	0.596	0.5	0.505	0.147	0.437	0.431	0.543
	Sent skip (r)	0.61	0.622	0.527	0.171	0.482	0.487	0.55	0.531	0.549	0.455	0.153	0.422	0.429	0.526	0.654	0.603	0.615	0.132	0.501	0.498	0.604
	Sent CBOW (r)	-	-	-	-	-	-	-	0.449	0.338	0.445	0.085	0.329	0.346	0.429	-	-	-	-	-	-	-
	Sent MCBOW (r)	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.596	0.5	0.505	0.147	0.437	0.431	0.543
	WV skip (o)	0.024	0.541	0	0.634	0.3	0.244	0.453	0.122	0.62	0.018	0.814	0.393	0.346	0.479	0.308	0.552	0.033	0.574	0.366	0.335	0.48
	WV skip (r)	0.024	0.541	0	0.634	0.3	0.244	0.453	0.122	0.592	0.009	0.797	0.38	0.332	0.474	0.308	0.483	0.077	0.515	0.346	0.32	0.49
	WV CBOW (r)	-	-	-	-	-	-	-	0.122	0.62	0.018	0.814	0.393	0.346	0.479	-	-	-	-	-	-	-
	WV MCBOW (r)	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.308	0.552	0.033	0.574	0.366	0.335	0.48
	Syn 3 skip (o)	0.659	0.459	0.176	0.195	0.372	0.337	0.508	0.367	0.577	0.173	0.237	0.339	0.318	0.553	0.442	0.69	0.231	0.176	0.385	0.357	0.546
	Syn 3 skip (r)	0.659	0.459	0.176	0.195	0.372	0.337	0.508	0.429	0.676	0.173	0.220	0.374	0.349	0.566	0.577	0.655	0.143	0.147	0.381	0.338	0.525
	Syn 3 CBOW (r)	-	-	-	-	-	-	-	0.367	0.577	0.173	0.237	0.339	0.318	0.553	-	-	-	-	-	-	-
	Syn 3 MCBOW (r)	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.442	0.69	0.231	0.176	0.385	0.357	0.546

Table 4. Summary of the BERT and fastText scores for the language varieties whose results deviate from those reported in the original study. For each embedding type (Vec) we report the original (o) and reproduced (r) accuracy scores of the four experiments, together with their macro- and micro-average results. Lastly, we indicate the micro-average on the whole data-set (F). Results for Galician are not included as they match those originally reported.