A Issues in the outlier identification script by Camacho-Collados and Navigli

The evaluation script released with the 8-8-8 dataset (Camacho-Collados and Navigli, 2016) has a problematic way of handling words not present in the vector space (out-of-vocabulary words, OOV). If a word does not exist in the vector space, then it is assigned the zero vector. This decision results in an issue when the outlier is OOV. The script sets the OP of the outlier to $|W|-1$, which means that the OP score is equal to a correct identification of the outlier. This behavior improves the OPP and Accuracy, while it should reduce it.

This error is a consequence of the concrete implementation choice of adding the outlier's compactness score as the last of the nine words into the dictionary storing them. The code then sorts (stable) based on the compactness score (which is zero for all words), where nothing will change, and the outlier will stay at position $|W|-1$.

In our evaluation, we avoid this issue by ensuring that all words are present in the vector space for all test cases. We release the corrected version of the script together with our data.

B Syntactic ambiguity and multi-word tokens in datasets

As the syntactic outlier identification task relies on clustering words with the same POS tag, we only use words with a single POS tag. This decision introduces a problem when creating the German 25-8-8-Syn, as there is no lexical difference between German adjectives and the corresponding adverbs. In contrast, in English and Italian, a suffix marks the modification of an adjective into an adverb. Since in German, all adjectives can also serve as adverbs, the German versions of the adjective tests have two parts of speech: adjective and adverb. We, therefore, make sure the outliers do not contain adverbs. This decision does not have an influence on the adverb categories as adverbs are not necessarily adjectives.

Italian constructs comparative adjectives by preceding the adjective by a modifier followed by the adjective (e.g., *veloce* becomes *più veloce*), making the word a multi-token word. The multi-token words are not a problem for our comparative adjacency categories, as the comparative is a constant added in all cases.

C Experimental Setup

For reproduction, we use the same hyper-parameters as Camacho-Collados and Navigli: dimensionality 300, minimum frequency 5, skip-gram based models (skip-gram, word2vecf, word2vecf+)[9] have negative sampling 15. To study the effect of the window size hyper-parameter, we trained models with window sizes of two, five, and 10, where applicable. We set the rest of the parameters to their default values. Each result presented is the mean of 10 training runs; we also provide variance information.

D 3CosMul evaluation

We use 3CosAdd to evaluate word analogies, where the vector offset of pair (a, a^*) and b combine to produce a vector b^* such that the pair (b, b^*) is analogous. In other words, the answer to the analogy us the result of optimizing:

$$\underset{b^* \in V}{\operatorname{argmax}}(cos(b^*, b) - cos(b^*, b) + cos(b^*, a^*))$$

As shown by Levy and Goldberg (2014b), certain aspects of vectors may dominate in the similarity measure, i.e., a *king* is more royal than masculine, causing the royalty aspect to dominate in the similarity measure. Levy and Goldberg therefore propose *3CosMul* as a better alternative to *3CosAdd*, which instead optimizes:

$$\underset{b^* \in V}{\operatorname{argmax}} \left(\frac{cos(b^*, b) \cdot cos(b^*, a^*)}{cos(b^*, a) + \epsilon} \right)$$

Where $\epsilon = 0.001$ and is used to prevent division by 0.

Table 5 shows the results of evaluating the word analogy task using both *3CosAdd* and *3CosMul*. As can be seen, every model scores higher when using *3CosMul* rather than *3CosAdd*. However, other than SG 2 passing SG 5 by a small margin, the ranking between the models stay the same, indicating that *3CosMul* is a superior function for computing word analogies. However replacing *3CosAdd* with *3CosMul* in the evaluation in the main body of the paper does not provide any insights into the quality of the vector spaces since its effects are limited exclusively to solving the analogy task.

[9] We use word2vec from https://code.google.com/archive/p/word2vec/source/default/source and word2vecf from https://bitbucket.org/yoavgo/word2vecf/src/default/

	CBOW 2	CBOW 5	CBOW 10	SG 2	SG 5	SG 10	W2VF	W2VF+
3CosAdd	0.2755	0.2753	0.2795	0.3533	0.3539	0.3521	0.3139	0.3401
3CosMul	0.2895	0.2873	0.2906	0.3681	0.3675	0.3637	0.3235	0.3526

Table 5: Results from running the UMBC trained models with the 3CosMul script provided by Levy et al. (2015).

E Detailed word analogy results

As explained by Gladkova et al. (2016), the dataset provided for the analogy task is unbalanced across different types of relations. Therefore, Gladkova et al. suggest including performance results on the individual categories to provide an understanding of performance on the various categories and not only the total, aggregate, performance. Based on this suggestion, we have included the results for each category in Tables 6 to 9.

Categories	**CBOW ws 2**	**CBOW ws 5**	**CBOW ws 10**	**Skip-gram ws 2**	**Skip-gram ws 5**	**Skip-gram ws 10**	**Word2VecF**	**Word2VecF+**
capital-common-countries	25.08 ± 1.16	43.62 ± 0.88	50.12 ± 0.38	78.20 ± 1.51	**83.06** ± 0.44	82.41 ± 1.11	20.36 ± 2.03	62.35 ± 1.80
capital-world	7.77 ± 0.05	14.40 ± 0.14	23.19 ± 0.07	56.64 ± 0.18	69.12 ± 0.12	**74.18** ± 1.03	5.18 ± 0.09	27.30 ± 0.42
currency	2.41 ± 0.05	4.04 ± 0.12	4.23 ± 0.10	9.49 ± 0.14	11.81 ± 0.12	**13.88** ± 0.35	1.31 ± 0.03	2.49 ± 0.02
city-in-state	6.81 ± 0.11	10.35 ± 0.38	18.53 ± 0.21	61.54 ± 5.11	66.50 ± 0.52	**67.74** ± 3.23	4.36 ± 0.29	31.71 ± 1.25
family	49.94 ± 0.71	50.08 ± 0.88	54.37 ± 1.08	85.79 ± 0.93	**86.60** ± 0.68	85.00 ± 1.32	74.51 ± 1.46	71.50 ± 0.38
adjective-to-adverb	10.58 ± 0.09	15.11 ± 0.15	19.63 ± 0.17	22.72 ± 0.13	25.72 ± 0.36	**30.52** ± 0.57	8.94 ± 0.92	13.75 ± 0.12
opposite	33.64 ± 0.25	32.86 ± 0.47	33.87 ± 0.20	40.03 ± 0.47	40.01 ± 0.62	40.02 ± 1.90	42.11 ± 1.11	**44.04** ± 0.20
comparative	89.26 ± 0.13	86.44 ± 0.23	84.83 ± 0.56	92.94 ± 0.10	92.82 ± 0.21	90.58 ± 2.72	90.65 ± 0.48	**93.57** ± 0.14
superlative	59.30 ± 0.44	55.52 ± 0.30	55.46 ± 0.57	**90.77** ± 0.28	85.19 ± 0.50	82.01 ± 5.93	68.84 ± 0.84	79.26 ± 0.82
present-participle	65.75 ± 0.29	66.05 ± 0.46	65.53 ± 0.43	76.77 ± 0.32	75.43 ± 0.65	**77.73** ± 1.52	63.11 ± 2.83	76.00 ± 0.37
nationality-adjective	30.01 ± 0.31	40.11 ± 0.45	47.36 ± 0.61	66.17 ± 1.07	76.67 ± 0.27	**80.66** ± 0.40	5.91 ± 0.08	48.57 ± 0.35
past-tense	46.39 ± 0.09	46.73 ± 0.54	45.49 ± 0.32	56.07 ± 0.44	55.59 ± 0.19	54.58 ± 0.36	58.26 ± 0.37	**64.02** ± 0.31
plural	67.66 ± 0.18	67.32 ± 0.53	69.34 ± 0.33	82.19 ± 0.62	**83.08** ± 0.35	82.67 ± 0.58	76.31 ± 0.93	79.64 ± 0.44
plural-verbs	58.73 ± 0.33	57.85 ± 0.97	58.63 ± 0.37	79.38 ± 0.51	74.93 ± 1.17	65.33 ± 10.90	85.96 ± 0.87	**87.13** ± 0.54
Total semantic	10.37 ± 0.05	15.96 ± 0.06	23.36 ± 0.05	56.29 ± 0.58	64.59 ± 0.13	**67.59** ± 0.56	9.39 ± 0.08	30.63 ± 0.23
Total syntactic	51.92 ± 0.03	53.01 ± 0.04	54.47 ± 0.06	68.72 ± 0.07	**69.51** ± 0.07	69.19 ± 0.80	54.75 ± 0.16	65.82 ± 0.03
Total Accuracy	33.06 ± 0.03	36.20 ± 0.02	40.35 ± 0.04	63.08 ± 0.13	67.28 ± 0.06	**68.46** ± 0.31	34.17 ± 0.03	49.85 ± 0.05

Table 6: Results from the different models trained on UMBC in the Word Analogy task.

Categories	**CBOW ws 2**	**CBOW ws 5**	**CBOW ws 10**	**Skip-gram ws 2**	**Skip-gram ws 5**	**Skip-gram ws 10**	**Word2VecF**	**Word2VecF+**
capital-common-countries	58.22 ± 1.47	64.35 ± 0.45	72.31 ± 0.79	91.58 ± 0.45	95.34 ± 0.39	**96.96** ± 0.59	42.65 ± 5.64	88.44 ± 0.77
capital-world	29.29 ± 0.21	44.28 ± 0.36	65.49 ± 0.14	84.55 ± 0.19	89.81 ± 0.17	**91.54** ± 0.19	11.25 ± 0.32	61.33 ± 0.68
currency	6.97 ± 0.22	12.49 ± 0.33	15.49 ± 0.26	15.91 ± 0.06	20.65 ± 0.43	**20.68** ± 0.19	3.83 ± 0.03	7.56 ± 0.11
city-in-state	7.57 ± 0.25	12.18 ± 0.35	30.79 ± 0.31	65.91 ± 0.71	**71.35** ± 0.45	70.46 ± 0.19	9.29 ± 0.27	39.07 ± 0.85
family	75.32 ± 0.33	75.41 ± 0.58	72.67 ± 1.06	**80.71** ± 1.31	80.44 ± 1.01	80.16 ± 0.20	71.76 ± 0.86	69.43 ± 1.79
adjective-to-adverb	7.53 ± 0.09	10.41 ± 0.10	12.50 ± 0.05	16.99 ± 0.10	19.25 ± 0.29	**26.77** ± 1.02	8.73 ± 0.25	9.47 ± 0.19
opposite	23.47 ± 0.20	23.77 ± 0.30	23.45 ± 0.23	33.69 ± 1.25	35.25 ± 0.52	32.12 ± 0.34	31.96 ± 0.67	**35.59** ± 0.25
comparative	70.15 ± 0.29	69.04 ± 0.11	70.11 ± 0.36	81.55 ± 0.35	**83.00** ± 0.13	79.57 ± 0.58	77.54 ± 0.20	79.92 ± 0.15
superlative	31.62 ± 0.62	31.54 ± 0.08	30.66 ± 0.15	52.26 ± 1.91	**52.77** ± 1.13	46.55 ± 0.57	39.40 ± 0.80	50.88 ± 1.08
present-participle	44.63 ± 0.36	49.26 ± 0.47	48.78 ± 0.33	**67.54** ± 0.36	65.85 ± 0.50	60.90 ± 1.45	53.96 ± 1.17	67.49 ± 0.27
nationality-adjective	41.25 ± 0.04	65.68 ± 0.53	80.01 ± 0.15	87.41 ± 0.27	**90.47** ± 0.06	90.21 ± 0.04	8.33 ± 0.59	76.39 ± 0.41
past-tense	51.97 ± 0.12	53.13 ± 0.22	55.76 ± 0.21	61.30 ± 0.18	61.22 ± 0.60	57.86 ± 0.79	64.24 ± 0.22	**73.03** ± 0.31
plural	54.27 ± 0.72	53.99 ± 0.28	52.51 ± 0.48	70.86 ± 0.59	73.08 ± 0.75	**74.74** ± 0.49	53.46 ± 0.48	68.63 ± 0.21
plural-verbs	53.98 ± 0.06	52.34 ± 0.27	55.52 ± 0.37	78.00 ± 0.42	71.87 ± 0.18	64.91 ± 0.73	77.81 ± 0.30	**79.66** ± 0.47
Total semantic	25.34 ± 0.09	35.17 ± 0.17	51.75 ± 0.02	72.84 ± 0.09	77.70 ± 0.14	**78.42** ± 0.04	15.22 ± 0.22	51.90 ± 0.35
Total syntactic	43.92 ± 0.03	48.18 ± 0.06	50.95 ± 0.02	63.74 ± 0.09	**64.36** ± 0.04	62.36 ± 0.09	46.05 ± 0.03	62.76 ± 0.03
Total Accuracy	35.49 ± 0.03	42.28 ± 0.02	51.32 ± 0.01	67.87 ± 0.07	**70.41** ± 0.03	69.65 ± 0.03	32.06 ± 0.04	57.83 ± 0.09

Table 7: Results from the different models trained on English Wikipedia in the Word Analogy task.

Categories	**CBOW ws 2**	**CBOW ws 5**	**CBOW ws 10**	**Skip-gram ws 2**	**Skip-gram ws 5**	**Skip-gram ws 10**	**Word2VecF**	**Word2VecF+**
capital-common-countries	46.68 ± 1.65	59.31 ± 1.15	63.40 ± 1.04	86.42 ± 0.70	91.90 ± 0.66	**93.24** ± 0.33	13.12 ± 0.34	52.71 ± 5.91
capital-world	18.88 ± 0.21	29.12 ± 0.31	37.31 ± 0.19	69.91 ± 0.42	85.41 ± 0.17	**88.76** ± 0.26	4.30 ± 0.04	22.00 ± 0.30
currency	2.97 ± 0.06	5.15 ± 0.26	6.88 ± 0.22	7.07 ± 0.31	11.15 ± 0.27	**11.22** ± 0.30	0.39 ± 0.01	1.08 ± 0.05
city-in-state	7.14 ± 0.92	6.00 ± 0.07	6.64 ± 0.15	33.73 ± 1.16	45.23 ± 1.10	**45.30** ± 0.45	3.30 ± 0.06	7.49 ± 0.21
family	30.20 ± 0.89	32.29 ± 0.52	29.98 ± 0.59	51.86 ± 0.53	59.98 ± 2.03	**61.07** ± 1.48	43.08 ± 0.39	50.34 ± 0.71
opposite	10.14 ± 0.55	10.72 ± 0.60	10.67 ± 0.72	10.40 ± 0.89	**16.31** ± 1.02	15.36 ± 0.75	1.57 ± 0.02	13.53 ± 0.45
comparative	29.77 ± 0.82	33.86 ± 0.51	33.80 ± 0.44	53.09 ± 0.57	**54.45** ± 0.85	52.34 ± 0.88	22.00 ± 0.29	50.20 ± 0.82
superlative	1.51 ± 0.03	1.84 ± 0.06	2.09 ± 0.05	4.80 ± 0.37	**6.67** ± 0.15	5.49 ± 0.25	4.63 ± 0.03	5.59 ± 0.12
present-participle	0.77 ± 0.02	0.94 ± 0.09	0.82 ± 0.03	1.86 ± 0.07	**3.16** ± 0.12	2.82 ± 0.08	0.65 ± 0.01	2.84 ± 0.10
nationality-adjective	11.52 ± 0.62	16.95 ± 0.24	19.69 ± 0.14	24.17 ± 0.16	32.10 ± 0.10	**33.93** ± 0.16	1.14 ± 0.02	8.57 ± 0.11
past-tense	14.60 ± 0.30	17.99 ± 0.37	19.10 ± 0.28	27.24 ± 0.65	26.92 ± 0.24	25.47 ± 0.41	15.43 ± 0.10	**36.55** ± 0.34
plural	10.51 ± 0.15	10.72 ± 0.50	12.98 ± 0.29	31.79 ± 0.96	36.67 ± 1.12	**44.24** ± 0.34	12.33 ± 0.49	13.89 ± 0.39
plural-verbs	38.89 ± 0.18	40.96 ± 0.09	38.27 ± 0.20	58.00 ± 0.89	**60.40** ± 0.56	59.69 ± 0.68	36.84 ± 0.29	54.28 ± 0.35
Total semantic	16.38 ± 0.07	22.36 ± 0.11	27.01 ± 0.11	53.93 ± 0.10	66.26 ± 0.29	**68.15** ± 0.08	6.40 ± 0.02	19.41 ± 0.17
Total syntactic	15.28 ± 0.06	17.70 ± 0.07	18.40 ± 0.04	28.45 ± 0.05	31.63 ± 0.08	**32.15** ± 0.03	12.30 ± 0.03	24.49 ± 0.06
Total Accuracy	15.83 ± 0.05	20.04 ± 0.07	22.71 ± 0.02	41.21 ± 0.05	48.98 ± 0.07	**50.19** ± 0.02	9.33 ± 0.01	21.95 ± 0.06

Table 8: Results from the different models trained on German Wikipedia in the Word Analogy task.

Categories	**CBOW ws 2**	**CBOW ws 5**	**CBOW ws 10**	**Skip-gram ws 2**	**Skip-gram ws 5**	**Skip-gram ws 10**	**Word2VecF**	**Word2VecF+**
capital-common-countries	13.14 ± 0.44	16.84 ± 0.19	20.22 ± 0.51	67.33 ± 2.15	83.81 ± 0.55	**88.36** ± 0.67	7.63 ± 0.40	26.17 ± 2.42
capital-world	4.36 ± 0.06	5.27 ± 0.03	7.13 ± 0.07	36.45 ± 0.42	58.15 ± 0.12	**67.02** ± 0.13	2.07 ± 0.01	6.88 ± 0.13
currency	1.09 ± 0.05	2.21 ± 0.08	2.99 ± 0.13	3.56 ± 0.07	7.04 ± 0.14	**9.92** ± 0.18	0.32 ± 0.01	1.48 ± 0.02
city-in-state	3.54 ± 0.03	3.15 ± 0.08	3.51 ± 0.10	14.42 ± 0.28	25.55 ± 0.50	**30.87** ± 0.77	2.32 ± 0.01	5.26 ± 0.17
regione-capoluogo	6.05 ± 0.39	7.37 ± 0.53	10.47 ± 1.30	25.38 ± 2.36	35.15 ± 1.31	**39.09** ± 2.14	2.81 ± 0.11	9.18 ± 1.46
family	25.32 ± 0.47	24.56 ± 0.84	22.72 ± 0.92	61.46 ± 1.32	65.67 ± 2.05	**66.17** ± 0.48	59.68 ± 1.26	59.47 ± 1.61
adjective-to-adverb	5.13 ± 0.03	6.52 ± 0.10	6.50 ± 0.19	10.45 ± 0.22	11.97 ± 0.56	**12.90** ± 0.23	3.42 ± 0.03	9.83 ± 0.25
opposite	8.26 ± 0.21	5.89 ± 0.31	6.00 ± 0.23	12.37 ± 0.23	13.26 ± 0.41	**14.53** ± 0.48	4.33 ± 0.04	14.38 ± 0.81
comparative	0.83 ± 6.25	0.83 ± 6.25	0.00 ± 0.00	0.83 ± 6.25	0.00 ± 0.00	0.00 ± 0.00	1.67 ± 11.10	**5.83** ± 14.57
superlative-(assoluto)	21.11 ± 0.38	16.59 ± 0.72	13.50 ± 0.25	**30.65** ± 1.99	25.88 ± 0.81	20.20 ± 0.64	4.42 ± 0.02	19.46 ± 0.49
present-participle-(gerundio)	47.65 ± 1.82	52.82 ± 1.15	50.86 ± 1.34	64.23 ± 1.46	65.83 ± 0.97	62.00 ± 0.61	41.90 ± 0.11	**66.59** ± 2.37
nationality-adjective	20.29 ± 0.12	37.44 ± 0.70	48.32 ± 0.19	69.61 ± 0.43	77.72 ± 0.33	**81.48** ± 0.30	4.72 ± 0.04	24.54 ± 0.37
past-tense	22.69 ± 0.32	28.74 ± 0.23	29.22 ± 0.49	**43.74** ± 0.53	38.96 ± 0.79	33.18 ± 0.74	17.35 ± 0.23	37.77 ± 1.21
plural	8.35 ± 0.11	12.92 ± 0.04	15.15 ± 0.32	26.02 ± 0.08	29.71 ± 0.55	**34.82** ± 0.99	21.79 ± 0.27	23.02 ± 0.99
plural-verbs-(3rd-person)	53.77 ± 1.29	56.37 ± 1.27	59.71 ± 2.15	80.17 ± 1.06	**85.74** ± 0.33	84.65 ± 1.28	75.93 ± 0.51	84.27 ± 3.22
plural-verbs-(1st-person)	0.67 ± 0.04	0.78 ± 0.05	0.41 ± 0.06	2.09 ± 0.07	**2.76** ± 0.09	2.38 ± 0.19	0.79 ± 0.02	1.69 ± 0.21
remote-past-verbs-(1st-person)	1.32 ± 0.56	1.60 ± 0.63	1.32 ± 0.50	2.97 ± 0.44	2.86 ± 0.41	**3.41** ± 0.47	0.26 ± 0.10	1.43 ± 0.50
noun-masculine-feminine-singular	23.98 ± 0.34	30.06 ± 1.15	32.34 ± 1.06	50.87 ± 0.66	**54.24** ± 3.10	52.58 ± 1.05	31.40 ± 0.48	46.92 ± 1.77
noun-masculine-feminine-plural	6.14 ± 0.65	4.01 ± 0.72	5.58 ± 0.71	23.83 ± 2.62	**28.69** ± 1.07	25.85 ± 1.48	3.27 ± 0.25	12.75 ± 1.34
Total semantic	4.38 ± 0.02	5.11 ± 0.02	6.57 ± 0.03	28.06 ± 0.13	44.01 ± 0.07	**50.77** ± 0.17	2.33 ± 0.01	7.10 ± 0.07
Total syntactic	21.42 ± 0.08	26.06 ± 0.02	28.06 ± 0.07	42.77 ± 0.11	**44.98** ± 0.07	44.64 ± 0.12	21.38 ± 0.01	33.41 ± 0.18
Total Accuracy	13.52 ± 0.03	16.35 ± 0.02	18.10 ± 0.02	35.95 ± 0.06	44.53 ± 0.02	**47.48** ± 0.09	12.40 ± 0.01	21.21 ± 0.04

Table 9: Results from the different models trained on Italian Wikipedia in the Word Analogy task.