Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Silent negatives: when partial annotation becomes supervision. A label integrity audit of the VU Amsterdam Metaphor Corpus benchmark

Gergely Tóth

toth.gergely@kre.hu

Károli Gáspár University of the Reformed Church in Hungary

Gábor Péli

Károli Gáspár University of the Reformed Church in Hungary

Miklós Gyorgyovich

Károli Gáspár University of the Reformed Church in Hungary

Bálint Duráczky

Károli Gáspár University of the Reformed Church in Hungary

Research Article

Keywords: metaphor detection, VU Amsterdam Metaphor Corpus, label noise, partial annotation, benchmark integrity

Posted Date: November 26th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10930300/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Silent negatives: when partial annotation becomes supervision

A label integrity audit of the VU Amsterdam Metaphor Corpus benchmark

Gergely Tóth[1,2] · Gábor Péli[1,2] · Miklós Gyorgyovich[1,2] · Bálint Duráczky[1,2]

[1] *Institute of Social and Communication Sciences, Faculty of Humanities and Social Sciences, Károli Gáspár University of the Reformed Church in Hungary, Budapest, Hungary*

[2] *Big Data and Social Analytics Research Group, Károli Gáspár University of the Reformed Church in Hungary, Budapest, Hungary*

Corresponding author: Gergely Tóth, toth.gergely@kre.hu

Gergely Tóth: ORCID 0000-0001-8040-1899, toth.gergely@kre.hu

Gábor Péli: ORCID 0000-0003-1839-6915, peli.gabor.lajos@kre.hu

Miklós Gyorgyovich: ORCID 0000-0002-2955-9647, gyorgyovich.miklos@kre.hu

Bálint Duráczky: ORCID 0009-0003-7991-7638, duraczky.balint@kre.hu

Abstract The VU Amsterdam Metaphor Corpus (VUAMC) underlies VUA-18/VUA-20, the most widely used metaphor-detection benchmark. We show that one of its 117 fragments – b1g, an academic text on geographic information systems – was annotated only in part: 3,006 lexical units according to the corpus creators' Appendix A, whereas the distributed TEI XML contains the full 20,170-word section. Because the encoding cannot distinguish the absence of annotation from a negative judgment, the benchmark conversion turned some 17,148 unannotated lexical units into negative gold labels. In benchmark units, this is 9,287 target tokens, 12.8% of the VUA-18 all_pos training set, carrying 3 positive labels, with a lemma prior predicting about 1,635 ($z = -53.66$). Eight independent layers of evidence point to this one section, among them the creators' documentation, a second annotation layer, the published academic metaphor rate, and the shared task's own unexplained register anomaly. None of the eight relies on a model or on training. Once this supervision is removed, a bert-base token classifier rises from 0.7070 to 0.7200 F1 (+0.0131, 95% CI [+0.0082, +0.0194]); removing an equal number of randomly selected tokens brings no such gain (−0.0030). The Verbs track replicates the result (+0.0207), and cross-validation over the corpus finds gains in all four registers rather than in the academic one alone (+0.0081). The test set is left intact, so every published test result still stands. We release the corrected split and argue that wherever a format cannot tell absence from negativity, partial annotation can silently become supervision

Keywords: metaphor detection, VU Amsterdam Metaphor Corpus, label noise, partial annotation, benchmark integrity

1 Introduction

In metaphor detection, the de facto standard is the VUA-18/VUA-20 benchmark, built on the VU Amsterdam Metaphor Corpus (VUAMC; Steen et al., 2010). The corpus consists of 117 fragments drawn from the British National Corpus, across four registers: academic text, news, fiction, and spoken language. When the shared task was organised, roughly 23% of the texts in each register were set aside for testing, and the remainder was kept for training (Leong et al., 2018). Both the 2018 and the 2020 competitions (Leong et al., 2018, 2020) are measured on this split, and the systems published since then – among them DeepMet with an all_pos result of 0.769, Go Figure! with 0.734, and illiniMet with 0.730 (Leong et al., 2020) – are comparable because they share the same test set. This is why the label integrity of the resource is no minor detail. The differences measured on it decide which architectures the field counts as progress, and those differences are often settled at the second decimal place of the F1 score.

The error we report did not originate from the annotators. A format conversion causes it, and the underlying logic can be explained in only a few sentences. Consider how many annotation formats record a positive judgment: by presence, and by presence only. Where the annotator judged a unit positive, the file carries a dedicated element or attribute; where no positive judgment was made, the file carries nothing. Two situations that could hardly differ more – a unit that was examined and judged negative, and a unit that was never examined at all – thus leave the same trace: an empty place. Inside the corpus, this ambiguity does no damage, since the extent of the coded region is fixed by the accompanying documentation rather than by the markup. The damage comes later, at conversion, when a procedure builds the machine-learning dataset from the corpus without opening that documentation. The unexamined section then flows into the training data as though it had been judged, negative label after negative label, and partial annotation quietly starts to act as supervision, the substitution leaving no trace anywhere. This mechanism is the one named in the paper's title.

The document in question is b1g-fragment02, the VUAMC academic text on geographic information systems. Its uncoded section – the part that the corpus creators never annotated – accounts for a substantial share of the document: 12.8% of the VUA-18 all_pos training set. This is 9,287 target tokens, of which exactly 3 carry a positive label, where a model-free lemma prior would lead one to expect around 1,635. The Verbs track shows the same picture on a smaller scale: the section covers 1,459 tokens and contains no positives at all. The phenomenon in question is therefore not noise scattered across the corpus but a single contiguous, identifiable span of text.

The novelty of the finding must be clearly distinguished from previous work. We are not concerned with inconsistent judgments within the annotated material (Zayed et al.,2020), nor with systematic biases in the annotation, as in Neidlein et al. (2020) on function words or Boisson et al. (2023) on constructional artefacts. The present case is of a different kind: an entire section was never annotated, and yet, because of the way labels are assigned, it still received negative labels. Li et al. (2023) note that the VUA's literal annotations are incomplete, but they do not identify the source of this incompleteness. We name, measure, and correct that source.

We also state at the outset what the finding does not claim, since it could otherwise be read as an accusation against the corpus or its annotators. The test set is unaffected: b1g lies entirely on the training side, so no published test result loses its validity, and every comparison between systems holds as before. Nor does the MIPVU annotation itself give any cause for concern. Where b1g does carry coding, that coding is perfectly calibrated once the register is taken into account; in both task tracks, the coded part is in fact somewhat richer in metaphor than the academic average. The error lies in the conversion step, not in the corpus and not in the work of the annotators. We maintain this framing throughout, and this is what makes the finding generalisable: a conversion error can recur anywhere the same conversion is applied – regardless of who the annotators were.

The study makes four contributions. The first is the finding itself, which we support with eight independent lines of evidence, none of which uses a model or training. Therefore, the conclusion does not depend on the experiments reported in Section 5. The second contribution asks what the error actually costs. The experiment designed to answer this question rests on a single comparison. If we remove the erroneous supervision, the BERT-based token classifier improves by +0.0131 F1. If we remove the same number of valid tokens, there is no measurable change. Taken together, the two runs leave only one interpretation: the improvement comes from the removal of erroneous labels, not from the reduction in the amount of training data. Third, we explain the shared task's own published but previously unexplained register anomaly, which is present in both task tracks. Fourth, we publish the corrected split and formulate the audit recipe as an algorithm for searching for the same pattern in other TEI- and standoff-based resources.

2 Related work

2.1 Label error as a benchmark problem

Northcutt et al. (2021) went looking for label errors in the test sets of 10 widely used benchmarks and found them at an average rate of at least 3.3%. The interesting part is not the errors themselves but what they do. On the original test sets, the models' relative rankings held up well. The errors affect the benchmark's role as a reference point. A larger model reproduces the systematic label errors more faithfully than a smaller one does, so once the proportion of originally mislabelled examples grows large enough, the ranking of the two models can reverse. The ImageNet results make this concrete: with corrected labels, once the share of such examples rises by six percentage points, ResNet-18 overtakes ResNet-50. The study became a turning point, since until it appeared, the field had regarded its benchmarks as error-free references.

The present finding differs from this work in three respects.

- First, it concerns an error in the training set, not the test set. The implication thus touches not the validity of the measurement but the performance that the measured systems can reach.
- Second, the error rate is of another order of magnitude. Northcutt et al. (2021) report an average of 3.3% across benchmarks. In contrast, our single document stands at 12.8%. Since the two values were computed on different bases, we offer the comparison at the order-of-magnitude level rather than directly. Even so, the gap is hard to overlook.
- Third, and most important: this error is not random noise but systematic bias, concentrated in one contiguous stretch of text. The two kinds of error do very different damage. Random noise averages out over a large sample, so its main cost lies in learning efficiency. By contrast, systematic bias shifts the model's decision threshold. Precisely this movement is what our experiment makes measurable.

For a formal treatment of the mechanism, one must turn to the multi-label image classification literature. Kim et al. (2023) examine the "unobserved labels assumed negative" setting (partially annotated data in which never-observed labels are quietly treated as negatives during training). They show that this assumption not only lowers accuracy but also biases the model's explanations. We demonstrate the same mechanism in the conversion of an NLP benchmark for the first time, to our knowledge. This provides the principal theoretical basis of the present study. We identify, in a new setting, a phenomenon already known and formally analysed, which had not previously been sought there and remained unnoticed because of the properties of the annotation format.

Our results fit into a well-known line of research. Nahum et al. (2025) show two things. Correcting label errors in the evaluation data shifts the reported performance upward and reorders the model rankings – their DeBERTa-v3 falls from second place to second-to-last. Filtering flagged errors from the training data, in turn, significantly improves the model's performance on the test set. The +0.0131 observed in our case is the token-classification counterpart of that second result. A difference is that the errors they consider must be detected statistically.

In contrast, the errors reported here trace back to a single document, so correcting them takes only a targeted removal. Merdjanovska et al. (2024) show that the real noise significantly more often distorts results than simulated noise. The point matters here, since our error has precisely the structure that simulated-noise models never produce: label flips that are neither independent of one another nor identically distributed across the corpus. What we have instead is one coherent stretch of text, silenced from its first sentence to its last. The closest methodological relative of our check is Merdjanovska and

Akbik (2025), who develop token-level indicators to detect mistaken "golden tags". Our check serves the same purpose without relying on any model.

Furthermore, Song et al. (2023) review noise-tolerant learning and describe a strategy that complements ours: these methods train models to perform despite errors. We take the opposite route and trace the error to its source, then remove it. Where an error is scattered beyond any hope of localisation, tolerance remains the only option; where it can be traced to a single identifiable cause, as here, removing it is better than tolerating it. Our case also needs separate treatment for a second reason. An error of this kind slips past the noise-oriented methods, since a text segment labelled negative from its first token to its last does not look like noise to them; it looks like a region with few positives and perfect internal coherence – in other words, like data.

2.2 Metaphor-specific antecedents

The quality of the VUA benchmark has been examined from several angles. Neidlein et al. (2020) analyse the bias caused by function words and show that language model performance largely depends on the handling of frequent, schematic elements. Zayed et al. (2020) uncover inconsistencies among annotated tokens in relation-level transformations of word-level datasets. Boisson et al. (2023) demonstrate constructional artefacts in metaphor datasets, that is, surface patterns that correlate with the label but do not follow from metaphoricity. Li et al. (2023) improve detection by explicitly modelling the basic meaning, and also note that the VUA literal annotations are incomplete.

The works listed above all examine the quality of the annotated material, asking how reliable or biased the annotators' judgments are. The present investigation concerns what happens to the unannotated material: text about which annotators never made a judgment becomes training material. In machine learning, supervision consists of the sequence of labels from which the model learns, so any defect in that sequence necessarily affects the model. If a passage that was never examined receives zero labels, the model learns that the passage contains no metaphor, even though it was never inspected. Silence, therefore, does not remain neutral: within the parameters the model learns, it begins to function as a judgment.

The difference is not merely one of nuance. If the annotators' markings are erroneous, the remedy is either re-annotation or a training procedure that tolerates label noise. Both address the annotation or the model, but if the unannotated text has been included as training material, there is nothing wrong with either. The error is introduced by the program that generates the benchmark label files from the corpus XML. We refer to this step as benchmark conversion (see Section 3.2). The program writes a zero for every token at which it finds no metaphor annotation, and thereby fails to distinguish the case "the unit was examined and is not a metaphor" from the case "the unit was never examined." The correction, therefore, occurs at the conversion step. This step must be given the boundaries of the annotated region, and where those boundaries are not available, it must write "unknown" rather than zero.

2.3 Primary sources

Several of the sources on which the argument rests are primary documentation. The appendix "Overview of annotated files from BNC-Baby", which closes the MIPVU volume (Steen et al., 2010, pp. 223–226), is an inventory. For every BNC fragment in the corpus, it names the section that was selected and states the number of lexical units that were actually coded in it. Table A1 of Herrmann's (2013) doctoral dissertation reports the same coded extent for the b1g fragment. Herrmann is a co-author of the MIPVU volume, and the first layer of evidence rests on two independent sources.

Krennmayr and Steen (2017) describe the TEI structure of the corpus, from which the presence-based logic of the annotation can be inferred. The shared task reports (Leong et al., 2018, 2020) provide the

split, the official baselines, and the basis for the eighth evidence layer. The primary data sources are VUAMC.xml (Oxford Text Archive 2541) and the v1.0 gold files of the EducationalTestingService/metaphor repository.

3 From VUAMC to VUA-18: how absence becomes zero

3.1 The MIPVU procedure and the structure of the TEI encoding

The VUAMC was developed using the MIPVU procedure (Steen et al., 2010), which is an extension of the Pragglejaz Group's MIP procedure. The basic idea of MIP is that a lexical unit is metaphorical if its contextual meaning differs from its basic meaning, and the two meanings can be understood through comparison. MIPVU extends this in three respects: it handles metaphor realised through reference, assigns a separate category to elements that signal the presence of metaphor, and provides explicit coding options for difficult cases where the annotator group did not reach consensus. According to the volume, inter-annotator agreement was $\kappa > 0.8$. The resulting tagset is rich and hierarchical, but the benchmark retains only its top-level distinction, metaphor-related word or not: every content word that carries the function="mrw" annotation counts as metaphorical, and every other content word counts as non-metaphorical.

The technical implementation of the annotation is decisive for what follows. The metaphorical unit is identified by the <seg function="mrw"> element. Non-metaphorical units have no annotation of their own, they simply do not receive a seg element. Consequently, the two states "we examined it, and it is not a metaphor" and "we did not examine it" have the same surface form: in both, the token carries no seg element. The format contains no markup indicating the boundaries of the coded span; that information is available only in the accompanying documentation, in the appendix of the MIPVU volume.

3.2 Benchmark conversion

The gold files of the shared task are lists of token identifiers paired with labels. The identifier has the form "{text}_{sentence_number}_{word_index}", and, according to the organisers' documentation, the index is obtained by splitting the sentence on whitespace, counting from 1 and including punctuation. Every target token receives the label 0 by default; a token receives the label 1 only where an mrw annotation is present.

The conversion is insensitive to whether a given passage was coded in the first place. On a fully annotated fragment, this rule raises no difficulty. On a partially annotated fragment, however, every target token in the uncoded passage receives a negative gold label. The error is therefore not a programming mistake but the consequence of a tacit assumption: the conversion takes it for granted that whatever the corpus contains has also been annotated. For 116 of the 117 fragments, this assumption does in fact hold.

Our own pipeline independently reconstructs this conversion. We compared the labels it produces against the ETS gold files. Across 117,920 tokens, the agreement is 1.000000, meaning that the reconstruction is bit-identical to the official benchmark. This had to be established first, because everything that follows depends on it: without an exact reconstruction of the split, the figures reported below could not be compared with the published baselines.

3.3 The case of b1g-fragment02

According to the appendix of the MIPVU volume, section 2 of the BNC B1G file was selected, and 3,006 of its lexical units were coded. The distributed VUAMC.xml, however, contains the entire 20,170-

word section. Of the 117 files, this is the only one in which such a discrepancy occurs; in the remaining files, the text contained in the file and the coded span essentially coincide. The consequence is approximately 17,148 uncoded lexical units that nevertheless received negative gold labels in the conversion.

The coded part falls into two contiguous sentence spans (Fig. 1). The first span runs from sentence 738 to sentence 765 and contains 724 words with 206 mrw annotations (28.5%); the second runs from sentence 1485 to sentence 1584 and contains 2,295 words with 428 annotations (18.6%). Together, the two spans contain 3,019 words and 634 annotations, that is, a metaphor proportion of 21.0%. Compared with the 3,006 lexical units given in the volume, the word count is 0.4% higher. A difference of this size is to be expected because words and MIPVU lexical units are not the same unit of counting. For the other 15 academic fragments, the ratio between the two counts ranges from 0.999 to 1.009, with a median of 1.004; for the coded spans of b1g, the ratio is 1.004, the median. The span boundaries are therefore not set heuristically; they are fitted to a figure taken from a primary source.

These numbers must be projected onto the benchmark, which counts differently from MIPVU: it counts only content words as target tokens and excludes the forms of the verbs be, do, and have (Leong et al., 2020). In this count, the uncoded section comprises 9,287 target tokens in the all_pos training set, that is, 12.8% of the 72,611 tokens in that set, and these carry a total of 3 positive labels. In the Verbs track, the same passage contains 1,459 target tokens, with no positive label among them. The all_pos track contains 10,925 labelled tokens from the whole b1g document, of which 1,638 lie within the coded spans. In the Verbs track, the corresponding number is 1,717 labelled tokens, of which 258 lie within the coded spans. None of this affects the test set, since b1g sits entirely on the training side. The test set consists of 22,196 tokens across 27 documents, with a prevalence of 0.1794.

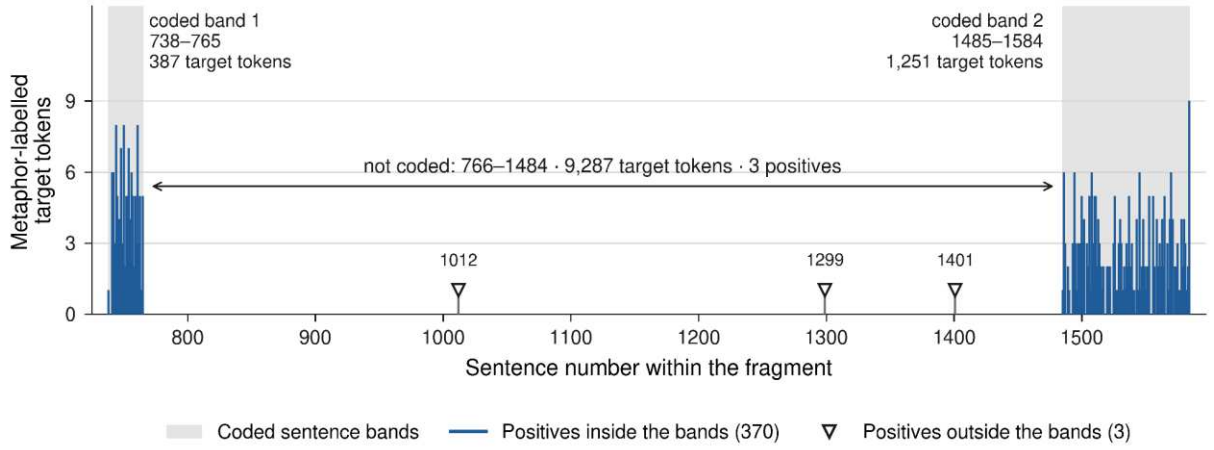


Fig. 1 The distribution of annotation in b1g-fragment02: how many target tokens per sentence received a metaphor label in the benchmark gold files. The two coded sentence spans (738–765 and 1485–1584) are highlighted. The annotation does not show a thinning frequency but rather a sharp boundary: within the spans, there are 370 positive target tokens; outside them, a total of 3, across the 9,287 target tokens between sentences 766 and 1484. The 634 mrw elements cited in Section 4.1 are a broader set, since they are counted on the TEI source, which also contains lexical units that are not target tokens.

4 Evidence: eight independent layers

Eight mutually independent layers of evidence support the finding. Three rely on the published corpus description, two on the XML itself, one uses all the remaining documents as a control, one relies exclusively on the gold labels, and one relies on the printed data table of the shared task. None of them uses a model or any training procedure, so the finding does not depend on the transformer experiments we report in Section 5. Table 1 summarises the structure.

Table 1 The eight evidence layers. None of the layers uses a model or training.

#	Layer	What it shows	Key figure	What it relies on
1	The volume's appendix	The creators themselves report the coded extent; Table A1 of Herrmann (2013) is an independent second source	For b1g, 3,006 lexical units are coded out of the full 20,170	published corpus description + the distributed XML
2	XML distribution	The annotation clusters into two contiguous sentence spans; a sharp boundary, not a thinning frequency	634 mrw elements fall within the delimited spans, and only 3 outside them	the distributed XML
3	corresp attribute	a second annotation layer, independent of the metaphor annotation, traces the same boundary	13 of 16 pairs within the spans	the distributed XML
4	Register statistics	counting the uncoded part, the academic register is larger than the published one	+17,238 lexical units, of which 17,148 are the uncoded part of b1g	published corpus description
5	Metaphor rate	exclusion, not inclusion, restores the published statistic	18.46% vs. published 18.5%	published corpus description
6	Corpus audit	all 117 documents by the same procedure: how many show the "two spans and a sharp boundary" pattern	1 (b1g); pattern indicator 68.9%, affectedness 85.0% at the token level	The other 116 documents are the control
7	Lemma-prior test	model-free, deterministic estimate of the expected number of positives	observed 3 vs. expected 1,634.9 → z = −53.66	the gold labels alone
8	Train/test prevalence gap	explains the shared task's own published and previously unexplained anomaly	academic train 14.41% vs. test 24.14%	the shared task's printed data table

4.1 The first six layers

I. The corpus creators' own statement. The appendix of the MIPVU volume gives 3,006 lexical units as the coded extent for the b1g fragment (p. 223). In Herrmann (2013), Table A1 reports the same. The distributed XML for the same fragment, however, contains 20,170 word elements (TEI w), more than six times the number that the creators report having coded. The two coded sentence spans contain 3,019 word elements, and the 719 sentences outside them contain 17,151 word elements. Of the latter, only three carry an mrw annotation, so that 17,148 lexical units are at once uncoded and unmarked.

We do not treat the 13-unit difference between 3,006 and 3,019 as evidence in itself. The two are not counted in the same unit: the appendix counts lexical units in the MIPVU sense, while our 3,019 counts TEI w elements. MIPVU treats a compound noun or a phrasal verb as a single lexical unit, and the corpus marks the second element of such a unit with a corresp attribute. Subtracting these continuations is therefore the rule that converts w elements into lexical units. The coded spans contain 13 such continuations. This is exactly where the 13-unit difference comes from, and the fragment contains 16 in all, so that the base of lexical units for the whole fragment is 20,170 − 16 = 20,154. We tested the same rule on the other 15 academic fragments, where it reproduces the appendix figures to within −2 and +21 units per file. The agreement should thus be read as accurate to roughly ±20 lexical units, not as exact. With that caveat in place, the affected share of the fragment comes to 85.1% of its lexical units (17,148 of 20,154) and 85.0% of its tokens (9,287 of 10,925) – the two measures, computed on different bases, land on practically the same figure.

II. The distribution of tags in the distributed XML. If the entire fragment had been coded, the mrw elements would be distributed evenly across the sentences, and their local density variation would follow the content structure of the text. Instead, 634 elements are concentrated in two contiguous sentence spans. Only 3 fall outside them (in sentences 1012, 1299, and 1401), each occurring in the middle of a stretch of several hundred unmarked words. It is the isolation of these three elements that matters for

the measurement: the pattern is not one of gradually declining frequency but of a sharp boundary. What we see is not “less frequent annotation” but “the end of annotation.”

III. A second annotation layer independent of metaphor annotation. In VUAMC, the corresp attribute couples two text locations, typically in an anaphoric or cross-reference relation. This annotation was applied by the annotators independently of the judgment of metaphoricity. In b1g, the attribute marks 16 such continuations, of which 13 fall within the coded spans. If the entire fragment had been coded, the continuations would be distributed in proportion to the size of the spans, so that approximately 15% of them, not 81%, would fall within the spans. Thus, an annotation layer independent of both our analysis and the metaphor annotation traces the same boundary.

IV. The register-statistical surplus. If the uncoded part of b1g is counted in the size of the academic register, then this register is 17,238 lexical units larger than the number the published corpus description reports. Of this surplus, 17,148 units are the uncoded part of b1g; the remaining 90 units are the residual of the counting method, of the same order as the +125, +86, and +554 measured in the three registers that contain no uncoded section. This figure quantifies the surplus: that much text would be present in the corpus without annotation and would nonetheless be counted in the register statistics. This layer alone would suffice to signal the problem, since it reveals a contradiction between the corpus description and the distributed file.

V. Reproduction of the published metaphor proportion. This test not only shows a discrepancy, but it also restores agreement. The corpus description gives the academic register a metaphor proportion of 18.5%. The same proportion can be computed in two ways. Leaving the uncoded part of b1g out of the calculation, our measurement yields 18.46%, an almost exact match to the published value; keeping the uncoded part, the proportion drops perceptibly below it. The published statistic is thus recovered by exclusion, not by inclusion. This is a strong indication that when the corpus creators computed their statistics, the uncoded portion was not part of the data they analysed.

VI. Corpus audit. We measured all 117 documents with the same procedure, examining how many show the “two spans and sharp boundary” pattern; the answer is a single document, b1g. The indicator is computed on the document’s TEI structure. We take every innermost section (div) of at least 100 word elements and ask what share of the words in those sections fall in sections that carry no mrw annotation at all. Of the 117 documents, 114 have at least one such section. b1g is the only document whose value is nonzero: 13,408 of its 19,468 words (68.9%) are in qualifying sections. This percentage must be kept clearly distinct from the affected share of the fragment, which is 85.0% by tokens (9,287 out of 10,925) and 85.1% by lexical units (17,148 out of 20,154). The two percentages answer different questions. The first is an indicator used to recognise the pattern, and it is comparable across documents. The second states the actual extent of the affected material. Taken together, they show that b1g is the only document in the corpus affected and that, within it, the affected portion is very large.

4.2 The seventh layer: the lemma-prior test

The seventh layer is the only piece of evidence that relies exclusively on the benchmark’s own gold labels and is completely model-free. The procedure we follow is deterministic. Leaving b1g out of the training set, we learn the metaphor proportion of every (lemma, part-of-speech) pair. This is a simple frequency estimate, without involving fitting or randomness. Each b1g token is then matched to its own (lemma, part-of-speech) pair. The proportion learned for that pair is taken as the probability that this particular token is a metaphor. Summing these per-token probabilities over the whole passage gives the number of metaphors the passage would be expected to contain. To ascertain the difference between the observed and expected values, we use a binomial z-statistic. One technical remark belongs here. The variance in this setting is, strictly speaking, Poisson-binomial, because the probability differs from token

to token. The binomial form slightly overstates the variance, making the test more conservative. At deviations as large as those measured here, the choice between the two statistics does not affect the conclusion.

The key to the test's validity is calibration. If the lemma prior were a poor estimator in general, it would produce large deviations on the genuinely annotated documents as well. To exclude this possibility, we ran the test on all 90 training documents, using the distribution of observed/expected ratios as a reference (Table 2).

Table 2 The lemma-prior test. The calibration row is decisive: the z-score of the weakest genuine document is −6.83, whereas that of the uncoded part of b1g is −53.66.

Part	n	Observed	Expected	Ratio	z
b1g coded (2 spans)	1,638	370	294.8	1.255	+5.91
b1g uncoded	9,287	3	1,634.9	0.002	−53.66
Weakest genuine document	297	9	45.6	0.198	−6.83
Median of the other 89 documents	–	–	–	1.037	–

The table supports two claims. First, the coded spans of b1g are perfectly calibrated and are even somewhat richer in metaphor than expected (ratio 1.255, z = +5.91). This part of the fragment is therefore properly annotated, refuting the objection that b1g, as a whole, was simply coded superficially. Second, the deviation of the uncoded part is not a matter of a significance threshold but of order of magnitude: the weakest genuine document reaches z = −6.83, whereas the uncoded part of b1g reaches z = −53.66. The median observed/expected ratio of the other 89 training documents is 1.037, the 5th–95th percentile band is [0.493, 1.590], and the ratio of 0.002 for the uncoded part of b1g is orders of magnitude away from this (Fig. 2).

The test was also run in two variants. In the first, register-matched variant, the prior was learned from academic documents only; its z = −55.92 rules out the objection that the deviation merely reflects idiosyncratic properties of the academic register. In the second, the test ran on the Verbs track, where the set of target tokens is narrower, and the prevalence is higher; there, it gave z = −30.60 and, with the academic prior, z = −35.45.

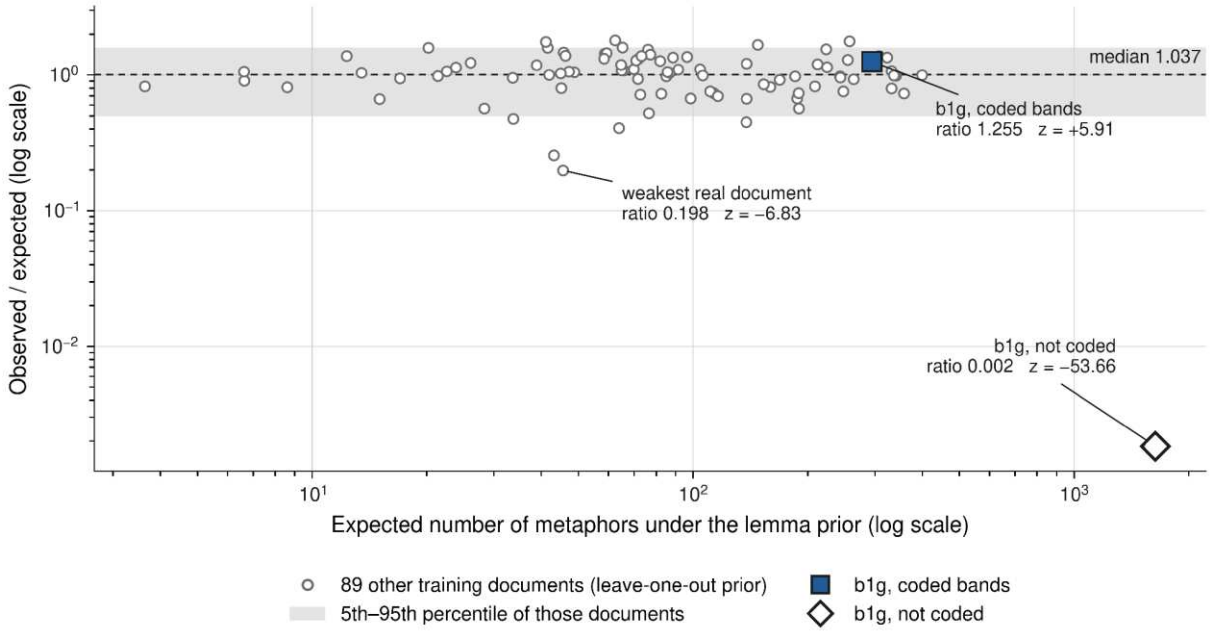


Fig. 2 Calibration of the lemma-prior test: observed/expected metaphor count ratio on the other 89 training documents, with leave-one-document-out prior, together with the two parts of b1g. Both axes are logarithmic. The grey band is the 5th–95th percentile of the 89 documents [0.493, 1.590]. The coded part of b1g lies within the reference band (1.255); its uncoded part is orders of magnitude outside (0.002), and it is even two orders of magnitude apart from the weakest real document (0.198).

4.3 The case of the three "whole" labels and the internal logic of the annotation

The three positive labels observed in the uncoded part require separate discussion, as they reveal how the annotation was produced. The passage contains five occurrences of the word form "whole" in total. All three adjectival occurrences (AJ0) received a label, whereas the two nominal occurrences (NN1, in the expression "as a whole") remained unlabelled. This distribution is not accidental. In the rest of the corpus, adjectival uses of "whole" are metaphorical in 68.8% of the cases. This is because, under MIPVU, its basic meaning is physical ("intact, in one piece"), so the abstract uses count as metaphorical by comparison with that basic meaning. The nominal expression "as a whole", by contrast, falls under a different judgment. The annotation, therefore, separated the adjectival from the nominal use of the same word form. This indicates that MIPVU makes a decision on a word by taking into account its context rather than by a mechanical search for the string.

The most plausible explanation of the pattern is a file-level consistency pass for a single lexeme: someone subsequently went through the file for one particular word form to make the judgments uniform and, in doing so, also added labels outside the coded spans. The same pattern appears in the corresp layer, where the three pairs outside the spans all involve the same phrase, "greenhouse effect". Two different types of annotation thus show the same manner of origin. This has a direct consequence for the lemma-prior test: the passage's true annotation count is zero, and the observed 3 are artefacts on the positive side. The z-value computed by the test underestimates the deviation. A more substantial aspect is what the passage reveals about the missing labels. Their absence does not record a judgment made by the annotators; it records where the coding work stopped –a distinction that the format is unable to express.

4.4 The eighth layer: the shared task's own prevalence anomaly

The eighth layer rests neither on the VUAMC XML and the MIPVU volume, nor on our measurement, but on the benchmark itself, on the data on which the field measures its systems. The strength of this

test is that it explains an anomaly reported and published by others that had remained unexplained. Hence, the symptom reported here was not discovered by us; it has been in print for eight years.

Table 2 in Leong et al. (2018) – which, according to its caption, shows "the number of text fragments from BNC, number of tokens and percentage of tokens marked as metaphor group by genres" – presents the number of texts, the number of tokens, and the proportion of metaphors by register, for both task tracks and on both sides of the split. In both the All POS and the Verbs track, the academic register is the only one in which the metaphor rates of the training and test sets differ substantially, but the report does not explain this outcome.

Table 3 presents the results of our re-measurement of the same split, rounded to two decimals, supplemented with rows for which there are no published equivalents. We compiled the register assignments independently; they match the shared task classification in all 117 documents.

Table 3 Metaphor proportion, token count, and text count by register on the two sides of the official VUA-18 split, in both task tracks. The source of the "ours" columns is the 94,807 target tokens of the ETS annotated set over the 117 documents; the register assignment comes from an independent reconstruction. The "published" columns are the values from Table 2 of Leong et al. (2018), unchanged; the rounding is that of the source, which reported metaphor proportions as whole percentages. The last three rows per track are the present paper's own addition. The gap is the difference between the unrounded proportions.

Register	Texts tr/te	n train	Train % ours	Train % published	n test	Test % ours	Test % published	Gap (pp)
All POS track								
academic	12 / 4	27,669	14.41	14	6,076	24.14	24	+9.73
conversation	18 / 6	11,994	9.78	10	5,302	10.26	10	+0.48
fiction	11 / 3	15,892	15.89	16	4,810	13.51	14	−2.38
news	49 / 14	17,056	19.69	20	6,008	21.99	22	+2.30
Total	90 / 27	72,611	15.21	–	22,196	17.94	–	+2.73
b1g-fragment02 (1 doc.)	1 / 0	10,925	3.41	–	–	–	–	–
academic without b1g	11 / 4	16,744	21.59	–	6,076	24.14	24	+2.55
academic without the uncoded part	12 / 4	18,382	21.68	–	6,076	24.14	24	+2.47
Verbs track								
academic	12 / 4	4,903	31.06	31	1,259	50.68	51	+19.61
conversation	18 / 6	4,181	14.97	15	2,001	14.54	15	−0.43
fiction	11 / 3	4,647	25.09	25	1,385	19.71	20	−5.38
news	49 / 14	3,509	42.12	42	1,228	45.52	46	+3.40
Total	90 / 27	17,240	27.80	–	5,873	29.98	–	+2.18
b1g-fragment02 (1 doc.)	1 / 0	1,717	7.05	–	–	–	–	–
academic without b1g	11 / 4	3,186	44.01	–	1,259	50.68	51	+6.67
academic without the uncoded part	12 / 4	3,444	44.22	–	1,259	50.68	51	+6.45

The first claim that can be drawn from the table is methodological. All 16 published percentages are identical to the rounded values of our measurement. All 16 token counts agree exactly: the four training-side rows sum to 72,611 and 17,240 tokens in the two tracks, the test-side rows sum to 22,196 and 5,873, and the number of texts is 90 + 27 = 117 in both tracks. The exact reproduction demonstrates that we are working on the official split, so the statements that follow apply to the reference dataset, not to a close reconstruction of it. A fourth source, independent of both the organisers and the corpus creators, confirms this: Stemle and Onysko (2018) report a label distribution of 11,044:61,567 in the training set. Our measurements yield exactly these numbers of positive and negative target tokens.

The second claim drawn from the table concerns the finding. In the All POS track, the academic train–test gap is +9.73 percentage points, while in the other three registers, it is at most 2.38 percentage points (computed from the unrounded proportions). If the uncoded part of b1g is removed, the prevalence of the academic training material rises from 14.41% to 21.68%, and the gap narrows to +2.47 pp. In other words, 74.7% of the gap is attributable to the uncoded part of a single document. Dropping b1g entirely yields 21.59% and a gap of +2.55 pp, that is, 73.8% of the gap. The metaphor rate of b1g as a whole is 3.41% (373 out of 10,925), compared with 21.59% for the other academic training documents. A genuine academic text does not contain six times fewer metaphors than the other academic texts; a partially annotated one, however, looks exactly like this.

The Verbs track repeats the same layer on a second set of target tokens, partially overlapping with the first. There, the contrast is even sharper. The academic gap is +19.61 pp, twice that of the All POS track, whereas it remains below 5.38 pp in the other three registers, and in fiction and spoken language, its value is negative. The uncoded part of b1g contributes 1,459 tokens with zero positives in the Verbs task. Excluding these raises the proportion of academic training from 31.06% to 44.22% and narrows the gap to +6.45 pp. Thus, 67.1% of the gap comes from a single document. The two tracks, therefore, yield the same (roughly two-thirds) order of magnitude across two sets of target tokens that overlap only partially within b1g (1,459 vs 9,287 tokens).

Finally, Table 3 also supports our interpretation of the finding: where b1g is coded, it is slightly richer in metaphor than the average across both tracks. The proportion in the coded part of the All POS track is 22.59%, compared with 21.59% in the other academic training documents (ratio 1.046), and that in the coded part of the Verbs track is 46.90%, compared with 44.01% (ratio 1.066). Both values are consistent with the z-value of +5.91 obtained in the lemma-prior test.

5 Consequence: a correction experiment

5.1 Design and pre-registered decision rule

The finding is in itself a claim about data quality. In benchmarking terms, two more things must be demonstrated: that the error has a measurable consequence, and that this consequence results from the incorrect labels rather than simply from 12.8% less training data. We therefore ran five conditions, each with three random seeds (42, 43, 44), with the same setting: bert-base-uncased token classification with a linear head, 3 epochs, batch size 32, learning rate 3e-5 with linear warm-up and decay, maximum length 192, decision threshold 0.5, bf16 mixed-precision computation. The label of target tokens is placed on the first subword, and the other subwords and non-target tokens are ignored in the loss computation.

We compared five conditions:

1. BASE is the uncorrected training set.

2. TRIM is the condition in which the uncoded part of b1g receives no supervision. This is the correct repair, since it removes only the incorrect labels and retains the 1,638 valid target tokens of b1g.

3. NOB1G is the coarse repair, in which b1g is omitted entirely. It shows whether retaining the valid part of the document makes any difference.

4. RAND is the size-matched control, which removes the same number of valid target tokens from other training documents.

5. TRIMW is TRIM on the same data, but with the positive class weight of BASE.

The design of the RAND control required particular care. Selecting individual tokens at random would break up the sentence structure. Moreover, the prevalence of the removed set might differ from the training set average, and this alone would degrade the model, thereby confounding the effects of "less data" and "altered class proportions." Therefore, the control setting removes entire sentences from other training documents and accepts sentences so that the prevalence of the removed set remains close to the training-set average. Because a sentence is an indivisible unit, the size of the removed set can be matched only up to the nearest sentence boundary (Verbs: 15,780 vs 15,781 training tokens; All POS: exact match). The procedure is deterministic with respect to the random seed. But the two conditions are not fully symmetric, and this warrants a note: TRIM keeps the uncoded tokens in place as context and withdraws only their supervision, whereas RAND removes entire sentences, including context. As the integrity check in Section 5.4 shows, the difference has no measurable effect.

The decision rule was fixed in advance, as documented by the commit hash of the code release. We denote by s the standard deviation of the F1 scores obtained with the three BASE seeds. The effect is considered real if TRIM − BASE > 2s while RAND − BASE remains substantially below that threshold. The rule takes this form because the seed-to-seed standard deviation, measured on the machine used, gives a natural estimate of the noise level: the 2s threshold therefore comes from the experiment itself rather than from an external convention. Confidence intervals for all differences are estimated by a document-level bootstrap, described in Section 5.3.

5.2 Results

Table 4 The five conditions. F1 is the average of the runs with the three random seeds, SD is the standard deviation of those same runs, and the labelling rate is the proportion of test tokens judged positive. pos_weight is the weight of the positive class in the loss function; it equals the negative/positive ratio of target tokens in the given training set, except for TRIMW, which deliberately keeps the BASE weight.

Condition	Train	Prevalence	pos_weight	F1 (mean)	SD	Labelling rate
BASE – uncorrected	72,611	0.1521	5.575	0.7070	0.0047	0.1987
TRIM – without the uncoded part	63,324	0.1744	4.735	0.7200	0.0034	0.2064
NOB1G – without the entire b1g	61,686	0.1730	4.781	0.7219	0.0051	0.1999
RAND – control: nearly the same number of random tokens, in whole sentences	63,324	0.1506	5.640	0.7040	0.0029	0.1994

Condition	Train	Prevalence	pos_weight	F1 (mean)	SD	Labelling rate
TRIMW – TRIM, but with the BASE weight	63,324	0.1744	5.575	0.7180	0.0034	0.2125

The standard deviation of the BASE seeds is s = 0.0047, hence 2s = 0.0094. The result is TRIM − BASE = +0.0131 (Table 4), corresponding to 2.78 s, and RAND − BASE = −0.0030, corresponding to −0.64 s. Thus, both pre-registered conditions are met. The direction of the effect is the same for all three random seeds: TRIM > BASE (+0.0164 / +0.0134 / +0.0094) and RAND < BASE (−0.0008 / −0.0039 / −0.0043). No run contradicts this pattern (Fig. 3).

The change in the labelling rate is also informative, provided that the reference point is the training set's prevalence. The false negatives reduce the positive rate in the training set to 0.1521, whereas the positive rate in the test set is 0.1794. Removing the uncoded part restores the training rate to 0.1744, effectively matching the test set, and the model's labelling rate rises from 0.1987 to 0.2064. The correction, therefore, does not simply make the model predict more positives. It removes the downward bias in the training distribution, and the model's decision behaviour shifts as a consequence.

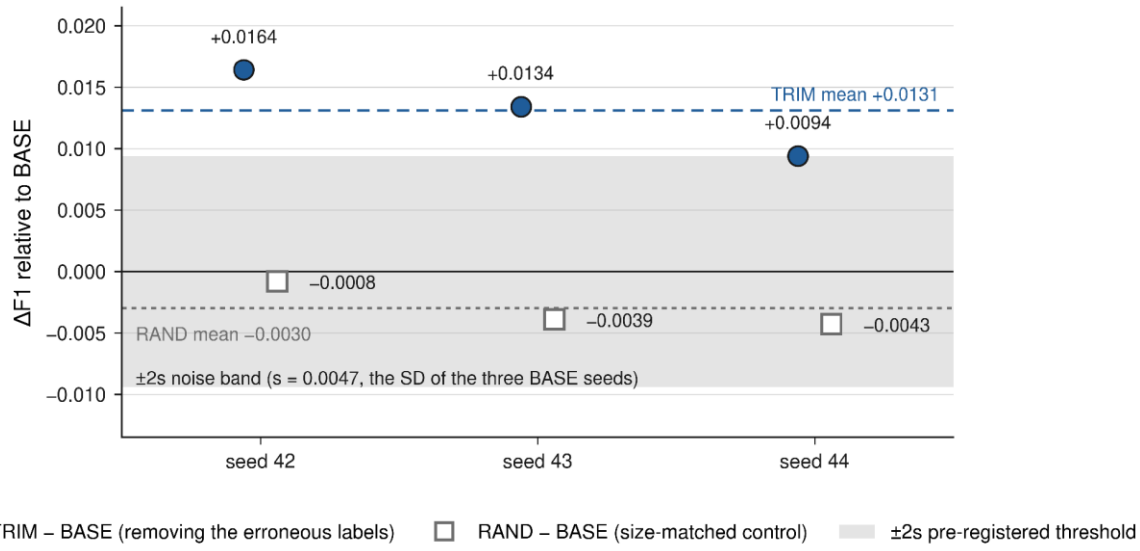


Fig. 3 The pre-registered decision rule by seed. The vertical axis shows ΔF1 relative to BASE; the grey band is the ±2s threshold (s = 0.0047, the F1 standard deviation of the three BASE seeds). All three TRIM points are outside the band, all three RAND points are inside the band, and no seed contradicts the direction.

5.3 Decomposition of the effect

Confidence intervals come from a percentile bootstrap with 4,000 resamples, and the unit we resample is the document, not the token – 27 test documents in all. Every interval refers to the difference. Within each resample, we compute F1 for every condition and seed, then average them to remove seed noise while preserving the uncertainty introduced by sampling the documents. Token-level intervals would be considerably narrower, which is why we do not work with them: their precision would be illusory. Tokens inside a document are interdependent – one author wrote them, often reaching for the same lexemes.

Table 5 Decomposition of the effect. Confidence intervals come from a document-level percentile bootstrap with 4,000 resamples.

Comparison	ΔF1	95% CI	What it measures
TRIM − RAND	+0.0161	[+0.0115, +0.0210]	purely the removal of the erroneous labels
TRIM − BASE	+0.0131	[+0.0082, +0.0194]	the full correction
NOB1G − BASE	+0.0149	[+0.0105, +0.0202]	discarding the entire b1g
RAND − BASE	−0.0030	[−0.0070, +0.0009]	mere data loss – zero
TRIM − TRIMW	+0.0020	[−0.0005, +0.0044]	weight recalibration – zero
TRIM − NOB1G	−0.0018	[−0.0052, +0.0014]	retaining the coded part – zero

The central row of Table 5 is TRIM − RAND = +0.0161. This part of the effect arises solely from the removal of incorrect labels, since RAND removes an equal number of valid tokens. The interval for RAND − BASE = −0.0030 contains zero, so the loss of data on its own has no detectable effect. The most obvious objection is thereby ruled out.

The interval for TRIM − TRIMW, +0.0020, also contains zero; consequently, recalibrating the positive class weight is not what produces the effect. The control is needed because in the corrected data, the prevalence rises from 0.1521 to 0.1744, and the class weight moves with it. TRIMW is trained on the same data with the old weight and ends up at essentially the same result. The comparison TRIM − NOB1G = −0.0018, in turn, shows that keeping the coded, valid part of b1g brings neither a measurable advantage nor a disadvantage. The finer correction is therefore defensible, and it is the one built into the corrected split we release.

5.4 Robustness checks

The first check ensures that the correction itself is implemented without errors. We implemented it twice, along two independent code paths. In one, the uncoded portion is physically deleted in a parallel working directory. In the other, every token stays in place as context and only its supervision is withheld, through a runtime exclusion list. Trained with the same seed, the two versions reached F1 scores of 0.7190 and 0.7180, and their predictions differed on just 42 of the 22,196 tokens (99.81% agreement). What remains of the difference comes from the non-determinism of GPU training rather than from the code, since both versions train on the same token sets and the same number of sentences.

The second check examines how a larger encoder performs on the corrected data. Under the TRIM condition, roberta-large reaches an F1 of 0.7062 (SD 0.0035), which is −0.0138 relative to bert-base under the same condition, CI [−0.0227, −0.0041]. The decline is significant and repeats the pattern already observed in the uncorrected data, where roberta-large likewise lagged behind bert-base (0.6944 vs 0.7070). Hence, the corpus error is not the reason why the larger encoder fails to help on this task. The cause of the decline is over-labelling: roberta-large assigns positive labels at a rate of 0.2341, bert-base at 0.2064, while the positive label prevalence in the gold data is 0.1794. The observation suggests that model capacity is not the limiting factor on this task.

The third check concerns the relation to the official baseline. On the uncorrected data, bert-base achieves 0.7070, which is 1.1 F1 points below the organisers' BERT baseline of 0.718 (Leong et al., 2020). On the corrected data, the same model with the same recipe yields 0.7200, which even slightly exceeds the official baseline without any change to the architecture. This observation provides a practical measure of the finding, since the difference produced by the data-side correction is of the same magnitude as the differences the literature commonly reports as modelling improvements.

5.5 Replication on the Verbs track

We then repeated the experiment on the VUA-18 Verbs track without making any changes. This is more than a robustness check. The Verbs data coincide with the original only partially. Hence, the run amounts to a confirmation on partly independent material: the Verbs target tokens overlap the All POS set only partially – 17,240 training and 5,873 test tokens, compared with 72,611 and 22,196 – and within the uncoded portion of b1g, there are 1,459 target tokens. There is no positive label among these 1,459, while there had been three in the All POS set. Everything else in the setting was fixed: the design, the random seeds (42, 43, 44), the number of epochs, the batch size, the threshold, and the pre-registered decision rule; the test set was left as is. Still, the two sets of runs used different hardware: All POS ran on NVIDIA GPUs, while Verbs on AMD Radeon/ROCm GPUs; any difference caused by this falls within the non-determinism range that the integrity check in Section 5.4 had already measured.

Table 6 The five conditions on the Verbs track. The columns have the same meaning as in Table 4.

Condition	Train	Prevalence	pos_weight	F1 (mean)	SD	Labelling rate
BASE – uncorrected	17,240	0.2780	2.597	0.7368	0.0029	0.2911
TRIM – without the uncoded part	15,781	0.3037	2.293	0.7575	0.0038	0.3000
NOB1G – without the entire b1g	15,523	0.3010	2.323	0.7555	0.0084	0.2956
RAND – control: nearly the same number of random tokens, in whole sentences	15,780	0.2779	2.598	0.7345	0.0049	0.2937
TRIMW – TRIM, but with the BASE weight	15,781	0.3037	2.597	0.7590	0.0007	0.3048

The standard deviation of the BASE seeds is s = 0.0029, so 2s = 0.0059. The result is TRIM − BASE = +0.0207 (Table 6), which is 7.03 s, and RAND − BASE = −0.0023, which is −0.78 s, meaning that both pre-specified conditions are met, and the direction is the same with all three seeds (+0.0138 / +0.0241 / +0.0242). The effect is nominally larger on the Verbs track than on the All POS track (+0.0207 vs +0.0131), but the two confidence intervals overlap, so no difference between the tracks can be established. What can be stated is that the Verbs track, that is, a set of target tokens that overlaps only partially with the first, replicates the result obtained on the All POS track.

Table 7 Decomposition of the effect on the Verbs track. Confidence intervals are from document-level percentile bootstrap with 4,000 replications (27 test documents).

Comparison	ΔF1	95% CI	What it measures
TRIM − RAND	+0.0230	[+0.0133, +0.0344]	purely the removal of the erroneous labels
TRIM − BASE	+0.0207	[+0.0112, +0.0307]	the full correction
NOB1G − BASE	+0.0186	[+0.0091, +0.0284]	discarding the entire b1g
RAND − BASE	−0.0023	[−0.0095, +0.0046]	mere data loss – zero
TRIM − TRIMW	−0.0015	[−0.0045, +0.0015]	weight recalibration – zero

Comparison	ΔF1	95% CI	What it measures
TRIM − NOB1G	+0.0021	[−0.0027, +0.0078]	retaining the coded part – zero

The controls (Table 7) give the same picture as on the All POS track. TRIM − RAND = +0.0230 is the amount that arises purely from removing erroneous labels, and the interval for RAND − BASE = −0.0023 contains zero, indicating that the mere loss of data has no detectable effect here either. TRIM − TRIMW = −0.0015 also contains zero: here too, the effect is not caused by the recalibration of the positive class weight. TRIM − NOB1G = +0.0021 shows that retaining the coded, valid part of b1g yields no measurable advantage or disadvantage either.

Changes in the labelling rate are even more pronounced in the Verbs track. Because of the false negatives, the positive rate on the training set is 0.2780, while it is 0.2998 on the test set. Removing the uncoded portion raises the training rate to 0.3037. The model's labelling rate changes from 0.2911 to 0.3000. The mechanism is therefore the same in both tracks: the uncoded segment skews the training distribution downward, and the model adjusts its decision behaviour to this skew. Removing the skew removes the effect.

5.6 Register localisation of the effect: cross-validation on the full corpus

All experiments so far have been performed on the official test set, which consists of 4 documents for the academic part and 3 documents for the fiction part. This is insufficient for concluding at the register level. We therefore ran document-level cross-validation on the whole corpus: a 10-fold partition, stratified by register and drawn with the same random seed, and run once on the uncorrected corpus and once on the corrected one. The comparison is paired, because the fold assignment was identical in the two runs. The evaluation was performed on the set of target tokens common to both corpora (85,520 tokens across 117 documents). The uncoded part of b1g is scored on neither side. Thus, any difference is due solely to the training data used by the models.

Table 8 The effect of the correction on the entire corpus, with document-level cross-validation. F1 values are out-of-fold scores calculated on the common evaluation set of the two corpora; confidence intervals are from a paired, document-level percentile bootstrap method with 4,000 replications.

Register	Doc.	n token	F1 uncorrected	F1 corrected	ΔF1	95% CI
Total	117	85,520	0.7060	0.7142	+0.0081	[+0.0031, +0.0140]
academic	16	24,458	0.7249	0.7352	+0.0103	[−0.0007, +0.0274]
conversation	24	17,296	0.6300	0.6413	+0.0113	[+0.0012, +0.0225]
fiction	14	20,702	0.6866	0.6898	+0.0032	[−0.0038, +0.0103]
news	63	23,064	0.7288	0.7344	+0.0056	[−0.0000, +0.0112]
Fold 7 – null control (b1g's fold)	10	7,233	0.7228	0.7229	+0.0001	–

On the full corpus, ΔF1 = +0.0081, CI [+0.0031, +0.0140] (Table 8), so zero lies outside the interval. All four registers move in a positive direction (academic +0.0103, conversation +0.0113, news +0.0056, fiction +0.0032), none of them stands out, and the per-register intervals largely overlap (Fig. 4). The prior expectation that the improvement would be localised in the academic register is therefore not

confirmed. The picture is consistent with the following interpretation: removing the erroneous labels does not add register-specific knowledge but removes a global bias in the training distribution via the same mechanism reflected in the labelling rates in Sections 5.2 and 5.5.

A built-in null control tests the validity of the design. b1g falls in fold 7, and when fold 7 is the test fold, the training data of the two corpora are identical, so their difference should be zero. The measured value is ΔF1 = +0.0001 (7,233 tokens, 10 documents). The token counts confirm this: fold 7 contains 16,520 tokens in the uncorrected corpus and 7,233 in the corrected one, and the difference gives exactly the 9,287 uncoded tokens of b1g.

The gain of +0.0081 F1 measured in cross-validation is smaller than the +0.0131 measured on the official test set. This is not a contradiction. In cross-validation, b1g is part of the test fold in one of the ten rounds, so in that round, the two corpora are trained on identical data. This 8.5% portion of the common evaluation set contributes nothing by construction. Moreover, the cross-validation training sets are larger than those in the official split (approximately 105 documents versus 90), and a given amount of incorrect supervision matters less the more sound data is available around it. The two values are not in competition with each other: the test value reports the effect as measured on the official benchmark. In contrast, the cross-validation value provides a more conservative estimate for the corpus as a whole.

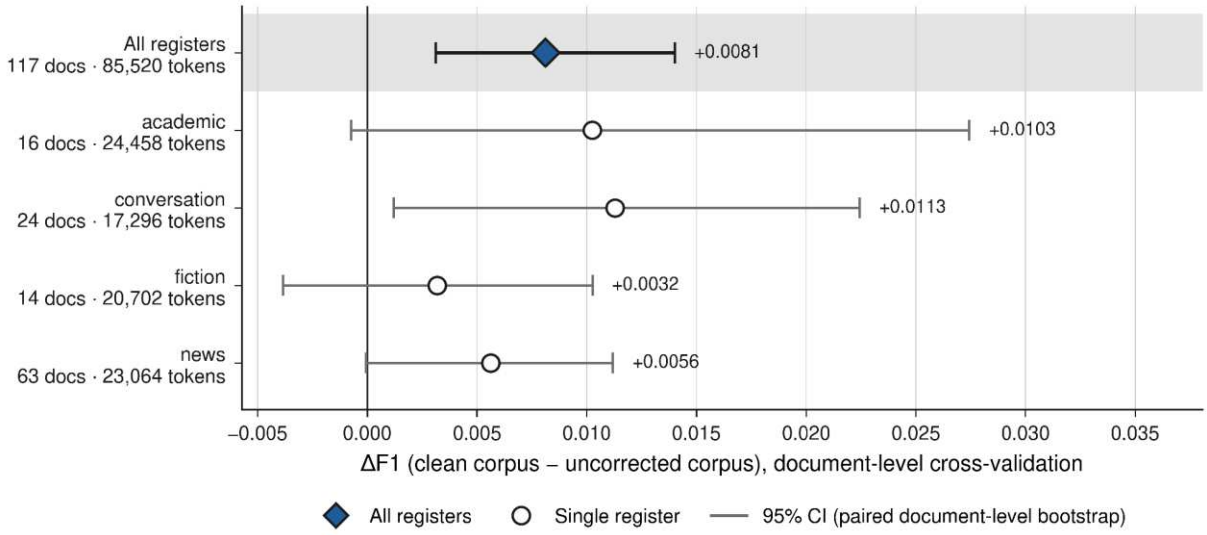


Fig. 4 Per-register effect on the entire corpus, with document-level cross-validation. The horizontal axis shows ΔF1 = corrected corpus − uncorrected corpus, measured on the two corpora's common evaluation set of 85,520 tokens. Error bars are 95% percentile confidence intervals from a paired, document-level bootstrap (4,000 resamples). The effect is positive in all four registers and is not localised to the academic register.

5.7 What the experiment does not settle

The question of generalisation across architectures remains open. We measured the effect on bert-base and corroborated it with roberta-large. We do not know how the two-branch architectures that top the published ranking (those that process the basic meaning and sentence context on separate branches) behave. With the release of the corrected split, such experiments can be carried out by anyone. We did not, however, rerun the full range of published systems; therefore, we make no claim about the published ranking.

We report the per-register breakdown measured on the official test as an exploratory observation, not as a result (All POS: academic +0.0073, news +0.0201, fiction +0.0106, conversation +0.0104, Verbs: academic +0.0220, news +0.0349, fiction +0.0055, conversation +0.0014). The 27 test documents are too few to support such a breakdown, and the figures for the two tracks do not point in the same direction.

The question of register localisation is therefore answered not by this breakdown but by the cross-validation in Section 5.6, which relies on 117 documents rather than 27.

6 Discussion

6.1 Why did all this go unnoticed

The symptom was published in Table 2 of Leong et al. (2018): the ten-percentage-point academic prevalence gap appears in the All POS track, and the twenty-point gap in the Verbs track. Since then, it has been visible to every researcher who has used the benchmark. The explanation was missing, as well as the question to which that explanation would have been the answer.

A deeper cause lies in how benchmarks are customarily used. The community treats the training dataset as a black box: the systems learn from it, attention is paid to the test results, and the checks for labelling errors are directed, as a rule, at the test set (Northcutt et al., 2021), where the validity of the measurement is directly at stake. Errors in the training set behave differently. They do not invalidate the measurement, but they reduce the performance that any measured system can achieve. They do so imperceptibly and to the same degree for all of them, and it is this uniformity that makes them invisible. When every participant learns from the same flawed data, the rankings show no visible distortion.

A further reason concerns format. The prevalence anomaly appears in the report as a per-register aggregate. Hence, tracing it back to a single document requires a document-level breakdown, which must be computed from public data. This step is not difficult, but it is not obvious either, since it is not part of the usual workflow for benchmark use.

6.2 An audit recipe for other TEI and standoff corpora

The methodological yield of our finding is a four-step procedure, applicable to other resources as well, which we release as a script.

I. The first step costs almost nothing: the documented extent of the coding is compared with the actual size of the distributed file. Wherever the documentation states such a figure, this is the cheapest and the strongest test. In the present case, a single comparison – 3,006 against 20,170 – would already have given grounds for suspicion.

II. The second step looks at how the tags are spread through the document. Working from the document's own TEI structure, we take every innermost section of at least 100 word elements and ask what share of the text lies in sections without annotation at all. Wherever such a section begins or ends, we then check whether the transition is abrupt or gradual. The 100-word threshold is a convention that offers comparability. The metric is computed in the same way for every document in which such sections exist (114 of the 117 documents here), and the distribution of its values over the documents that are known to be annotated throughout supplies the standard against which any other document is judged.

III. The third step brings in a control: a second, independent annotation layer, wherever the corpus provides one. A layer of this kind traces the boundary of the coded region on its own, untouched by the primary annotation.

IV. The fourth step is a prior test on the gold labels that involves no model: the gap between the lemma-based expectation and the observed count gives an order-of-magnitude signal. This requires neither training nor any random component.

The strength of the procedure lies in its ordering. The steps become progressively more expensive. The first step requires only the documentation; the second, the raw corpus; the third, the richer annotation of

the corpus; and the fourth, the converted gold labels. Therefore, it is economical to begin with the first step and to proceed further only if that step has found something.

6.3 Generalising the mechanism and a format-level proposal

This mechanism is not specific to any single text corpus. It arises in any process in which positive annotations exist solely as token labels, with no record of which parts of the text the annotator examined. This description applies to most TEI-based annotation schemes and many standoff formats. Kim et al. (2023) formalise exactly this mechanism. In the conversion, a label that was never assigned becomes a negative label that appears to have been assigned, and the converted data end up carrying information that the original never contained. The risk increases when the corpus and the reference dataset derived from it are maintained by different groups: those who know which segments were annotated are not the ones who wrote the conversion script, so the only knowledge that could prevent the error may never reach the point at which it occurs.

This leads to our normative recommendation, which we state as a requirement. Any conversion step that produces a machine-learning dataset should require asking for the boundaries of the annotated segment. Where those boundaries cannot be supplied, the token should be marked unknown rather than negative. Training should exclude unknown tokens from the loss, which is exactly what the TRIM condition does in the present experiment. Since the cost of this change is small, it can be applied retrospectively to existing resources to close off the class of error documented here. One further practice to adopt is ensuring that a per-document prevalence summary accompanies every conversion output. With such a summary in place, the present anomaly could not have been overlooked.

6.4 What the finding does not say

Certain things remain exactly as they were. It is worth stating them one by one, so that no more is read into the results than they support. The test set remains unchanged. Not a single published test result loses its validity, and the comparisons between systems remain as they were before. Nor is the quality of the MIPVU annotation questioned. Where b1g actually carries coding, it is perfectly calibrated once register is taken into account: it stands slightly above the academic average in both task tracks (the ratios are 1.046 and 1.066, respectively), and the lemma-prior test yields $z = +5.91$ for that part. The error occurs in a single place: the conversion step.

What the finding implies for the ranking of systems, however, must be stated with caution. An effect of +0.013 is comparable in magnitude to the differences that the literature attributes to architectural improvements, and, on the corrected data, a plain bert-base reaches the organisers' BERT baseline. This does not imply that the published ranking of systems is incorrect, since establishing a revised ranking would require rerunning all systems on the corrected data, which is beyond the scope of this paper. What follows is that a data-side factor of the same magnitude has not yet been controlled for. Future comparisons should therefore be conducted on the corrected split.

7 Limitations

The finding concerns a single document out of 117, and according to the corpus audit, it is indeed a unique case: the mechanism can be generalised, but not its magnitude. In other corpora, the same mechanism may operate on a smaller or a larger scale.

The correction experiment shows an effect for one architecture (bert-base), corroborated by one large encoder (roberta-large). Therefore, the result cannot be extended to the full range of systems. In particular, we do not know how the correction affects those two-branch architectures that process basic

meaning and sentence context on separate branches, and that stand at the top of the published ranking. The experiment was repeated on the Verbs track (Section 5.5), again on bert-base only.

Whether the effect is localised to a particular register cannot be decided on the official test set, since its academic portion consists of only 4 documents and its fiction portion of only 3. We therefore report the per-register figures measured on the test set as exploratory observations rather than as results, and we answer the question by cross-validation on the full corpus (Section 5.6). The per-register intervals obtained there, however, remain wide, especially in the academic and fiction registers (16 and 14 documents, respectively), so the only findings we report are the common direction of movement across the four registers and the aggregate effect. We do not report a ranking of the registers.

At the corpus level, a residual of 855 lexical units (0.46%) remains between the values computed from the XML and those reported in the appendix of the MIPVU volume. The direction of the discrepancy can be explained by the difference between tokenisation-based counting and counting by lexical units, but we were unable to reduce it to zero. A related limitation is that the match of the coded extent is not exact but holds only to within approximately ±20 lexical units.

The correction experiment was run on GPUs, where bit-identical reproduction cannot be expected. The integrity check in Section 5.4 gives a measure of this: two independent implementations of the same correction, trained with the same random seed, produced predictions that differed on 42 of the 22,196 test tokens. The All POS and the Verbs experiments were run on GPUs from different manufacturers, so what should be reproduced from the reported results is the sign and the magnitude of the effect, not the last decimal place. The provenance files associated with the runs record the hardware, library versions, and input fingerprints, so that the source of discrepancies can be traced.

8 Conclusion

Of the VUA-18 training set, 12.8% is text that the corpus creators never annotated, yet the benchmark conversion assigned negative gold labels to every target token in this text. That the latter had taken place is supported by eight independent layers of evidence, all converging on the same section: among them, the creators' own documentation, a second annotation layer, a model-free prior test, and the anomaly that the shared task itself had published. None of these eight depends on a model or on training.

The error also causes measurable damage. Correcting the error raises F1 by +0.0131, while removing an equal number of valid tokens changes nothing, allowing the gain to be attributed solely to the removal of the incorrect labels. On the corrected data, a plain bert-base model matches the organisers' official BERT baseline without any architectural changes. The experiment behaves the same way on the Verbs track, whose target tokens overlap with the original set only partially (+0.0207). A document-level cross-validation over the whole corpus finds a positive effect in each of the four registers (+0.0081), a result that a null control further supports. We did not touch the test set: the published results stand unchanged.

The lesson generalises to any format in which the absence of annotation and a negative judgment are indistinguishable, because in such formats, partial annotation silently becomes supervision.

A conversion step should therefore be supplied with the boundaries of the annotated span. Where those boundaries are unavailable, it should mark the token as "unknown" rather than assign it a zero, as is standard practice today. We release the corrected split, the audit recipe, and the full code of the experiments.

Statements and Declarations

Funding

The research group is supported by the internal research grant RH-TUD/2608/2024 of Károli Gáspár University of the Reformed Church in Hungary ("Social scientific analysis of sermons by means of traditional qualitative and advanced (NLP, AI) text-analytic tools"). Principal investigator: Gergely Tóth, associate professor.

Competing interests

The authors have no financial or non-financial interests directly or indirectly related to the work.

Ethics approval

The research involved no studies on humans or animals, so ethical approval is not applicable. Before submission, we notified the maintainers of the affected resources: on 1 September 2026, we opened issue no. 8 in the EducationalTestingService/metaphor repository (https://github.com/EducationalTestingService/metaphor/issues/8), and on the same day, we sent an information letter to the VU Amsterdam Metaphor Lab, the group that maintains the VUAMC. The corrected split and the audit materials were published on the same date under DOI 10.5281/zenodo.22216960.

Data availability

The corrected split is deposited in a public repository under a persistent identifier; the deposit contains the list of token identifiers to be excluded, the boundaries of the two coded spans of b1g, and the conversion script. DOI: https://doi.org/10.5281/zenodo.22216960 (concept DOI, which resolves to the most recent version; version 1.1.0 at the time of submission). The primary data sources are VUAMC.xml (Oxford Text Archive, 2541) and the v1.0 gold files of the EducationalTestingService/metaphor repository. The official test set was not modified. The provenance file records the cryptographic fingerprint of the test-side gold label vector, so that the integrity of the test can be verified.

Code availability

The full recipe of the lemma-prior test, the corpus audit, the register-by-register recomputation, and the correction experiment is available in the same repository: every condition and seed, with commit hashes – among them the commit that records the pre-registered decision rule. The document-level cross-validation of Section 5.6 is released alongside it: the out-of-fold predictions for the uncorrected and the corrected corpus, the per-fold and per-register statistics, and the paired comparison reported in Table 8. Every run carries a provenance sidecar file with the SHA-256 fingerprints of inputs, code, and outputs, the versions of the runtime environment, and the run timestamps.

Author contributions

Gergely Tóth: conceptualisation, data analysis, implementation of the experiments, and writing of the original draft. Gábor Péli, Miklós Gyorgyovich, and Bálint Duráczky: methodological review, validation of the findings, and revision of the manuscript. All authors read and approved the final manuscript.

References

Boisson, J., Espinosa-Anke, L., & Camacho-Collados, J. (2023). Construction artifacts in metaphor identification datasets. In *Proceedings of the 2023 Conference on Empirical Methods in Natural*

Language Processing (pp. 6581–6590). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.406

Herrmann, J. B. (2013). *Metaphor in academic discourse: Linguistic forms, conceptual structures, communicative functions and cognitive representations* [Doctoral dissertation, Vrije Universiteit Amsterdam]. LOT. https://www.lotpublications.nl/Documents/333_fulltext.pdf

Kim, Y., Kim, J. M., Jeong, J., Schmid, C., Akata, Z., & Lee, J. (2023). Bridging the gap between model explanations in partially annotated multi-label classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 3408–3417). IEEE. https://doi.org/10.1109/CVPR52729.2023.00332

Krennmayr, T., & Steen, G. J. (2017). VU Amsterdam Metaphor Corpus. In N. Ide & J. Pustejovsky (Eds.), *Handbook of linguistic annotation* (pp. 1053–1071). Springer. https://doi.org/10.1007/978-94-024-0881-2_39

Leong, C. W., Beigman Klebanov, B., & Shutova, E. (2018). A report on the 2018 VUA metaphor detection shared task. In *Proceedings of the Workshop on Figurative Language Processing* (pp. 56–66). Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-0907

Leong, C. W., Beigman Klebanov, B., Hamill, C., Stemle, E., Ubale, R., & Chen, X. (2020). A report on the 2020 VUA and TOEFL metaphor detection shared task. In *Proceedings of the Second Workshop on Figurative Language Processing* (pp. 18–29). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.figlang-1.3

Li, Y., Wang, S., Lin, C., & Guerin, F. (2023). Metaphor detection via explicit basic meanings modelling. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (pp. 91–100). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.acl-short.9

Merdjanovska, E., & Akbik, A. (2025). Token-level metrics for detecting incorrect gold annotations in named entity recognition. In *Findings of the Association for Computational Linguistics: EMNLP 2025* (pp. 15292–15304). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.findings-emnlp.827

Merdjanovska, E., Aynetdinov, A., & Akbik, A. (2024). NoiseBench: Benchmarking the impact of real label noise on named entity recognition. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* (pp. 18182–18198). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.emnlp-main.1011

Nahum, O., Calderon, N., Keller, O., Szpektor, I., & Reichart, R. (2025). Are LLMs better than reported? Detecting label errors and mitigating their effect on model performance. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing* (pp. 26770–26797). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.emnlp-main.1360

Neidlein, A., Wiesenbach, P., & Markert, K. (2020). An analysis of language models for metaphor recognition. In *Proceedings of the 28th International Conference on Computational Linguistics* (pp. 3722–3736). International Committee on Computational Linguistics. https://doi.org/10.18653/v1/2020.coling-main.332

Northcutt, C. G., Athalye, A., & Mueller, J. (2021). Pervasive label errors in test sets destabilize machine learning benchmarks. In *Proceedings of the 35th Conference on Neural Information Processing Systems (NeurIPS 2021)*, Datasets and Benchmarks Track. https://doi.org/10.48550/arXiv.2103.14749

Song, H., Kim, M., Park, D., Shin, Y., & Lee, J.-G. (2023). Learning from noisy labels with deep neural networks: A survey. *IEEE Transactions on Neural Networks and Learning Systems, 34*(11), 8135–8153. https://doi.org/10.1109/TNNLS.2022.3152527

Steen, G. J., Dorst, A. G., Herrmann, J. B., Kaal, A. A., Krennmayr, T., & Pasma, T. (2010). *A method for linguistic metaphor identification: From MIP to MIPVU*. John Benjamins. https://doi.org/10.1075/celcr.14

Stemle, E. W., & Onysko, A. (2018). Using language learner data for metaphor detection. In *Proceedings of the Workshop on Figurative Language Processing* (pp. 133–138). Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-0918

Zayed, O., McCrae, J. P., & Buitelaar, P. (2020). Adaptation of word-level benchmark datasets for relation-level metaphor identification. In *Proceedings of the Second Workshop on Figurative Language Processing* (pp. 154–164). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.figlang-1.22

Primary data sources

VUAMC.xml. VU Amsterdam Metaphor Corpus. Oxford Text Archive, 2541.

EducationalTestingService/metaphor, v1.0 gold files. NAACL-FLP shared task dataset.