Advertising Is Not Noise: Sponsored Speech Outranks New Terminology in Automatic Term Extraction

Martin Macak
Jožef Stefan Institute
Ljubljana, Slovenia
martinmacak@protonmail.com

Abstract

Long-form podcasts are an attractive corpus for terminology extraction: unedited, precisely dated and speaker-attributed. They are also sponsored, and a sponsor read is a script performed word for word across dozens of episodes, which is exactly what one of the two standard term-extraction signals is built to find. Collocation strength rewards phrases whose words recur together far above chance; rarity rewards words that are unusual in general English. We show on 788 episodes of one science-and-current-affairs podcast (10.2 million words, 2017–2026) that the two signals treat advertising in opposite ways, even after a cleaning step has removed most of it. Of the 100 phrases collocation ranks highest, the 87 that occur in episodes with third-party advertising labels spend a median 100% of their uses inside a marked advertisement, against a base rate of 4.7%; the top 100 by rarity and frequency-matched random phrases sit at 0%. The reason is mechanical: fixed wording raises collocation, while the brand names that would raise rarity are what a cleaner removes first, when it catches them. This is a claim about the head of the ranking, where a pipeline takes its cut, not about whether either signal separates advertising from terminology across their whole distributions; we could not test that cleanly, and we say why. The collocation result repeats on seven of eight other podcasts. We also report the cleaning step's recall and precision, and release a hand-checked lexicon of 59 coined terms.

Keywords

terminology extraction, spoken corpora, corpus contamination, collocation, neologism, podcast mining

1 Introduction

Podcast and video transcripts have become a standard corpus type in computational social science. Long-form spoken corpora are attractive because they are unedited, precisely dated, speaker-attributed, and available at a scale that written interview data never reaches.

They are also sponsored. Advertising is read aloud from a script, repeated close to verbatim across dozens of episodes, and built around brand names that do not occur in general English. Fixedness, repetition and lexical distinctiveness are what automatic term extraction looks for. The standard treatment of advertising in corpus work is to remove it as noise before analysis begins and to validate the removal by a gain on a downstream task. We argue that this understates the problem: removal is never complete, and what survives is not inert.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
Information Society 2026, Ljubljana, Slovenia
© 2026 Copyright held by the owner/author(s).
https://doi.org/10.70314/is.2026.sikdd.103

Automatic term extraction rests on two signals. *Collocation* asks how much more often the words of a phrase occur together than chance would predict; the usual measure is pointwise mutual information (PMI). *Rarity* asks how unusual those words are in everyday English. Both are in routine use, and a given pipeline usually ranks its candidates by one or the other. Combined scores exist [8], but their weighting is a free parameter, so which signal dominates a ranking remains a choice.

Claim. On a sponsored podcast corpus that has already been cleaned, collocation ranks the advertising that survives cleaning above genuine coined terminology, and rarity does not, because the cleaner has already taken the brand names out. The claim is about the top of the ranking, which is where a pipeline takes its term list. The phrase collocation ranks highest in our cleaned corpus is `midnight luxe hybrid mattress won`, a mattress advertisement. We are aware of no prior work that evaluates term-extraction signals against advertising as a labelled competing class on a spoken corpus, or that reports the two signals disagreeing about it.

Any study of terminology, new words or discourse on a sponsored corpus will therefore carry advertising in its results without saying so, and how much depends on which signal it ranked by, a choice usually made on unrelated grounds.

Contributions. (1) The asymmetry, measured: of the 100 phrases ranked highest by collocation, the 87 falling in labelled episodes spend a median 100% of their uses inside a segment a third party marked as advertising, while the top 100 by rarity and two sets of random five-word control phrases sit at 0%, against a 4.7% base rate (Section 4.1). (2) The mechanism: a sponsor read is a fixed expression repeated verbatim, which raises collocation, while the brand tokens that would raise rarity are what a cleaner removes (Section 5). (3) A cleaning step with both error rates: 84.9% recall against 318 third-party segments, 4.9 times a repetition-only baseline, at 47.4% calibrated precision, so about 4 percentage points (pp) of the corpus is genuine content removed with the advertising (Section 4.2). (4) A released lexicon of 59 coined terms in 64 surface rows, each dated to its first use (Section 4.3). (5) Ranking is not separation: whether either signal separates coinages from advertising is decided by how the comparison class is built (Section 4.1). (6) A replication on eight other podcasts (Section 6).

Section 2 reviews term extraction and corpus cleaning, Section 3 the corpus, labels and detector, Section 4 the ranking comparison and the detector's cost, and Sections 5 to 7 the mechanism, its limits and what follows.

2 Related Work

Term and keyphrase extraction. Contrast against a reference corpus (keyness [12], weirdness [1]) assumes a term is vocabulary the general language does not use. Internal cohesion (PMI [4], likelihood-ratio collocation tests [6], C-value/NC-value [8]) assumes its parts co-occur far above chance.

Neologism detection. New-word detection is dominated by *written* monitor corpora: newspapers [7] and large social-media pipelines [16]. Closest is Rossini and van der Plas [16], who reduce a very large token set to a hand-verified shortlist; their corpus is Reddit, written and unsponsored, so this contamination cannot arise there. The marker detector used here follows the pattern-extraction tradition [9].

Podcast and speech-recognition corpora. Large podcast corpora are recent: the Spotify Podcast Dataset [5] and the Text Retrieval Conference (TREC) Podcasts Track [10], built for search and summarization [11].

Corpus cleaning, and where this work departs from it. Removing non-content is a mature subfield: boilerplate detection [13], CleanEval [3], web-corpus construction [2]. In podcasts, Reddy et al. [15] detect the same material as the detector of Section 3.2 and validate it by showing that removal improves summarization. Across all of this work advertising is *noise* to be removed. That framing is right but incomplete: removal is never total, and one of the two standard signals prefers what survives.

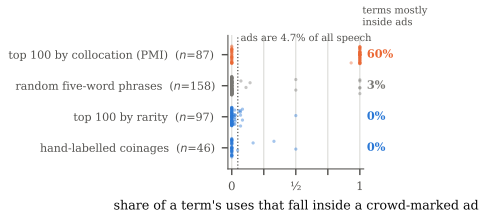


Figure 1: The two standard signals disagree about the same cleaned corpus. Each dot is one term, placed by the share of its uses that fall inside advertising as marked by SponsorBlock's volunteers. Ranked by collocation, in orange, the median term spends every one of its uses inside advertising; ranked by rarity, in blue, none does. Grey is the control: random five-word phrases matched on how often they occur. The right-hand column counts the terms that fall mostly inside advertising. The label comes from outside this project, not from the detector of Section 3.2.

3 Data and Method

3.1 Corpus and labels

The corpus is 788 dated episodes of a single long-running science-and-current-affairs podcast with two regular hosts: 829,989 diarized utterances, 10.23 million words and 1,042 hours of speech, recorded between 10 June 2017 and 13 May 2026. The audio was transcribed with Deepgram's `nova-3` model with speaker diarization.

We measure advertising against a label we did not produce. SponsorBlock [14] is a browser extension whose users mark the sponsor segments of a video so that others can skip them. Its contributors had marked 318 sponsor segments in 292 of these episodes before this work began, and those segments cover 4.7% of what is said in those episodes. That 4.7% is the share of its uses a phrase would spend inside an advertisement by chance. The label supports recall but not precision, because an unmarked stretch is not evidence that it is not an advertisement: contributors mark what they want to skip. Coverage is 37% of episodes and none before 2021, so every label-derived figure describes the later part of the corpus.

3.2 Contamination Removal

The detector is cue-anchored rather than repetition-based. It anchors on promotional formulae such as `brought to you by`, `use code` and `percent off`, and on brand tokens learned from the corpus, and treats a window around each anchor as an advertisement. A token enters the brand vocabulary only when it is absent from general English, because a first version that admitted merely rare words removed genuine scientific vocabulary along with the brands. The window is eight minutes of speech, 240 s either side of the cue, because the human-marked segments have a median length of 423 s with the cue anywhere inside them; past 240 s, a grid search shows, each further point of brand removal costs disproportionately more genuine vocabulary. The detector removes 9.6% of the corpus, 8.0 pp of it flagged as sponsor reads and the rest as boilerplate such as intros and sign-offs, event promotion, and clips; 90.4% is kept.

3.3 Term Discovery

Four detectors run over the cleaned corpus and their outputs are unioned. Each over-captures by design, because precision is bought in the hand-labelling step below. D1 fires on metalinguistic coining markers, surface patterns in which a speaker announces a term as they invent it (*what I call*, *a term I coined*), following [9]; it is the only detector that can find a term used once and then dropped. D2a selects rare multi-word phrases, with a phrase's rarity taken as the Zipf frequency of its rarest content word on the Zipf scale [18] as implemented in `wordfreq` [17], on which everyday words sit above 4 and we require 3.0 or below. D2b selects phrases with high collocation strength, PMI of 8.0 or more, because coinages built from ordinary words, such as `blue team`, are not rare and can be found only by cohesion. D3 selects rare single words at Zipf 2.5 or below. For a phrase $w_1 \dots w_n$ we compute PMI as $\log_2 p(\text{phrase}) - \sum_i \log_2 p(w_i)$, without length normalisation; each added word subtracts another term, so PMI rises with phrase length, which Section 4.1 controls for. A term must occur in at least three episodes to enter the candidate set, except when D1 fires, which admits a term on its first and only use. That exemption is what makes stillborn coinages findable at all, and Section 4.3 says what it does to one figure. The union yields 255,690 candidates.

3.4 Hand Labelling

A score combining rarity and collocation strength, with marker-detected terms admitted unconditionally, cut the candidates to a 700-term shortlist, because a quarter of a million candidates cannot be read. The author labelled the shortlist by hand on two questions: where the term came from (`origin`: coined here, borrowed from a technical field, ordinary language, or a transcription artifact) and how it is used (`use`: literally, to frame a claim, or both). Each row showed the term, its first use in context and the number of episodes it appears in, but nothing about whether it later caught on, so the labels are blind to a term's fate.

4 Results

4.1 The Ranking Result

On the cleaned corpus, after the detector of Section 3.2 has run, the five phrases collocation ranks highest are all sponsor copy, three of them variants of one mattress advertisement, led by `midnight luxe hybrid mattress won` (PMI 66.3) and `tj robinson's fresh pressed olive` (66.0). All five sit far above

the podcast's highest-scoring coined term, `time traveling money printer`, at 36.2.

Measured against the SponsorBlock label, the median term in the top 100 by collocation spends every one of its uses inside a marked segment, and 60% of them spend more than half (Figure 1). The median is over the 87 of the 100 that occur in a labelled episode; the other 13 are not counted. Ranked by rarity, 97 of the top 100 are attested and the median is 0%, as it is for the 46 hand-labelled coinages attested there. PMI rises with phrase length (Spearman $\rho = 0.56$) and 99 of the top 100 are five words long, so the controls are random five-word phrases, matched or not to the top 100 on how often they occur. Both sit at a median of 0%, with 3 to 4% above half, which is what a 4.7% base rate predicts for phrases seen once or twice, as most controls are. A top-collocation phrase is 17.8 times as likely to sit mostly inside marked advertising as a pooled control phrase (52 of 87 against 11 of 328, pooling the matched control in Figure 1 with a second, unmatched random set of 170; 95% confidence interval (CI) [9.7, 32.7]).

What this does not show. Whether either signal *separates* advertising from terminology across their whole distributions is a different question, and its answer depends on the comparison class. A signal can be captured by advertising at its top and useless as a classifier overall, and only the first matters to a pipeline that takes the top of a list. We scored the 26 well-attested coinages against every 1–5-word phrase inside the eight-minute window around a promotional cue, which we call the ad-window class ($n = 52{,}930$ phrase types occurring at least five times). That class is ordinary speech rather than advertising vocabulary: 95.9% of it contains none of the 118 brand words the detector learned. AUC, the area under the receiver operating characteristic (ROC) curve, is here the chance that a randomly picked coinage outscores a randomly picked phrase from the class. We report it undirected, as $\max(a, 1-a)$, because we had no prior about the direction; it therefore cannot fall below 0.5, and chance alone produces 0.545 at $n = 26$. Rarity reads 0.647 and collocation 0.545 against that class, neither a separation. Against the 2,177 phrases in the window that do carry a learned brand word, 2,064 of them with a defined PMI, collocation reads 0.605 (95% CI [0.51, 0.71], permutation $p = 0.07$), and rarity reads 0.981, a tautology, because brand membership is defined by absence from general English. No comparison class available to us tests separation cleanly, which is why the claim of this paper is about ranking. The test covers 26 of the 64 coinage rows, because scoring a term needs a minimum number of occurrences and that excludes the stillborn coinages.

4.2 Detector Cost

Cleaning is not free, and both of its error rates are reported as results (Table 1). *Recall* is how much of the real advertising the detector finds: 84.9% of the crowd-marked speech by utterance count (95% CI [84.4, 85.4]; 86.5% by duration), against 17.3% (24.5% by duration) for a baseline that removes only verbatim repetition, a ratio of 4.9. A blind check of 50 kept utterances found one advertisement (Table 1). Recall is an in-sample fit: the eight-minute window was chosen by a grid search over the same 318 segments it is scored against, so the true figure is lower. *Precision* is 47.4%. That figure corrects the author's raw blind judgement, 37.5% (33/88) of removed utterances judged to be advertising, for the fact that the judge is an imperfect detector: measured against the SponsorBlock marks, the judge recognised advertising inside marked segments 70.3% of the time (26 of 37) and called speech outside them advertising 7.9% of the time (5 of 63). We correct for both errors, treating unmarked speech as non-advertising, which Section 3.1 warns overstates the second correction and makes 47.4% a lower bound; the interval is wide, 95% CI [28.0, 71.7], because the calibration rests on 37 marked utterances. At that precision, 8.0 pp times one minus 0.474, roughly 4.2 pp of the corpus is genuine content removed with the advertising (95% CI [2.3, 5.8]). The two cautions cut in opposite directions: in-sample recall overstates the catch, so more advertising survives cleaning than the figure implies, while over-removal thins the pool the advertising is ranked against, which bounds the size of the effect but not its sign.

Table 1: Contamination detector cost. Recall is an in-sample fit (Section 4.2); by duration, recall reads 86.5% and the baseline 24.5%. Intervals are percentage points: Wilson for proportions, parametric bootstrap for the calibrated row. Bold marks the precision figure the paper uses.

	value	95% CI
recall vs. 318 segments, in-sample	84.9%	[84.4, 85.4]
repetition-only baseline	17.3%	[16.8, 17.9]
boilerplate precision	100% (8/8)	[67.6, 100.0]
sponsor precision, raw	37.5% (33/88)	[28.1, 47.9]
calibrated, two-sided	**47.4%**	[28.0, 71.7]
false-negative rate, kept utterances	2.0% (1/50)	[0.4, 10.5]

4.3 The Lexicon

The lexicon contains 59 distinct coinages across 64 surface rows, with transcription and inflectional variants aggregated. Twenty-four of the 59 (40.7%) are stillborn, used in the episode that introduced them and never again. That figure is a floor set by the discovery route rather than an estimate of how often a coinage dies: a term used once and not announced on air cannot be discovered at all, and every one of the 24 was found by the marker detector, the only route exempt from the three-episode threshold. The full lexicon is released with the paper.

5 Discussion

Why the two signals disagree. A sponsor read is written once and then performed dozens of times, word for word. That is exactly what collocation rewards, so advertising reaches the top of the collocation ranking ahead of anything a speaker coined in conversation. Rarity behaves the other way round, and the cleaning step is why. Brand names look extremely rare because they do not occur in ordinary English, and that absence is the very rule the detector uses to learn them, so cleaning strips out the words that would otherwise top the rarity ranking. What cleaning does not strip out is the sponsor's phrasing where it drifts outside a detected block, and phrasing is what collocation measures.

The problem appeared in this project's own pipeline. An early shortlist for this paper, ranked with these same two signals, behaved exactly as described here: leftover advertising copy outranked genuine coinages, and two of the podcast's best-known terms, `omega principle` and the short form `money printer`, dropped off the list. What brought them back into discovery was not a better advertising filter but fixing three defects in the marker detector, after which the known-answer check reads 12

of 12 at discovery. Both terms are lost again at the shortlist cut (Section 6), a separate failure we have not fixed.

How reliable the labels are. All hand labelling was done by one person, so a language model re-labelled all 700 rows blind and the two sets were compared. On the four-way `origin` split they agree at Cohen's $\kappa = 0.271$, which is poor; on the binary question the lexicon rests on, coined or not, at $\kappa = 0.447$, moderate at best, with the model recovering 75% of the hand-marked coinages.

6 Limitations

The main corpus is one podcast, two speakers and one subject area, and its advertising label covers 37% of its episodes and none before 2021. Of the paper's two products, the ranking asymmetry can be tested elsewhere and the lexicon cannot. The asymmetry needs a sponsored podcast with SponsorBlock coverage, the same cleaner and the two rankings, so we replicated it on eight other podcasts: a physics channel, a science outreach show, a health-science show, a wellness show, a business show and three political interview shows, between 50 and 316 episodes and 35 to 129 hours each. On seven of the eight, the median phrase in the top 100 by collocation spends 100% of its uses inside marked advertising. On the eighth, the health-science show, the eight of the top ten that occur in a labelled episode are all sponsor reads. But the top 100 also holds the host's own opening and closing lines, performed word for word in every episode, which the label marks as the host's own promotion rather than as advertising, and some genuine neuroscience terms, so the median there is 5%, or 68% with the host's promotions counted. On three shows, the physics channel and two of the political shows, nearly every five-word phrase that repeats across episodes is advertising, so the control phrases score as high as the top 100 and those three results say more about the pool than the ranking. The business, science outreach, wellness and remaining political show test the ranking cleanly, and all four replicate. On the rarity side, the top of the ranking on every show is a brand name or web address that survived cleaning. Where many survived, on the physics channel and the three political shows, the median of the top 100 reads 100% (four shows); where few did, on the business, science outreach and wellness shows, it reads 0% (three shows); on the health-science show it reads 8%. So the full asymmetry, collocation at 100% and rarity at 0% with the control at 0%, replicates on three shows. Rarity finds advertising wherever cleaning has left the brand names in, which is the mechanism of Section 5 from the other side. The lexicon does not travel, because judging whether a phrase is a coinage of these hosts requires knowing their vocabulary, and no second human annotator checked ours.

Against twelve terms fixed before the run, all of which survive discovery, the shortlist keeps 8 of 12; among the four it drops is `omega principle`, the hosts' best-known coinage, so it is missing from the released lexicon.

7 Conclusion

Advertising on a sponsored corpus is not noise for term extraction but a competing class, and the two standard signals treat it in opposite ways once a cleaner has done its work: collocation ranks what survives first, and rarity does not once the cleaner has taken the brand names out. From that point, which signal a pipeline ranks by decides how much advertising its term list inherits; before it, both deliver it. The mechanism is independent of the corpus by construction, because a script performed verbatim is a fixed expression in any corpus; whether the effect is as large elsewhere is what the replication in Section 6 tests. A practitioner who ranks by collocation on sponsored speech should check the top of the list against an advertising label before reading it as terminology.

References

[1] Khurshid Ahmad, Lee Gillam, and Lena Tostevin. 1999. University of Surrey participation in trec8: weirdness indexing for logical document extrapolation and retrieval (wilder). In *The Eighth Text REtrieval Conference (TREC 8)*. NIST Special Publication 500-246, 717.

[2] Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The wacky wide web: a collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43, 3, 209–226. DOI: 10.1007/s10579-009-9081-4.

[3] Marco Baroni, Francis Chantree, Adam Kilgarriff, and Serge Sharoff. 2008. Cleaneval: a competition for cleaning web pages. In *Proceedings of the Language Resources and Evaluation Conference*. DOI: 10.63317/4y7ib5ikczny.

[4] Kenneth Ward Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16, 1, 22–29. https://aclanthology.org/J90-1003/.

[5] Ann Clifton et al. 2020. 100,000 podcasts: a spoken English document corpus. In *Proceedings of the 28th International Conference on Computational Linguistics*, 5903–5917. DOI: 10.18653/v1/2020.coling-main.519.

[6] Ted Dunning. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19, 1, 61–74.

[7] Ingrid Falk, Delphine Bernhard, and Christophe Gérard. 2014. From non word to new word: automatically identifying neologisms in French newspapers. In *Proceedings of the Language Resources and Evaluation Conference*, 4337–4344. DOI: 10.63317/3tnyws3xnn2a.

[8] Katerina Frantzi, Sophia Ananiadou, and Hideki Mima. 2000. Automatic recognition of multi-word terms: the c-value/nc-value method. *International Journal on Digital Libraries*, 3, 2, 115–130. DOI: 10.1007/s007999900023.

[9] Marti A. Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th conference on Computational linguistics -*. Vol. 2, 539. DOI: 10.3115/992133.992154.

[10] Rosie Jones, Ben Carterette, Ann Clifton, Maria Eskevich, Gareth J. F. Jones, Jussi Karlgren, Aasish Pappu, Sravana Reddy, and Yongze Yu. 2021. Trec 2020 podcasts track overview. arXiv:2103.15953. (2021). DOI: 10.48550/arxiv.2103.15953.

[11] Rosie Jones et al. 2021. Current challenges and future directions in podcast information access. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1554–1565. DOI: 10.1145/3404835.3462805.

[12] Adam Kilgarriff. 2001. Comparing corpora. *International Journal of Corpus Linguistics*, 6, 1, 97–133. DOI: 10.1075/ijcl.6.1.05kil.

[13] Christian Kohlschütter, Peter Fankhauser, and Wolfgang Nejdl. 2010. Boilerplate detection using shallow text features. In *Proceedings of the third ACM international conference on Web search and data mining*, 441–450. DOI: 10.1145/1718487.1718542.

[14] Ajay Ramachandran and the SponsorBlock contributors. 2026. Sponsorblock: skip sponsorships on youtube. Browser extension and crowdsourced segment database. https://sponsor.ajay.app, source at https://github.com/ajayyy/SponsorBlock. Segments retrieved 28 July 2026. (2026).

[15] Sravana Reddy, Yongze Yu, Aasish Pappu, Aswin Sivaraman, Rezvaneh Rezapour, and Rosie Jones. 2021. Detecting extraneous content in podcasts. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 1166–1173. DOI: 10.18653/v1/2021.eacl-main.99.

[16] Diego Rossini and Lonneke van der Plas. 2026. From 124 million tokens to 1,021 neologisms: a large-scale pipeline for automatic neologism detection. arXiv:2605.06426. Accepted at the NeoLLM 2026 workshop, co-located with LREC-COLING 2026. (2026). DOI: 10.48550/arxiv.2605.06426.

[17] Robyn Speer. 2022. Wordfreq: a python library for looking up the frequencies of words in many languages. Version 3.0.2, Zenodo. Version 3.1.1 used here. (2022). DOI: 10.5281/zenodo.7199437.

[18] Walter J. B. van Heuven, Paweł Mandera, Emmanuel Keuleers, and Marc Brysbaert. 2014. SUBTLEX-UK: a new and improved word frequency database for British English. *Quarterly Journal of Experimental Psychology*, 67, 6, 1176–1190. DOI: 10.1080/17470218.2013.850521.