

RESEARCH ARTICLE

The Anatomy of a Press Release

A pre-registered census of 246,777 English-language releases on 1,151 websites

Émile Bolduc^{1*} 

¹ PPN World Research Center, PPN World, Montréal, Québec, Canada

* Corresponding author: ppnw@ppnsource.com · ORCID <https://orcid.org/0009-0005-7882-4741>

Citation: Bolduc, É. (2026). The anatomy of a press release: A pre-registered census of 246,777 English-language releases on 1,151 websites (PPN World Research Paper No. 2, Edition 1). PPN World. <https://www.ppnworld.com/research/anatomy-of-a-press-release>

Registered: October 11, 2026

Data collected: October 11, 2026 – October 13, 2026

Published: October 15, 2026

Corrected: October 17, 2026

Copyright: © 2026 PPN World. This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International licence (CC BY 4.0), which permits unrestricted use, distribution and reproduction in any medium, provided the original author and source are credited.

Funding: This study was funded by Bollé Communications inc., which operates PPN World and PPN Source. It received no external funding.

Competing interests: The author is affiliated with PPN World (Bollé Communications inc.), which published this paper, operates a commercial press-release monitoring service and supplied the archive the population was drawn from. PPN Source, a distributor the same company operates, is one of the population's websites (6 releases).

Abstract

Background. Press releases are among the most-written texts in professional communication and the resource journalists say they find most useful, yet their form has been measured mostly in samples of hundreds to a few thousand from one sector, or in vendor counts of distributor releases.

Methods. We set out to read every English-language release with a stated publication date from July 1, 2026 to October 6, 2026 in a press-release archive — 246,777 releases on 1,151 websites of distributors, companies, national, state and local governments, universities, sports leagues and nonprofits — following each website's robots.txt, and measured 176,492 of them on 877 websites with detection rules registered, with the analysis code, before any page was read, and revised once, in the open, after a first reliability test. The rules were checked against a coder reading the pages, and four hypotheses tested with a cluster bootstrap over websites and a Holm correction.

Results. The press release has split in two. On distributors' websites, 64% of releases end with an About section and 76% open with a dateline; on organizations' own websites, 3% and 17% (4.3% and 11% without the sports websites, whose pages are mostly news stories). Distributors' releases are longer by 155 words at the median (H1 supported), mostly outside the story: the body, from the lead to the About section, is not detectably longer. Executives' quotations carry a word of feeling more often than officials' (11.8% against 4.1%; H4 supported). The dateline (H2) and media-contact (H3) rules fell short of the reliability the registration required, so those two hypotheses are reported, not decided. FOR IMMEDIATE RELEASE survives on 3.2% of releases, 77% of them on a single website, the U.S. Department of Justice's.

Conclusions. The About section, a convention of the newswire era, belongs to the distributors' copy of a release; the dateline and the media contact point the same way, but their rules could not decide it. The copy an organization publishes on its own website has kept the story and little of the apparatus around it.

Keywords: press release; public relations; genre; content analysis; census; dateline; media contact; boilerplate; newswire; pre-registration

1. Introduction

A press release is a statement an organization publishes about itself, addressed to the news media and to the public. It is one of the oldest forms of public relations — the genre is conventionally dated to the statement Ivy Lee issued for the Pennsylvania Railroad after a train wreck in October 1906 (Hiebert, 1966) — and still the material journalists say they find most useful: in a 2025 vendor survey of more than 3,000 journalists in 19 markets, 72% named the press release the most useful resource a communications team can provide (Cision, 2025).

How releases are written is taught in every public-relations course: a dateline, a lead that answers who, what, when and where, a quotation from a leader, an About paragraph, a media contact, an end mark. Linguists describe the result as a hybrid genre, informative in form and promotional in purpose (Catenaccio, 2008), and as text written to be copied — "preformulated" for the journalist's article, with quotations ready to lift (Jacobs, 1999). Whether the form survives the move from the newswire to the web, and who keeps it, has not, to our knowledge, been measured at scale. Content analyses of the genre have mostly covered hundreds to a few

thousand releases from one sector (Sumner et al., 2014; Bossema et al., 2019; McLaren & Gurău, 2005; Lassen, 2006); an academic corpus of about 24,600 releases from one distributor served another question — whether news reuses their text (Nicholls, 2019); the largest count of the form itself we found is commercial, a distributor's analysis of 138,901 releases from ten newswires, limited to distributors (PR Gun, 2026). This paper sets out to measure every English-language release in a defined population over one quarter, with the hypotheses and the analysis registered before any page was read.

2. Materials and methods

2.1 Population

The population is every release in PPN World's archive with a stated publication date from July 1, 2026 to October 6, 2026, in English, from the sources the archive served: 246,777 releases on 1,151 websites. A website is an origin — scheme, host and port — and takes the publisher type of the sources that link to it. Distributors (commercial newswires) published 142,073 of the releases (58%); the rest were published by national governments and their agencies, state, provincial and city governments, sports leagues, companies, universities, nonprofits, political organizations and a few others on their own websites. National governments here include international organizations such as the United Nations and the European Union's institutions (13% of that group's weight), and sports leagues include their clubs and other sports bodies. The archive's sources are newsrooms, and not everything they publish is a release in the strict sense: the sports websites publish mostly news stories by their own writers, some companies publish blog posts, and 92% of the political organizations' weight is an archive of Donald Trump's posts on Truth Social. The census keeps what the sources published, as registered; the limitations give the own-website figures without the sports websites. The study is a census: every release of the population was included, none sampled.

2.2 Fetching and extraction

Pages were fetched from October 11, 2026 to October 13, 2026. Each website's robots.txt was read first and applied to every release (Koster et al., 2022); requests to each website went one at a time, at least a second apart, from a client that named itself, and once more as a browser if refused. A bot check was never solved and a certificate error never bypassed. The text was extracted with Mozilla Readability, the algorithm of Firefox's Reader View (Barbaresi, 2021; Bevendorff et al., 2023). A release is measured if its page answered with at least 50 words of English carrying the archive's headline: 176,492 releases (72%) on 877 websites. The rest are counted by reason — the largest are pages refused or gone (19,795), invalid certificates (17,428, nearly all on one distributor's website), bot checks (11,059) and unreachable robots.txt files (6,077). Within each website, measured releases are weighted so that every website keeps its share of the population.

2.3 Measures

One registered program reads each page: the headline (length, a number, an announcement verb); the dateline (a place in

capitals, a date or a distributor's tag, then a dash, at the top of the text); the sub-headline; a byline; the length in words, overall and of the body from the lead to the About section; quotations of eight words or more with a speech verb, attributed to an executive when a corporate title is beside the speaker and to an official when a public title is, and whether they contain a word of feeling (pleased, excited, thrilled, delighted, proud, honored, happy, glad, elated); the About section; a media contact — a contact cue with an email address or a phone number; an end mark (###, -30-); FOR IMMEDIATE RELEASE; images, video and links.

2.4 Hypotheses and analysis

Four hypotheses were registered: distributors' releases are longer than releases on organizations' own websites (H1, the difference of weighted median word counts); the dateline is a distributor's convention (H2, the difference of shares); most releases name no one to call (H3, the share with a contact, against one half); and emotion words are more frequent in executives' quotations than in officials' (H4, commercial against public bodies' releases, by quotation). Intervals and one-sided p-values come from 2,000 bootstrap replicates that resample websites within publisher type; the four tests carry a Holm correction (Holm, 1979). A measure could decide a hypothesis only if it agreed with a coder reading the page on at least 90% of releases with Krippendorff's $\alpha \geq .800$ (Krippendorff, 2004; Lombard et al., 2002). What had been looked at before registration — the archive's identifiers, addresses and headlines, and two pilots of 100 and 1,001 releases outside the population — is listed in the registration, as recommended for analyses of existing data (Weston et al., 2019).

2.5 Reliability of the measures

Reliability. Coder A, a language model (Claude Opus 5.5, Anthropic) — models of this kind have matched or outperformed crowd workers on comparable annotation tasks (Gillardi et al., 2023) — coded 200 releases drawn at random — at least 60 from distributors and 60 from public bodies, as registered; the draw gave 106, 70 and 24 from other publishers — from the page itself, on the registered codebook, without the detector's output. A second, human coder was registered for a subset and withdrew after five releases; the paper therefore rests on one coder, a language model, and says so. Where the registered rules failed, they were revised on those 200 releases and tested once, unchanged, on 100 new ones drawn the same way (at least 30 and 30; the draw gave 56, 37 and 7) and coded blind (Figure 2). The extraction gave the whole release on 78% of the test pages and part of it on most of the rest.

2.6 Deviations from the registration

1. Pace. The measuring step ran in parallel and more websites were fetched at once than registered; the courtesy rules were unchanged, and re-measuring the second pilot in parallel gave identical results. The research computer slept for part of the census; releases that failed while it slept were fetched again with the computer awake (both counts reported).
2. Refusals with 429 Too Many Requests were fetched once more, at least 30 seconds apart per website; the hypotheses

are decided on the registered count, the retried count is a robustness row.

3. The second, human coder was registered for 50 releases on twelve questions, then for 20 on four; the coder stopped after five, and no agreement between coders is computed.
4. The registered contact and executive-quotation rules failed the reliability test on the first 200 releases. The contact rule now also reads the page around the extracted text; executive titles are read in any letter case and linked to a speaker named later by surname; quotation marks are paired within a paragraph. GOV.UK pages that are not releases (statistics notices, consultations, publications) were set aside: 3,655 pages, counted in the population among those not measured. The revised rules were frozen before a second, blind test on 100 new releases, whose

result was recorded before the census was re-measured. The registered results are reported beside the revised ones.

3. Results

3.1 Two press releases

Figure 1 reads a release from top to bottom on both kinds of website. The parts the newswire made standard — the dateline at the top, the About section and the contact at the bottom — are the norm on distributors' websites and rare on organizations' own: an About section closes 64% of distributors' releases against 3%, a dateline opens 76% against 17%, a contact appears on 65% against 14%. The quotation is the part both keep: 59% of distributors' releases and 42% of the others quote someone.

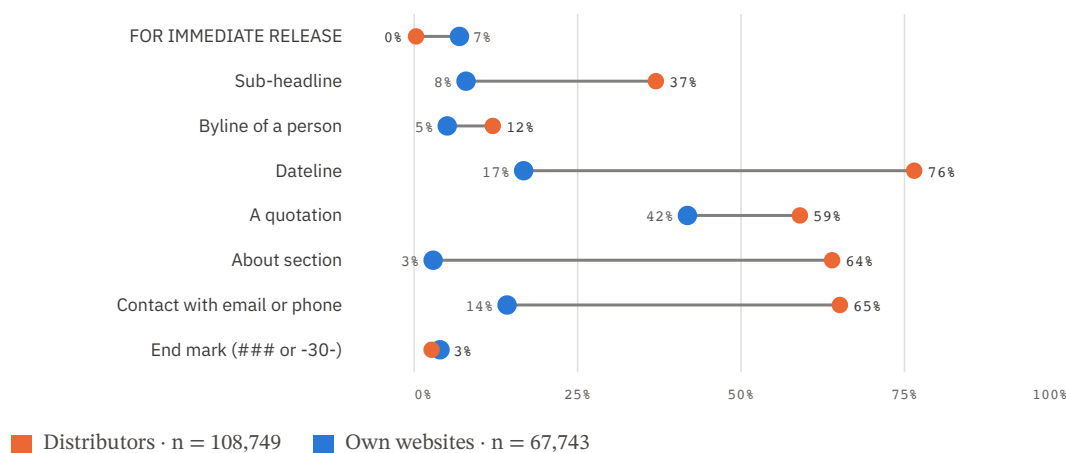


Figure 1. Two press releases. Share of releases carrying each part, on distributors' websites and on organizations' own websites, weighted for the releases that could not be read. Parts in the order a release is read. The sub-headline, end-mark and byline rules fell short of the reliability rule (Figure 2); read those rows as indicative.

The split is not a matter of companies against governments. Companies publishing on their own newsrooms keep more of the apparatus than public bodies — 21% carry an About section and 24% a contact — but far less than the same kind of text carries on a distributor's website (Table 1). State and local

governments are the exception among public bodies: 46% of their releases name a contact. Sports leagues, the largest group of own websites measured, almost never do (0.6%): their pages are mostly news stories, not releases.

Table 1. The anatomy of a press release, by publisher. Releases measured, median length in words, and the share carrying each part, weighted for the releases that could not be read. Distributors are commercial newswires; every other row publishes on its own websites. National governments include international organizations; sports leagues include clubs and other sports bodies, whose pages are mostly news stories; most of the political organizations' weight is an archive of Donald Trump's posts on Truth Social.

Publisher	Measured	Median words	Dateline	About	Contact	Quotation	For immediate release
All releases	176,492	588	51%	38%	43%	52%	3%
Distributors	108,749	637	76%	64%	65%	59%	0%
Companies	4,120	718	17%	21%	24%	36%	0%
National governments	23,152	401	9%	2%	19%	35%	15%
State and local governments	7,106	405	21%	4%	46%	43%	13%
Universities	4,134	678	2%	2%	17%	66%	0%
Sports leagues	26,362	594	27%	1%	1%	47%	0%
Nonprofits	1,621	513	15%	10%	11%	46%	0%
Political organizations	873	128	2%	0%	0%	5%	0%
All own websites	67,743	482	17%	3%	14%	42%	7%

3.2 Length

The median release runs 588 words, at a Flesch–Kincaid grade of 15 (Kincaid et al., 1975) — written for a university

reader. H1 is supported: distributors' releases are longer than releases on organizations' own websites by 155 words at the median (95% interval 55 to 244; Holm-adjusted $p = 0.005$).

The registration asked the same question of the body alone, from the lead to the About section, contact or end mark (R5). There the difference all but vanishes: 468 words on distributors' websites, 465 on organizations' own (difference 3, 95% interval -84 to 96). Without the sports websites, whose news stories are long, it is 57 words (95% interval -2 to 117; a check that was not registered), against 204 for the whole release. A distributor's release is longer mostly because of what surrounds the story — the dateline, the About section, the contact block, the legal notices — and not detectably because the story is.

3.3 Whom to call

A contact block is the one part of a release written for the journalist alone. H3 asked whether most releases now name no one to call. As measured, 43% of releases carry an email address or a phone number under a contact cue (95% interval 24% to 57%) — 65% on distributors' websites, 14% on organizations' own.

The measure could not decide the hypothesis. On 100 releases drawn to test it, the rule agreed with a coder reading the page on 89% of them ($\alpha .779$), under the registered 90%. It is rarely wrong when it finds a contact (precision 96%) but missed 16% of those the coder found, most often a contact under a cue it does not know ("Questions?", "Address for clarification") or one the extraction left out. The census share is therefore a floor, and close enough to one half that the registered question stays open: H3 is reported, not decided. Misses

at that rate would not, however, close the gap between the two kinds of website.

3.4 The dateline

A news story filed from away opens with the place it was written in, in capitals (Associated Press, 2026); distributors apply the convention to releases, often with the date and their own name. 51% of releases open with one: 76% on distributors' websites and 17% on organizations' own, a difference of 60 points (95% interval 23 to 73).

H2 is reported, not decided. The dateline rule passed the reliability test on the first 200 releases coded (90.5% agreement, $\alpha .810$) and failed it on the 100 drawn to test the revised rules (87%, $\alpha .739$), and the registration applies the test to the second draw. Its errors run mostly one way: when it finds a dateline, it is almost always right (precision 98%); it missed 20% of the datelines the coder found.

3.5 Quotations and feelings

52% of releases quote someone in a sentence of eight words or more. H4 compares executives quoted in commercial releases with officials quoted in public bodies' releases. Executives' quotations carry a word of feeling — pleased, excited, thrilled, delighted, proud, honored, happy, glad, elated — in 11.8% of cases (132,328 quotations); officials' in 4.1% (13,313). H4 is supported: the difference is 7.6 points (95% interval 4.9 to 8.8; Holm-adjusted $p = 0.002$).

The emotion is a formula rather than the rule: most executives' quotations contain none of the nine words. Where one appears, it is in the register Jacobs (1999) describes — a quotation written by the communications team, to be lifted whole into an article.

Table 2. The four registered hypotheses. Estimates with the revised detection rules, cluster-bootstrap 95% intervals and Holm-adjusted p-values; the last column gives the estimate with the rules as registered. H2 and H3 rest on measures that fell short of the registered reliability rule and are not decided.

Hypothesis	Estimate	95% interval	p (Holm)	Verdict	As registered
H1 · Distributors' releases are longer (median words)	155	55 to 244	0.005	Supported	169
H2 · The dateline is a distributor's convention (difference, points)	60	23 to 73	0.002	Not decided	60
H3 · Most releases name no one to call (share with a contact)	43%	24% to 57%	0.185	Not decided	29%
H4 · Executives' quotations carry more emotion words (difference, points)	7.6	4.9 to 8.8	0.002	Supported	8.1

3.6 The marks of the fax era

FOR IMMEDIATE RELEASE told a newsroom that a release on paper could run at once. On a web page it has no reader left to instruct, and it is rare: 3.2% of releases carry it, 0.3% on distributors' websites. Where it survives is government, and mostly one website: the U.S. Department of Justice prints it on every release measured, and its releases are 77% of all those that carry it — an estimate resting on the 365 of its 4,862 releases that could be read (almost all the rest were behind a bot check), which the weights let stand for all of them. National governments' share is 14.9% with it and 0.5% without it; state and local governments keep it on 13.0% of releases. The end mark — ### or -30-, "end of copy" — is on 3.2% of releases, a quarter of state and local governments' (26%) and few others; the end-mark rule, like the sub-head-

line and byline rules, fell short of the reliability rule (Figure 2), so these shares are indicative.

Headlines are long. The median is 81 characters, 93 on distributors' websites and 67 on organizations' own; 75% of all headlines and 86% of distributors' exceed 60 characters, close to the 90.2% the distributor count found on newswires (PR Gun, 2026). An announcement verb (announces, launches, unveils...) opens 31% of distributors' headlines and 8% of the others.

3.7 Robustness checks

Robustness. Weighting every website equally, leaving out the five largest websites, leaving out bylined stories (as far as the byline rule, which fell short of the reliability rule, finds them), pages read only after rendering, or pages the identified client could not read, dropping the weights, and counting the releases first refused with 429 Too Many Requests (1,853 more

measured) or keeping as first recorded the releases lost while the research computer slept (917 fewer measured) each re-run the four tests. H1 and H4 are supported in every one of them; H1's difference ranges from 94 to 163 words, and H4's

from 5.5 to 7.7 points. With the rules as registered, 179,606 releases were measured and the estimates were 169 words (H1), 60 points (H2), 29% with a contact (H3) and 8.1 points (H4); every registered test was then nominally supported.

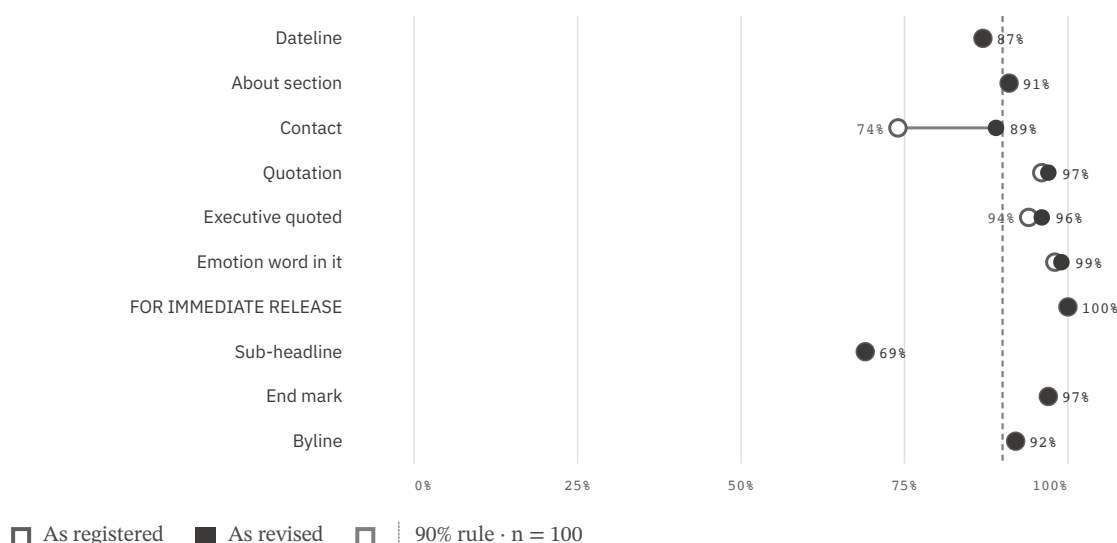


Figure 2. How far each measure can be trusted. Agreement between the detection rule and a coder reading the page, on 100 releases drawn after the rules were revised (hollow: the rule as registered; filled: as revised). A measure decides a hypothesis only at 90% agreement with Krippendorff's $\alpha \geq .800$.

4. Discussion

4.1 Key findings

- 1. The About section belongs to the distributors.** 64% of releases on distributors' websites end with a standard paragraph about the organization; 3% of releases on organizations' own websites do. The same split holds for the dateline (76% against 17%).
- 2. Distributors' releases are longer — mostly around the story (H1).** At the median, a release runs 637 words on a distributor's website and 482 on the organization's own (95% interval of the difference 55 to 244; H1 supported). The body, from the lead to the About section, is not detectably longer: 468 against 465 words (468 against 411 without the sports websites' news stories). Most of the difference is the apparatus around it.
- 3. The typical release runs 588 words; a company's, 718; a national government's, 401.** At the median, a press release runs 588 words: 637 on distributors' websites and 482 on organizations' own. By who publishes it, a company's runs 718 words, a university's 678, a sports league's 594, a state or local government's 405 and a national government's 401. Most of the gap is in the story itself: the body of a company's release runs 637 words, a national government's 387.
- 4. Executives sound pleased more often than officials (H4).** A word of feeling — pleased, excited, thrilled, delighted, proud — appears in 11.8% of executives' quotations in commercial releases and 4.1% of officials' quotations in public bodies' releases (difference 7.6 points, 95% interval 4.9 to 8.8; H4 supported).

5. FOR IMMEDIATE RELEASE survives on one federal website — and in state and local governments.

The line written for the fax machine survives on 3.2% of releases, and 77% of those are on one website: the U.S. Department of Justice prints it on every release we could read (365 of 4,862). Without it, national governments' share falls from 14.9% to 0.5%, and the overall share to 0.8%. State and local governments keep it on 13.0% of releases; distributors on 0.3%.

6. Headlines are long, and longest on distributors.

The median headline runs 81 characters — 93 on distributors' websites, 67 on organizations' own. 75% of headlines are longer than 60 characters, 86% on distributors.

4.2 What it means for communicators

How long is a typical press release? 588 words at the median across 176,492 English-language releases measured in 2026: 637 on distributors' websites, 482 on organizations' own. The story itself — lead to About section — runs 468 words on distributors' websites and 465 on organizations' own.

Does a press release still need a dateline? On a distributor's website, it is the norm: 76% of releases open with one. On an organization's own website, 17% do (11% without the sports websites). These are estimates: the dateline rule fell short of the reliability the registration required, and it misses datelines more often than it invents them. The dateline belongs to the copy that goes out on the wire.

Should a press release include a media contact? It is the only part written for the journalist alone, and the copy on the organization's own website usually leaves it out: as measured, 14% carry an email address or phone number (23% without the sports websites), against 65% on distributors' websites. These are estimates: the contact rule fell short of the reliability the registration required. A newsroom page without a name to

call sends the reporter to the distributor's copy — or elsewhere.

Is FOR IMMEDIATE RELEASE still used? Rarely: 3.2% of releases carry it, and 77% of those come from one website, the U.S. Department of Justice's. Leave it out and the share falls to 0.8%, and state and local governments are then the ones that keep it most (13.0%); on distributors' websites it is almost absent (0.3%).

Do executives' quotes sound more emotional than officials'? Yes, though it is not the rule: 11.8% of executives' quotations in commercial releases contain pleased, excited, thrilled, delighted, proud or a similar word, against 4.1% of officials' quotations in public bodies' releases.

4.3 Limitations

- The population is one archive's sources, chosen for a commercial monitoring service, over one quarter, in English. Distributors carry more than half of its releases; organizations that publish only on their own websites, and those the archive does not follow, are under-represented. Results are reported by publisher type so that no group's weight hides another's.
- Not everything the archive's sources publish is a release in the strict sense. Sports websites carry 38% of own websites' weight and publish mostly news stories; without them, 4.3% of own websites' releases carry an About section, 11% a dateline and 23% a contact (against 3%, 17% and 14%), 52% of all releases carry a contact (95% interval 33% to 64%), and H1's difference is 204 words (122 to 269). The split between the two kinds of website holds. These checks were made after publication and were not registered.
- Two distributors' websites carry 68% of distributors' weight, so the distributor figures describe them above all; without the five largest websites (R4), H1 and H4 are still supported.
- Releases behind bot checks, behind invalid certificates, refused to our client or on websites whose robots.txt could not be read are counted, not guessed at. Weights restore each website's share but cannot restore a website with nothing measured — 17% of the population's releases are on such websites: 19% of distributors', nearly all on two websites, and 20% of national governments', half on one military news website — and where few of a website's releases could be read, those few stand for all of them.
- Publisher types are the archive's own, and a few websites are mistyped: the Government of Alberta and the Australian Bureau of Statistics among companies (5% of that group's weight), CERN and the SANS Internet Storm Center among universities. A page reached both with and without www counts on two websites (245 measured pages). Both are left as registered.
- The measures are rules. The dateline and contact rules fell short of the reliability the registration required on the test draw, and the hypotheses resting on them are not decided.
- The only coder is a language model. Its codings will be deposited with the data so that anyone can check them against the pages.

- Extraction is imperfect: on about one test page in five, the text extracted was only part of the release.
- The study is funded and published by a company that sells press-release monitoring; the results report distributors in aggregate and name none of their websites but the publisher's own.

5. Conclusions

There are two press releases now. The copy on a distributor's website still has the newswire's parts: a dateline, an About section, a contact to call. The copy on an organization's own website mostly does not. The difference in length is mostly that apparatus, not the story.

Pre-registration. The hypotheses, population, measures and analyses were registered on 11 October 2026, before any page of the population was read (commit 4b769927). The manifest of the registered files' hashes was submitted to OpenTimestamps as the census began; the Bitcoin block that anchors it (970,932) was mined 21 minutes after the first page was requested, and the manifest and its proof will be deposited with the data. The registration, its addendum, the codebook and the deviations log are published unedited. <https://www.ppnworld.com/research/anatomy-of-a-press-release/registration/press-release-anatomy.md>

Data availability. The anatomy table by publisher type is published as CSV under CC BY 4.0; its groups keep the archive's names (wire: distributors; government: national governments; municipality: state and local governments; league: sports), and its dateline, contact, sub-headline, byline and end-mark columns carry the reliability limits stated above. The per-release data — every release of the population with its address, website, publisher type, outcome, weight and measures, without page text, quotations or contact details — will be deposited on Zenodo with the code of every step; until then, the Research Center sends it on request. The registration, its addendum, the codebook, coder A's instructions and the deviations log are published unedited. <https://www.ppnworld.com/research/anatomy-of-a-press-release.csv>

Ethics statement. The study read public web pages and involved no human participants; no ethics review was sought. It followed each website's robots.txt, identified itself first and retried once as a browser when refused (a robustness row leaves those pages out), limited its pace, and never solved a bot check or bypassed a certificate error. Contact blocks contain personal data: only their presence and type enter the data; the pages kept to re-run the measures are not published.

Use of AI tools. Claude (Anthropic) was coder A in the validation and, used through Claude Code with Claude Opus 5.5, prepared the data-collection and analysis software, drafts of the text and this layout, under the author's supervision. The author is responsible for the content of the paper.

Corrections. Corrected on October 17, 2026, after publication on October 15, 2026; the corrections are listed, dated, on the paper's web page, and no estimate of the four hypotheses changed. <https://www.ppnworld.com/research/anatomy-of-a-press-release>

References

- Associated Press. (2026). The Associated Press stylebook 2026–2028 (58th ed.). Basic Books.
- Barbaresi, A. (2021). Trafalatura: A web scraping library and command-line tool for text discovery and extraction. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, 122–131. <https://doi.org/10.18653/v1/2021.acl-demo.15>
- Bevendorff, J., Gupta, S., Kiesel, J., & Stein, B. (2023). An empirical comparison of web content extraction algorithms. Proceedings of the 46th International ACM SIGIR Conference on Research and

- Development in Information Retrieval.
<https://doi.org/10.1145/3539618.3591920>
- Bossema, F. G., Burger, P., Bratton, L., Challenger, A., Adams, R. C., Sumner, P., Schat, J., Numans, M. E., & Smeets, I. (2019). Expert quotes and exaggeration in health news: A retrospective quantitative content analysis. *Wellcome Open Research*, 4, 56.
<https://doi.org/10.12688/wellcomeopenres.15147.2>
- Catenaccio, P. (2008). Press releases as a hybrid genre: Addressing the informative/promotional conundrum. *Pragmatics*, 18(1), 9–31.
<https://doi.org/10.1075/prag.18.1.02cat>
- Cision. (2025, May 7). 2025 State of the Media report [Survey of more than 3,000 journalists in 19 markets].
- Gilardi, F., Alizadeh, M., & Kubli, M. (2023). ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30), e2305016120.
<https://doi.org/10.1073/pnas.2305016120>
- Hiebert, R. E. (1966). *Courtier to the crowd: The story of Ivy Lee and the development of public relations*. Iowa State University Press.
- Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2), 65–70.
<https://www.jstor.org/stable/4615733>
- Jacobs, G. (1999). Preformulating the news: An analysis of the metapragmatics of press releases. *John Benjamins*.
<https://doi.org/10.1075/pbns.60>
- Kincaid, J. P., Fishburne, R. P., Rogers, R. L., & Chissom, B. S. (1975). Derivation of new readability formulas for Navy enlisted personnel (Research Branch Report 8-75). *Naval Technical Training Command*.
- Koster, M., Illyes, G., Zeller, H., & Sassman, L. (2022, September). Robots Exclusion Protocol (RFC 9309). Internet Engineering Task Force.
<https://doi.org/10.17487/RFC9309>
- Krippendorff, K. (2004). *Content analysis: An introduction to its methodology* (2nd ed.). Sage.
- Lassen, I. (2006). Is the press release a genre? A study of form and content. *Discourse Studies*, 8(4), 503–530.
<https://doi.org/10.1177/1461445606064830>
- Lombard, M., Snyder-Duch, J., & Bracken, C. C. (2002). Content analysis in mass communication: Assessment and reporting of intercoder reliability. *Human Communication Research*, 28(4), 587–604.
<https://doi.org/10.1111/j.1468-2958.2002.tb00826.x>
- McLaren, Y., & Gurău, C. (2005). Characterising the genre of the corporate press release. *LSP and Professional Communication*, 5(1), 10–30.
- Nicholls, T. (2019). Detecting textual reuse in news stories, at scale. *International Journal of Communication*, 13, 4173–4197.
- PR Gun. (2026, August 25). [Analysis of 138,901 press releases from ten newswires, 2021–2026] [Press release].
- Sumner, P., Vivian-Griffiths, S., Boivin, J., Williams, A., Venetis, C. A., Davies, A., Ogden, J., Whelan, L., Hughes, B., Dalton, B., Boy, F., & Chambers, C. D. (2014). The association between exaggeration in health related science news and academic press releases: Retrospective observational study. *BMJ*, 349, g7015. <https://doi.org/10.1136/bmj.g7015>
- Weston, S. J., Ritchie, S. J., Rohrer, J. M., & Przybylski, A. K. (2019). Recommendations for increasing the transparency of analysis of preexisting data sets. *Advances in Methods and Practices in Psychological Science*, 2(3), 214–227.
<https://doi.org/10.1177/2515245919848684>

Appendix. Data behind the figures

Each table gives the numbers drawn in the figure of the same number.

Table A1. Two press releases

Part	Distributors	Own websites
FOR IMMEDIATE RELEASE	0.3%	6.9%
Sub-headline	37.0%	7.9%
Byline of a person	12.0%	5.1%
Dateline	76.5%	16.7%
A quotation	59.0%	41.8%
About section	63.9%	2.9%
Contact with email or phone	65.1%	14.2%
End mark (### or -30-)	2.7%	4.0%

Table A2. How far each measure can be trusted

Measure	Agreement, as registered	Agreement, as revised	α , as revised	Can decide
Dateline	87%	87%	.739	reported only
About section	91%	91%	.811	decides
Contact	74%	89%	.779	reported only
Quotation	96%	97%	.940	decides
Executive quoted	94%	96%	.910	decides
Emotion word in it	98%	99%	.928	decides
FOR IMMEDIATE RELEASE	100%	100%	1.000	decides
Sub-headline	69%	69%	.326	reported only
End mark	97%	97%	.558	reported only
Byline	92%	92%	.162	reported only