



# Multi-Word Combinations in Science Communication: A Corpus-Driven Study of University Press Releases

Emilia Petrocelli <sup>1\*</sup>

 0000-0001-6800-9659

Laura Di Ferrante <sup>2</sup>

 0000-0003-4182-6872

Elisa Ghia <sup>3</sup>

 0000-0002-0353-2913

<sup>1</sup> Department of Human Studies, University for Foreigners of Siena, Siena, ITALY

<sup>2</sup> Department of Communication and Social Research, Sapienza University of Rome, Rome, ITALY

<sup>3</sup> Department of Political and Social Science, University of Pavia, Pavia, ITALY

\* Corresponding author: [petrocelli@unistrasi.it](mailto:petrocelli@unistrasi.it)

**Citation:** Petrocelli, E., Di Ferrante, L., & Ghia, E. (2022). Multi-Word Combinations in Science Communication: A Corpus-Driven Study of University Press Releases. *Online Journal of Communication and Media Technologies*, 12(4), e202224. <https://doi.org/10.30935/ojcm/12247>

## ARTICLE INFO

Received: 30 Apr 2022

Accepted: 13 Jul 2022

## ABSTRACT

Entrepreneurial universities sponsor their research through the media to increase their popularity. This in turn may engender a biased representation of scientific knowledge in the media. We present the results of a quali-quantitative analysis of phraseology characterizing press releases (PRs) issued by universities and focusing on published research. Our investigation aims to identify typical n-grams in university PRs and their main purposes. The study is based on a developmental 68,913-word collection of university PRs. 5-word n-grams were assigned categories and functions. Results contribute to the characterization of PRs as a hybrid genre at the crossroads between news and ads. Moreover, findings add knowledge about this specific type of PRs, which cover some of the same functions as scientific discourse while at the same time enhancing promotional features. This is shown by the recurrent use of phrasal units fulfilling both informative and promotional needs to the detriment of others only fulfilling informative functions.

**Keywords:** science communication, academic research, press releases, phraseology, n-grams, academic discourse

## INTRODUCTION

The utmost goal of research is to contribute to relevant changes in society and the first step to achieving this goal is to disseminate data results effectively. Scientific dissemination has a twofold objective: on the one hand, it aims to reach a primary audience of experts, to enhance awareness in the research community, create new study networks, and boost the reproducibility of research; on the other hand, it aims to tackle a secondary audience, made of laypeople and practitioners who could benefit from using research findings. As with virtually all areas of life, access to scientific knowledge has been rearranged and, to some degree, eased by the internet, to the point that the majority of scholarly journals nowadays are online, and many scholarly books have a printed and/or an online copy. However, digital science dissemination has a multitude of other modes beyond these more traditional ones: academic and non-academic social media, blogs, and forms of interaction like TED talks or YouTube videos. During the recent COVID-19 pandemic, the circulation of research has been ever more necessary. The breakout of the virus and its continuous spread have pushed every single person to eagerly look for information from all different sources. The hyper-stimulation of communication has provoked what the Director General of the World Health Organization has called *Infodemic* (Munich

Security Conference, February 15, 2020), which is the overflow of news, including false or misleading information. What is more, studies have shown that false-based information spreads much faster than evidence-based one (e.g. Pulido et al., 2020). Those who aim at an informational hygienic approach by looking for trustworthy information most presumably try to rely on scientific, popularized news including data issued by universities on their websites, as higher education institutions are perhaps the most authoritative sources of research-based news.

Universities use dedicated pages to their press releases (henceforth PRs), which are usually about the research that their affiliates carry out either as single authors/group, or in collaboration with scientists from other institutions. Laypeople's reliance on PRs has automatically increased access to science and research-related vocabulary, procedures, and, by way of that, has put research agents and institutions in the foreground. The entrepreneurial urge of universities to increase their popularity (Clark, 1998; Santiago et al., 2008) is indeed well satisfied by PRs, which not only are nurtured by the deontological need to maximize the uptake of research, but also to enhance the visibility and reputation of the institution, in line with what Catenaccio (2007, 2008), Jacobs (1999), and Jacobs and Van Hout (2009) outlined when analyzing corporate PRs.

Given that the shift from science to science communication involves the use of less technical language, alongside recontextualization and reformulation (Garzone, 2014, 2021; Gotti, 2014; Linell, 1998), we argue that some specific language traits of academic PRs can be identified. In particular, differently from corporate PRs, university ones are often based on a reference text, namely the original scientific article. In other words, university PRs are the result of a centripetal motion, where textual, discursive and ethical dimensions converge: from a textual perspective, PRs respond to both informational and promotional functions, but at the same time have also absorbed some characteristic elements and formulae of academic discourse. In addition, these dimensions interact with the whole ethical sphere of perceptions and expectations attached to universities as reliable institutions and producers of science. The general aim of this paper is that of exploring the linguistic building blocks that tie all these dimensions together in PRs. In particular, many studies showed that genre-specific features can be disclosed through n-grams and lexical bundles (e.g., Biber et al., 1999; Stubbs, 2002), which are also relevant in discriminating between given registers and discourse types. For this reason, we identified those fixed multi-word combinations that mostly recur in research-focused university PRs, to investigate their functions and to see whether there is some degree of overlap with the phraseology of scientific articles. The study is part of a larger project based on two matching corpora: one comprising academic articles and the other comprising their respective popularized pieces. In this paper, we focus on PRs issued by the universities themselves.

## LINGUISTIC PATTERNS OF SCIENCE DISSEMINATION

---

Together with science journalism and science museums, PRs are a pivotal tool for science communication, in that they offer free access to research data, in a form of language that is more accessible and less technical than the one used in scientific articles.

Academic PRs, being issued by educational institutions whose prior mission is to disseminate scientific knowledge, are authoritative sources of information for both researchers and laypeople. Historically, media have played a key role in aiding the diffusion of information contained in PRs, which boosted the process by offering pre-formulated chunks of language to be later copied by journalists in their news (Jacobs, 1999). Textual traits of pre-formulation have become a constituent feature of all types of PRs, including e-releases, which grant direct access to information and therefore do not necessitate the filtering of journalism (Sleurs & Jacobs, 2005; Strobbe & Jacobs, 2005). What is more, in the case of academic PRs, pre-formulation is the product of a "reformulation" process of the source text: the scientific study. Being the latter an extended work presented on a significant number of pages, clearly the information given in a one-pager format of a PR must be condensed and some parts of it may have to be erased for the sake of others.

Our interest in the study of PRs derives from the consideration of the hybrid position of PRs along what Catenaccio (2008) calls the "informative-promotional continuum" (p. 9, but see also Bhatia, 2004; Fairclough, 1992), in that they carry characteristics of both news and ads. In addition, the hybridity of PRs as a genre is reinforced by an allure of opacity, since their authors do not claim authorship. What is more, when PRs

become news in the hands of journalists, not only is their ideal authorship totally lost, but they also directly move into the realm of news articles as a genre, even when no single word is changed. With university PRs, the subtle line between PRs and news articles is even thinner as the “press release” caption is often omitted from university websites, and - when present - it may be replaced by other labels like “news”, “story”, “news release”. Consistently with this, in a recent study, Di Ferrante et al. (2021) showed how academic PRs tend to overshadow external collaborations to their own benefit. This was evident not only from the more frequent mention of authors affiliated with the university issuing the PRs, even when they are not first or corresponding authors, but also from the strategic omission of external participation and the use of vague language to convey content.

The general objective of this investigation is to identify the content elements that are preserved from the scientific articles that the PRs reformulate. This aspect was already explored in other studies (see for example Di Ferrante et al., 2021; Yavchitz et al., 2021) and approached through different perspectives; all such studies highlighted the use of what Yavchitz et al. (2021) define the *spin*, i.e., the twist of information to the advantage of the institution issuing the PR. Nevertheless, no investigation to our knowledge has been carried out with regard to academic PRs in the field of phraseology.

The objective of the present research is addressed through the analysis of n-grams for several reasons. Firstly, it has been shown how the textual essence of PRs lies within the pre-formulation of chunks to be reused in news articles. Secondly, Catenaccio (2008, p. 24) outlined some distinctive moves of PRs: press/news release caption, contact details, and company logo, as well as a headline, lead, and some form of expansion with evaluative elements. These peripheral but meaningful features are mostly vehicled by patterned language with relative syntactic and semantic stability. What is more, Biber et al. (1999) point out how lexical bundles (also referred to as n-grams, clusters, or lexical phrases) can be considered extended collocations regardless of their structural status and they claim that these are “the sequences of words that most commonly co-occur in a register” (p. 989). This is why lexical bundles have become a powerful magnifying glass to spot systematic patterns of use in a genre and, potentially, in its hybrid nuances. Also Barbieri (2018, p. 251) argued that lexical bundles “can reveal a great deal about the unique linguistic characteristics and communicative functions shaping registers”.

Our direction is evermore strengthened by the evidence brought about by Gray and Biber (2015, p. 135), who highlight how lexical phrases in academic writing not only serve common discourse functions like expression of stance and discourse organization, but also carry out peculiar functions of local referential framing, related to research orientations. Cortes (2013) identified the peculiarity of phraseological academic discourse by segmenting published research articles into moves and steps. She found that the most recurring bundles were those associated with three moves which are distinctive of the genre: claiming the relevance of the field, enhancing the gap the study aims to fill, and showing the purpose of the study itself. In line with that, Hyland (2012) analyzed variation across four different disciplines (electrical engineering, science, business, and applied linguistics) by subdividing lexical bundles into three labels: “participant-oriented”, “text-oriented”, and “research-oriented”, which make direct reference to the textual context, but also to physical or abstract entities, such as location, procedures, quantification, and so on. He saw that research-oriented bundles were particularly used in discourse related to engineering and science, which has been the object of investigations highlighting how science popularization might manipulate information, predisposing to bias and misinterpretation (see for example Bratton et al., 2019; Sumner et al., 2014; Woloshin et al., 2009).

If the previous and other studies (Biber & Barbieri, 2007; Biber et al., 2004; Csomay, 2013; Nesi & Basturkmen, 2006) looked at lexical phrases in academic research, to our knowledge there is no investigation to date on the use of n-grams in its popularized counterpart, academic PRs—and very little within the realm of PRs in general. De Cock and Granger (2021a, 2021b) studied key bundles in corporate PRs, by comparing them with business news reporting. What they found is that the two genres make use of distinct units, coherently with the distinct communicative purpose of each genre. That is to say that the n-grams in business news reporting tend to cover informing functions prevalently, as opposed to corporate PRs, which instead cover both informational and persuasive functions. In other words, PRs make use of more self-centered, well-routinized n-grams, to bring forth the intentions and promises of the firm and to stress out positive elements.

The scope of the present analysis is limited exclusively to 5-word n-grams for the following reasons. Biber et al. (1999) claimed that the lexical bundles which mostly pertain to academic prose are “more commonly parts of noun phrases and prepositional phrases” (p. 991) and “four-word, five-word, and six-word bundles are more phrasal in nature” (p. 992), so we assumed they were also more prevalent in academic prose. Of these three types, we opted for five-word n-grams, given their “considerable positional flexibility in formulaic sequences” (Hyland, 2012, p. 152), a phenomenon that Biber et al. (1999) outlined when describing the various pattern combinations in long n-grams. They stressed that commonly, “lexical bundles of a given length are incorporated into longer lexical bundles”, for example, “the four-word bundle *I don't know what+* is incorporated into two different five-word bundles: *I don't know what to+* and *I don't know what it+*. These five-word bundles are in turn incorporated into six-word bundles: *I don't know what to do* and *I don't know what it is*” (p. 1001). As a consequence, in order to have as close as possible a view of the presumably average length of the most recurrent typical n-grams in academic PRs, we opted for 5-word units. However, we must also add that this decision is mostly based on practical rather than theoretical considerations because, as Gray and Biber (2015) claim, there are yet no specific recommendations to address the length of lexical phrases in research (p. 135-136).

## THE STUDY

The current sections outline the research questions and the methodological approach for data analysis, followed by the phases along which it progresses. In particular, choices are detailed and motivated for the analysis of key n-grams in university PRs in order to investigate the emergence of any patterns in the use of typical multi-word combinations characterizing this discourse type. Specifically, this study is a qualitative analysis of phraseology in PRs issued by universities and follows up on previous research by expanding on the study of language use in university PRs in relation to their status as a hybrid genre (Di Ferrante et al., 2021). By drawing on a corpus of University PRs, 5-word n-grams are extracted to explore their main discourse functions. The study moves from three main research questions:

1. Are there any typical n-grams in university PRs?
2. What are the main overlapping areas between PRs and the typical structure of scientific articles?
3. What are the main purposes of typical n-grams?

### Methodology–The Corpus

The analysis is based on the University Research Press Release (UNREP) corpus, which is a developmental collection of University PRs issued in English. It currently consists of 68,913 total words and 90 University PRs that disseminate findings of published scientific articles. All the PRs were retrieved in the news/PR sections of 64 Universities from all over the world. Corpus compilation involved several steps. While a complete version of UNREP was stored for qualitative analysis, several elements were eliminated for the corpus to be later inputted into corpus analysis software. The first phase involved the deletion of all images and figures, as well as graphic elements such as menus, logos, page names, and layout ornaments, which frequently featured in the original PRs. Written text was also deleted when it consisted of notes to editors, notes to journalists, social media sharing buttons, and captions to pictures and figures. Although they form a structural part of the PRs (see Catenaccio, 2008), media contacts were similarly eliminated because they mainly included names, email addresses, and telephone numbers. Finally, most PRs were followed by links to related news and topics that were also deleted as they do not belong to the PR texts. Where present, boilerplates were preserved as they are a structural component of the PR genre and contain relevant information.

### Methodology–Data Analysis

Data analysis drew on both quantitative methods and follow-up, qualitative observations on a set of target n-grams. The selection of relevant n-grams proceeded along three phases:

- a) extraction of keyword n-grams from the corpus;
- b) definition of categories and functions for grouping and classifying n-grams; and
- c) interrater reliability for final n-gram selection.

**Table 1.** Categories for the classification of n-grams

Category	Operational definition
Research group	The phrase refers to one or more people involved in the research group (full listing, full affiliation, and expertise).
Publication procedure and venue	The phrase refers to the publication process and outcome (peer review, editors' review, non-peer reviewed, scientific panel review, and specific journal).
Facilities	The phrase refers to the actual facilities where research was carried out (laboratories, research centers) or indicates the affiliation of the mentioned researchers/professors/authors. Sponsors are also included in this category.
Study object	The phrase explicitly points to general issues concerning the object of a study (hypothesis, thesis statement, research question, scope and confines).
Methodology (if experimental)	The phrase refers to the methodology followed in the study (analysis protocol, alternative. If experimental: features of experimental subject(s) or object(s)).
Features of experimental subject(s) or object(s)	The phrase refers to characteristics of the participants in the study.
Innovation	The phrase refers to the novelty of the contribution (new perspective, findings, and conclusions).
Implications	The phrase refers to the possibility for data to be generalized to wider populations/objects (inferential procedures, limitations).
Not relevant	The phrase refers to the specific topic of the press release or matters that are too general or non-specific enough to be classified.
Not enough data	The researcher is not able to identify function. Function is strictly context-bound.

In the first phase, we analyzed the UNREP corpus through the Sketch Engine software (Kilgarriff et al., 2014; <http://www.sketchengine.eu>) in order to obtain token frequencies and n-grams. In particular, we extracted 5-grams through the keywords function. In phase two, we used metadata to classify the n-grams based on the structural elements of academic discourse and drawing on previous studies on scientific communication and academic PRs. Starting from these premises, and following preliminary observations and close reading of the PRs in the corpus, we identified eight categories corresponding to elements of scientific discourse: research group, publication procedure, facilities, study object, methodology, features of experimental subject/-s or object/-s, innovation, implications. Two extra categories, “not relevant” and “not enough data”, were added to include all cases in which the unit was too unclear to categorize, too specific to the PR topic, or too strictly linked to the surrounding linguistic context. The categories, along with their respective operational definitions, are listed in **Table 1**.

The items were also classified according to their function. Functions are based on the assumption that PRs are a hybrid genre, and were identified and operationally defined, as follows (**Table 2**). The labels “not relevant” and “not enough data” were also included in the functions list.

Phase three involved the attribution of a category and a function to the extracted five-word key n-grams with a frequency of occurrence of at least three tokens in the corpus (which correspond to a frequency of 43 per million words<sup>1</sup>), leading to a total of 61 items. Our aim was to identify lexical and phrasal units that were expected to account for the structural and procedural requirements of scientific research design and publication. Four raters (three of whom are the authors of this paper) worked independently assigning categories and functions to each of the target five-grams. Subsequently, each one of the authors analyzed the outcome independently. All the items selected by the four researchers underwent interrater reliability assessments. Subsequently, the selected items with full (100%) and 75% agreement rate were retained. 15 items were left out as they did not fit any of the relevant categories (being classified as either not relevant or not providing enough information for classification) or showed lower agreement rate, resulting in a total of 46 target five-grams. A discourse analysis was carried out on these target items by looking at their broader syntactic co-text and concordances.

<sup>1</sup> Studies have usually set a threshold frequency ranging from 10 to 40 times per million words (Biber, 2006; Biber et al., 1999, 2004; Hyland, 2008a, 2008b). However, research drawing on smaller corpora often sets lower cutoffs (DeCock, 1998; Nesi & Basturkmen, 2006).

**Table 2.** Functions for the classification of n-grams

Function	Operational definition
Promotion	The phrase has the function of highlighting the expertise and scientific prestige of the study—and of the actors who made the study possible—by mentioning publication venues, the proper names of scholars and universities/institutions involved, or academic and professional roles.
Informativeness	The phrase has function of pointing at new data the study offers to current knowledge.
Explicit audience engagement	The phrase has the function of explicitly engaging the readers.
Not enough data	The researcher is not able to identify the function. The function is too dependent on the context to be identified.
Not relevant	The phrase is not relevant to academic phraseology.

**Table 3.** Target n-grams after interrater reliability and total occurrences in the UNREP corpus

Item	F	Item	F
published in the journal of	9	the study's lead author and	3
according to a new study	8	in the journal science advances	3
the research was supported by	6	the ub school of management	3
nyu school of global public	5	the electric power research institute	3
school of global public health	5	a new study by researchers	3
research was supported by the	5	faculty of arts and social	3
author of the study and	5	a co-author of the study	3
researchers at the university of	5	of arts and social sciences	3
in the journal of marketing	4	was recently published in the	3
assistant professor of marketing at	4	to new research from the	3
in the journal of consumer	4	assistant professor at the university	3
professor of marketing at the	4	according to new research from	3
a co-author of the paper	4	at the university of notre	3
an assistant professor at the	4	the department of computer science	3
the study also found that	4	lead author of the study	3
to a new study by	4	of the university of pennsylvania	3
the university of notre dame	4	researchers at the university of	3
the national academy of sciences	4	by a grant from the	3
at nyu school of global	3	university and the university of	3
group was shown images of	3	the university of southern california	3
barber faculty of arts and	3	professor in the department of	3
was shown images of the	3	professor at the university of	3

## RESULTS

### Results: 5-Grams Categories and Functions

As mentioned above, the five keyword n-grams were extracted and ranked by their frequency of occurrence in the UNREP corpus. Those which occurred three or more times were analyzed individually by the researchers who attributed each item to one of the eight categories listed in [Table 2](#). The four raters' decisions were finally compared: only the items with a 100% and 75% agreement were selected for the subsequent analysis. Moreover, those items that were unanimously rated as "not relevant" and "not enough data" were discarded. The resulting 5-grams are the focus of the analysis and are listed in [Table 3](#).

The next step consisted in looking at the distribution of the n-grams in the different categories, as shown in [Table 4](#). This was aimed at understanding what elements of scientific discourse are given prominence in PRs. It should be noted, though, that some elements connected to specific aspects of the methodology or the object of the study at hand might have been lost: as a matter of fact, 5-word stretches were often not transparent enough to be attached to one of the categories; it turns out that strings like "they were more likely to" or "how we use social media", which were high in terms of keyness, would be rated with "not enough data". While this clearly is a limitation of this type of approach, these pieces of information could still be retrieved through a more in-depth, qualitative analysis.

**Table 4.** Distribution of the n-grams within the categories

Categories	Absolute frequency	Percentage
Facilities	18	39.2%
Research group	13	28.2%
Publication procedure	7	15.2%
Innovation	6	13%
Methodology	2	4.3%
Study object	0	0%
Features of experimental subject(s) or object(s)	0	0%
Implications	0	0%
Total	46	100%

The analysis shows that most 5-grams are connected to four main general categories among those listed in [Table 2](#), i.e. research group, publication procedure/venue, facilities (which also includes sponsors), and innovation (novelty of the findings), as shown in [Table 4](#).

In particular, 39% of these n-grams are to be addressed to the facilities category, which includes units like “the university of Notre Dame” or “The UB school of management”. These types of units are very typical of research-based PRs, where the name of the University is used repeatedly for multiple purposes, such as providing information about authors’ affiliation or stating ownership of the study; the repetition of universities’ names has also been found to serve a general self-promotion goal of the institution issuing the PR (Di Ferrante et al., 2021).

The second most populated category is research group (28%): on the one hand, this datum is consistent with the genre of PRs that have a strong informative component and hence cannot overlook the so-called *who* of Lasswell’s (1948) five Ws. On the other hand, mentioning research authors is also functional to promotion: PRs are often provided with colored pictures of the authors affiliated with the university issuing the PR, regardless of their role in the study.

15% of the five keyword n-grams were attributed to the category of publication procedure/venue. This element frames research output in several ways<sup>2</sup> and the peer-review process or the prestige of the venue where the article is published are quite prominent features. Interestingly, this process, which does not tend to be portrayed in research studies, is used quite consistently across university PRs. Also in this case, we think that this is functional for two purposes. Firstly, it provides fundamental information to the reader who intends to learn more about the topic of the research by retrieving the original scientific article. In this regard, most press releases include a link to the original research—and, interestingly enough, “hyperlink” is the most frequently occurring single keyword in the corpus. Secondly, a scientific journal is a reputable venue for research publication and making reference to it supports the idea that such research is relevant and newsworthy.

The fourth category by number of connected strings is Innovation and novelty of the findings or innovativity of the research. This category includes 13% of the n-grams. Research-based PRs usually present articles that have just been published or are about to be, so the emphasis on the novelty of the study was expected in this type of text. Only 4% of the n-grams pertain to methodology, while no 5-grams emerged related to Implications, the study object, and features of experimental subject(s) or object(s).

As far as functions are concerned, they are often inseparable in that both promotional and informative aims might coexist within the same unit—especially those which tend to be directed towards a self-centered goal that enhances the university’s involvement (research group, affiliation, publication procedure/venue).

For this reason, raters agreed, beforehand, to rate the unit by the prevailing function of the n-grams. Hence, the functions should not be interpreted as mutually exclusive and exhaustive. The rating outcome is presented in [Table 5](#).

<sup>2</sup> For example, the academic journal’s logo or name are on the first page of the published article, sometimes anonymous reviewers are thanked in the acknowledgement sections of articles, etc. Also, it should be noted that very often articles are majorly or minorly modified according to reviewers’ comments. Nonetheless, this stage of academic discourse is not explicitly narrated anywhere in the articles.

**Table 5.** Distribution of the n-grams by functions

Prevailing function	Absolute frequency	Percentage
Promotion	39	84.8%
Informativeness	7	15.2%
Explicit engagement	0	0%
Total	46	100%

noun (4,039 items | 24,608 total frequency)

Lemma	Frequency ? ↓	Lemma	Frequency ? ↓	Lemma	Frequency ? ↓	Lemma	Frequency ? ↓
1 [url]	963 ...	14 school	109 ...	27 student	81 ...	40 ad	66 ...
2 hyperlink	914 ...	15 company	103 ...	28 news	80 ...	41 image	66 ...
3 study	438 ...	16 participant	96 ...	29 way	80 ...	42 post	65 ...
4 people	349 ...	17 woman	94 ...	30 year	79 ...	43 covid-19	64 ...
5 medium	270 ...	18 time	92 ...	31 brand	78 ...	44 energy	64 ...
6 research	260 ...	19 health	90 ...	32 use	78 ...	45 state	63 ...
7 researcher	221 ...	20 percent	89 ...	33 user	74 ...	46 country	63 ...
8 university	183 ...	21 impact	89 ...	34 advertising	72 ...	47 change	62 ...
9 professor	136 ...	22 effect	89 ...	35 result	70 ...	48 pandemic	60 ...
10 food	127 ...	23 datum	85 ...	36 child	70 ...	49 friend	60 ...
11 marketing	115 ...	24 finding	84 ...	37 author	70 ...	50 value	59 ...
12 consumer	112 ...	25 group	83 ...	38 science	68 ...		
13 business	111 ...	26 product	83 ...	39 facebook	68 ...		

**Figure 1.** Frequency list for nouns

The wide majority of the target 5-grams (almost 85%) are believed to have a promotional orientation. Only 15% fulfill a strictly informative function. No n-grams in the corpus ascribe to an explicit audience engagement function.

Informative 5-grams include references to the methodology of the study presented in the PR (e.g. “was shown images of the”) or are associated with the Innovation category (“according to new research from”). Although informative and promotional functions often coexist, 15% of the n-grams were rated as informative only. To enforce the investigation, a discourse analysis on larger portions of the texts provides more insights into those passages where the informative and promotional functions coexist within the same phrasal unit.

**Results: Frequent Items and Patterns in 5-Grams**

As observed by De Cock and Granger (2021a), one weakness of the multi-word approach is that while it extracts contiguous words, it fails to pull out discontinuous or non-fixed sequences. However, the authors suggest that “this weakness can be overcome by using the extracted sequences as prompts to explore the range of variants they display”. Hence, we decided to conduct a more detailed qualitative analysis of a set of recurrent items within the 5-grams focusing on their left and right collocates and their specific function in the PR. This seems also consistent with Partington and Morley’s idea (2004) that themes and topics which occur frequently might be hinted at by the investigation of frequently occurring bundles in a genre.

We selected the target items based on the frequency lists for nouns and verbs in the UNREP Corpus. After generating the lists, we extracted frequent items that were also recurrent in our key 5-grams. The target lemmas thus include: nouns *study*, *professor*, *university*; verbs *publish* and *support*—to which we added *say* given its considerably high frequency in the corpus (as the third most frequent verb; **Figure 1** and **Figure 2** show the frequency lists for nouns and verbs respectively, including absolute frequency values).



verb (1,140 items | 10,946 total frequency)

Lemma	Frequency ? ↓	Lemma	Frequency ? ↓	Lemma	Frequency ? ↓	Lemma	Frequency ? ↓
1 be	2,184 ...	14 get	67 ...	27 give	50 ...	40 want	40 ...
2 have	538 ...	15 understand	67 ...	28 become	48 ...	41 create	39 ...
3 say	466 ...	16 take	66 ...	29 come	47 ...	42 improve	37 ...
4 use	247 ...	17 look	65 ...	30 work	47 ...	43 pay	37 ...
5 do	233 ...	18 publish	64 ...	31 feel	46 ...	44 change	36 ...
6 find	193 ...	19 accord	63 ...	32 compare	45 ...	45 ask	36 ...
7 make	162 ...	20 need	62 ...	33 focus	44 ...	46 live	34 ...
8 show	124 ...	21 suggest	61 ...	34 share	44 ...	47 study	33 ...
9 see	115 ...	22 think	61 ...	35 affect	43 ...	48 conduct	32 ...
10 help	78 ...	23 know	61 ...	36 identify	43 ...	49 promote	32 ...
11 lead	78 ...	24 reduce	59 ...	37 develop	42 ...	50 require	31 ...
12 include	76 ...	25 provide	55 ...	38 support	41 ...		
13 increase	68 ...	26 report	55 ...	39 go	40 ...		

Figure 2. Frequency list for verbs

Sample concordance lines for professor

doc#0 /humsci.stanford.edu/], a co-author of the paper who is the Luke Blossom <b>Professor</b> in the [HYPERLINK: https://humsci.stanford.edu/]School of Humanities and
doc#0 'humsci.stanford.edu/) [HYPERLINK: https://humsci.stanford.edu/]Stanford <b>professor</b> of physics and SLAC professor of particle ph [HYPERLINK: https://humsci.
doc#0 _INK: https://humsci.stanford.edu/]Stanford professor of physics and SLAC <b>professor</b> of particle ph [HYPERLINK: https://humsci.stanford.edu/]ysics and astroph
doc#0 .edu/steven-allen) [HYPERLINK: https://profiles.stanford.edu/steven-allen/], <b>professor</b> of [HYPERLINK: https://profiles.stanford.edu/steven-allen] physics at Stanf
doc#0 show that children are not protected from online marketing of alcohol," said <b>Professor</b> Theresa Marteau, Director of Cambridge's Behaviour and Health Research
doc#0 ----- <b>Professor</b> Marteau said that it is the interactive nature of these adverts which may ma
doc#0 ration files cries out for greater transparency," said Huber, the Forst Family <b>Professor</b> of Political Science in the Faculty of Arts & Sciences. </s></s> "Our work sh
doc#0 [HYPERLINK: http://www.human.cornell.edu/people/fb38])assistant <b>professor</b> of fiber science and apparel design and director of the Body Scan Researc
doc#0 j [HYPERLINK: https://publichealth.nyu.edu/faculty/marie-bragg], assistant <b>professor</b> of public health nutrition at NYU School of Global Public Health and assista
doc#0 public health nutrition at NYU School of Global Public Health and assistant <b>professor</b> in the Department of Population Health at NYU Langone. </s></s> "We nee
doc#0 "ERLINK: https://publichealth.nyu.edu/faculty/jennifer-pomeranz], assistant <b>professor</b> of public health policy and management at NYU School of Global Public He
doc#0 ople/eklein], [HYPERLINK: https://cph.osu.edu/people/eklein] an associate <b>professor</b> of public [HYPERLINK: https://cph.osu.edu/people/eklein] [HYPERLINK:
doc#0 p and so it can't become a deciding factor in their shopping choices," says <b>Professor</b> Achim Spiller, Head of the working group "Marketing for Food and Agricultu
doc#0 ced consumer creativity, according to new research co-written by business <b>professor</b> Ravi Mehta. ----- F
doc#0 s.edu/profile/ravi-mehta/) [HYPERLINK: https://business.illinois.edu/ba/], a <b>professor</b> of [HYPERLINK: https://business.illinois.edu/ba/]business [HYPERLINK: ht
doc#0 mselves and the person in the post," said Benjamin Johnson, an assistant <b>professor</b> of advertising in the University of Florida College of Journalism and Commi
doc#0 "ERLINK: https://www.uvm.edu/gund/profiles/meredith-1-niles], an assistant <b>professor</b> of Nutrition and Food Sciences [HYPERLINK: https://www.uvm.edu/cals/nf
doc#0 ted as a major definer of female identity," said report lead author Associate <b>Professor</b> Teresa Davis, of the University of Sydney Business School.
doc#0 the most recurring images of femininity in post-war advertising," Associate <b>Professor</b> Davis said. -----
doc#0 " ----- -- Associate <b>Professor</b> Teresa Davis -----

Figure 3. Sample concordance lines for professor

**Nouns: Professor**

Professor is a very relevant word in the UNREP corpus. Not only is it included in the first ten most frequent nouns in the corpus (136 raw occurrences, i.e. 197 hits per 100,000 words), but it is also frequent in the list of 5-word n-grams (25 n-grams)<sup>3</sup> and belongs in the research group category.

The noun *professor* was looked at in its context of use through a concordance analysis. The ultimate goal was to identify its patterns of use within the target PRs.

The analysis revealed that *professor* is usually associated with determined entities which show considerable positional flexibility, occurring both to the right of *professor* or to its left. Some linguistic items have also been found to consistently collocate with it (Figure 3).

<sup>3</sup> For this reason, *Professor* is here analyzed instead of 'researcher', which is more frequent in the UNREP corpus overall, but shows fewer occurrences in the target 5-grams (11 occurrences).

**Table 6.** Relevant patterns for professor

Professor: Syntactic context to the left			Professor: Syntactic context to the left		
a	<b>Professor</b>	[PROPER NAME]	g	[PROPER NAME]	<, a> <b>Professor</b>
b	<b>Professor</b> <at>	[UNIVERSITY NAME]	h	[UNIVERSITY NAME]	<b>Professor</b>
c	<b>Professor</b> <of>	[DISCIPLINE]	i	[DISCIPLINE]	<b>Professor</b>
d	<b>Professor</b> <in>	[COLLEGE/DEPARTMENT/SCHOOL]	l	<Said co-author>	<b>Professor</b>
e	<b>Professor</b> <and>	<co-author>	m	<Said>	<b>Professor</b>
f	<b>Professor</b> <and>	<research scholar>	n	<is a>	<b>Professor</b>
			o	<Assistant/Associate>	<b>Professor</b>

**Table 6** shows relevant patterns/co-text to the left and to the right of this word, which appears in 2-to-5-word n-grams. The items were organized in contiguous columns in the table to better show the symmetric patterns between some of them.

In **Table 6**, the linguistic co-text of *Professor* is in angle brackets (<>) when reported verbatim; the items in square brackets specify the type of co-text when it is characterized by high variability (as in the case of Professors' and Universities' names).

As it was mentioned above, one interesting result is the positional flexibility of some of the recurring patterns. As it can be observed in **Table 6**, the right-hand patterns in a., b., c., mirror g., h., and i. in the left-hand patterns, with the addition or deletion of the appropriate syntactic material. Specifically, it was found that *professor* is either followed by the Professor's proper name, as in "Professor Theresa Marteau", or preceded by the proper name, a comma, and the indefinite article: "Jeremy Baileson, a Professor (...)".

The pattern *Professor* + PREPOSITION (patterns b., c., and d.) is quite frequent and the three most common prepositions are *of*, *in*, and *at*, with the following distribution in the whole corpus (frequencies are normalized per 100,000 words): *Professor* + *of* -112 occurrences; *Professor* + *in*-23 hits; *Professor* + *at*-12 occurrences. All prepositions mainly introduce disciplines, roles and institutions. "At" usually collocates with the name of an institution, "in" with the name of a College, Department, or School and "of" with the name of the discipline - as in "Professor of communication".

As shown in patterns b. and h., *Professor* co-occurs with proper names of institutions, both as right and as left modifiers (e.g. "UCLA Professor"); when on the right, the name of the university is preceded by the preposition "at", as in "Professor at New York University". When followed by the conjunction "and", *professor* usually collocates with "research scholar" and "coauthor". As far as additional left-hand linguistic collocates are concerned, *professor* might be preceded by labels indicating academic rank, such as "assistant" or "associate". Frequent left modifiers also include present and past forms of the verb *to say*, which is used to introduce or end direct quotations. Consistently with this, it should be noted that *say* is the third most frequent verb in the corpus, just after *be* and *have*-and is discussed in more detail below.

In order to broaden the analysis of the context of use, we also looked at longer stretches of text where these linguistic patterns would appear. Firstly, *professor* is used to describe the research group composition, as shown in example 1 below:

#### Example 1: Research groups' composition

- 1a. The research group consisted of **Professor** Ohtsubo Yohsuke
- 1b. In a paper co-authored by Gaurav Jain, an assistant **professor** of marketing in the Lally School of
- 1c. The research, with co-authors **Professor** Grahame Dowling and Dr Edward Wei

Secondly, *professor* is often found in contexts where a verbatim quote from a researcher is introduced; in such contexts, *professor* collocates with a form of the verb *say* either as a verb phrase or as discontinuous frames as shown in example 2 below:

#### Example 2. Direct quotations

- 2a. everywhere else," **said** co-author Peter Bobkowski, assistant **professor** of journalism. "It's a reflection of how
- 2b. wealthy need alternative ways to signal their status," **said Professor** Berger.
- 2c. are 'green materialists,'" **said** Helm, an associate **professor** in the Norton School of Family and Consumer Sciences

black hole at the center of a galaxy 800 million light-years away, Stanford University astrophysicist [HYPERLINK: https://profiles.stanford.edu/wilkins] Dan Will  
 Co-authors of this research are from Saint Mary's University (Canada), Netherlands Institute for Space Research (SRON), University o  
 s University (Canada), Netherlands Institute for Space Research (SRON), University of Amsterdam and The Pennsylvania State University. This work  
 Research (SRON), University of Amsterdam and The Pennsylvania State University. This work was supported by the NASA NuSTAR and XMM- New  
 XMM- Newton Guest Observer programs, a Kavli Fellowship at Stanford University, and the V.M. Willaman Endowment at the Pennsylvania State University.  
 I University, and the V.M. Willaman Endowment at the Pennsylvania State University. Research -----  
 use, age restrictions. Researchers from RAND Europe and the University of Cambridge investigated the marketing campaigns of five alcohol compa  
 their registration." The study's co-authors are Marc Meredith of the University of Pennsylvania, Yale Law School graduate Michael Morse, and Katie Ste  
 rvania, Yale Law School graduate Michael Morse, and Katie Steele of the University of Pennsylvania. Augmented reality can improve online shopping  
 Baytar and collaborators from Iowa State University and Virginia Polytechnic Institute focused on the latter technology, asking  
 is for the environment or for society. A team of scientists from the University of Göttingen has investigated to what extent factors which affect consume  
 n the working group "Marketing for Food and Agricultural Products" at the University of Göttingen studied how two groups - one from Germany and one from th  
 still serve a purpose. According to new research co-written by a University of Illinois expert in new product development and marketing, resource sca  
 post," said Benjamin Johnson, an assistant professor of advertising in the University of Florida College of Journalism and Communications who led the study. <  
 www.uvm.edu/cals/nfs [HYPERLINK: https://www.uvm.edu/cals/nfs] at the University of Vermont and a fellow at the university's [HYPERLINK: https://www.uvm.  
 s://www.uvm.edu/cals/nfs] at the University of Vermont and a fellow at the university's [HYPERLINK: https://www.uvm.edu/gund] Gund [HYPERLINK: https://www  
 s://www.uvm.edu/gund]. [HYPERLINK: https://www.uvm.edu/gund] Led by University of Vermont researchers, the study examines diet diversity among 107,000  
 [HYPERLINK: https://www.uvm.edu/gund/profiles/molly-brown] of the University of Maryland. "If it's more erratic and intense, as is predicted with c  
 ded Meredith Niles, Taylor Ricketts, Brendan Fisher and Serge Wiltshire ( University of Vermont), Molly Brown (University of Maryland), and Benjamin Emery  
 Brendan Fisher and Serge Wiltshire (University of Vermont), Molly Brown ( University of Maryland), and Benjamin Emery

Figure 4. Sample concordance lines for university

Table 7. Relevant patterns for university

University: Syntactic context to the left		University: Syntactic context to the right	
a	<b>University</b> <of> [UNIVERSITY NAME]	g	[UNIVERSITY NAME] <b>University</b>
b	<b>University</b> <at> [SCHOOL NAME]	h	<at>/<from> [UNIVERSITY NAME] <b>University</b>
c	<b>University</b> <s> [COLLEGE/ DEPARTMENT/ SCHOOL]	i	<the> <b>University</b>
d	<b>University</b> [DIVISION/ LABORATORY/ SCHOOL]	l	[PERSON ROLE] <in the> <b>University</b>
e	<b>University</b> [PERSON ROLE] <professor> / <researcher>		
f	<b>University</b> [PERSON, DISCIPLINE] <astrophysicist>		

Thirdly, *professor* is used in sentences that are syntactically coordinated with other sentences containing *lead*, used either as a noun or as a verb. Noun phrases like “lead author”, “corresponding author” or “lead supervisors” are used as synonymous phrases for *professor*. *Lead* ranks 55th in the adjective frequency list of the corpus, with 21 occurrences (30 hits per 100,000 words); in most cases, it collocates with *author* (19 occurrences out of 21). Furthermore, *lead* is ranked 11th in the verb frequency list and is used mostly in the passive voice, while *professor* appears in the by-phrase. As a matter of fact, elicitation of the leading roles of the mentioned professor in the research project is a relevant aspect of PRs and a promotional strategy.

Example 3. Leading role

- 3a. engage,” said **lead author** Chris Vargo, an assistant **professor** of Advertising
- 3b. study's **lead author** and an assistant **professor** of marketing at the USC Marshall School
- 3c. according to a study **led** by a Texas A&M university **professor**
- 3d. A new study **led** by an MIT **professor**

Nouns: University

The noun *university* shows 183 raw occurrences in the corpus (266 occurrences per 100,000 words). Typically, it appears within n-grams related to the research group and facilities categories—more specifically, 16 facilities-related 5-grams and 14 research group related ones, for a total of 30 five-grams. Figure 4 shows some concordances for the item *university*, while Table 7 illustrates typical syntactic patterns.

Most of the patterns developing both to the left and to the right of the noun typically identify the institution’s name (e.g. “the University of Edinburgh”, “Rutgers University”).

Considering right-hand collocates, out of 183 occurrences 101 are part of the pattern “University + of”, followed by the proper name of the university (147 out of 266 occurrences per 100,000 words). In fewer cases, right modifiers define a more specific institution within the university (“University Office of the Provost”, “University Division of Marketing & Communications”) or refer to roles (“university researchers”, “university neurobiologists”). The genitive construction “university’s” usually precedes the names of departments and institutes within the university (“university’s Institute for Environment”).

The university’s name and the roles of its affiliated scholars also expand to the left in the corpus. One obvious case includes such instances as “Yale University” or “Columbia University”, which are often embedded in prepositional phrases introduced by “at” and “from”.

Most prepositional left modifiers link the noun *university* to its researchers, giving rise to such patterns as “researchers from the University of Pennsylvania” or “scientists from the University of Oxford”. Finally, the definite article *the* commonly introduces the noun *university* when naming specific institutions (“the University of Technology Sydney”).

Longer stretches of text around the word *university* often aim at highlighting the prestige of a given institution, as evident in example 4:

**Example 4. Prestige of the institution**

4a. The **University** of Kansas is a **major** comprehensive research and teaching **university**

When co-occurring with *study*, *university* also appears in strings that state or remark ownership of the research, as in example 5 below:

**Example 5. Stating ownership**

5a. a new Ohio State **University** study suggests

5b. such as depression and anxiety, finds a new Michigan State **University** study.

Finally, and as observed above, *university* is found in contexts, where scholars are linked to their specific affiliations (example 6):

**Example 6. Link between scholar and affiliation**

6a. a fellow at the **university’s** Gund Institute for Environment

6b. professor of communication at North Carolina State **University**

6c. according to a study led by a Washington State **University** researcher

6d. colleagues from Yale **University** and Arizona State **University**

**Nouns: Study**

The word *study* has 438 raw occurrences (636 per 100,000 words). It occurs within 34 5-grams, most of which ascribe to the novelty of the findings category (20) (Figure 5 and Table 8).

As already mentioned, PRs in our corpus all revolve around new research being carried out and published. Therefore, *study* is often repeated multiple times within each PR, with an average of 4.8 occurrences per individual text. As far as the right context of *study* is concerned, the string “<a new> study + by” occurs 10 times per 100,000 words (seven raw occurrences), whereas in terms of right-hand context, *study* often collocates with the preposition “by” followed by the role of the authors as in “study by researchers at Oxford’s Leverhulme Centre”. Similarly frequent is also “study led by” (five raw occurrences and seven occurrences per 100,000 words), which is usually followed by the proper name of an author or by a common name preceded by the name of the university as in “according to a study led by a Washington State University researcher.”

The 5-gram “the study <also found that>” shows six hits per 100,000 words (four raw occurrences). In line with the category, n-grams featuring the noun *study* generally highlight the novelty of the contribution, as evident in example 7:

**Example 7. Novelty of the contribution**

7a. according to **a new study** by researchers in the University at Buffalo School

7b. **a new study** by researchers at the University of Georgia suggests

... use requires a flattening of global and national inequalities. In our study, we allow for inequalities in energy consumption only where need dictates. Areas you couldn't on Facebook," said Eleanor Winpenney, a co-author of the study from RAND Europe, who also pointed out that you don't need to be logged in the study, recently published in the journal [HYPERLINK: In the study, co-author. Marteau, Director of Cambridge's Behaviour and Health Research Unit, a study suggest that either the CAP Code is not being followed, or these guidelines percentage of children accessing these social media sites, the results from this study showed that fans of the products engaged with the sites via 'Likes', this included promotional videos, recipes, games and competitions. The study showed that fans of the products engaged with the sites via 'Likes', this Study uncovers flaws in process for maintaining state voter rolls whether people have moved to prune their voter rolls. A Yale-led study of this process in Wisconsin shows that a significant percentage of registered changed addresses, potentially endangering their right to vote. The study, published in the journal Science Advances [HYPERLINK: ial address of registration after the state marked them as having moved, the study showed. The findings suggest that states should more clearly com here twice as likely to be incorrectly identified as having moved. The study likely undercounts the number of registered voters incorrectly listed as havin registration before the poll books were published. In addition, the study examined low-turnout elections, making it likely that many people wrongly lie effort on contacting people at risk of losing their registration." The study's co-authors are Marc Meredith of the University of Pennsylvania, Yale Law School of Pennsylvania. Augmented reality can improve online shopping, study finds By E.C. Baytar, "Evaluating Garments in Augmented Reality When Shopping Online ( Results for the AR technology were mixed. According to Baytar, study participants were able to find the right size with AR try-on, but had trouble ev toward the AR garment and purchase intentions. As expected, the study showed that higher telepresence corresponded to more positive attitudes toward garment. The study was completed in 2016 but has new significance, Baytar said, with brick and rol of Global Public Health and NYU Grossman School of Medicine in a new study published in the journal Pediatrics. Food and beverage companies spend \$1.6 billion on task of having young kids at home because of COVID-19," said Bragg, the study's senior author.

Figure 5. Sample concordance lines for study

Table 8. Relevant patterns for study

Study: Syntactic context to the left		Study: Syntactic context to the right	
a	study <by> [PERSON ROLE]	g	<a new> study
b	study <by> [PERSON ROLE] <in/><at> [UNIVERSITY NAME]	h	<according to> <a new> study
c	study <by> [UNIVERSITY NAME] [ROLE]		
d	study <led> <by> [PERSON ROLE]		
e	study <published in > [JOURNAL NAME]		
f	<the> study <also found that> [OBJECT COMPLEMENT CLAUSE]		
g	study <and> <research scholar>		

Quite often, these longer stretches include the publication venue (example 8) or enhance the premises or research team involved in the research (example 9):

**Example 8. Introducing the publication venue**

- 8a. according to **a new study published in** Nature Scientific Reports.
- 8b. According to **a new study in the journal** Humanities and Social Sciences

**Example 9. Enhancing a research institution or team**

- 9a. **a new study** by the University of Sussex Business School reveals
- 9b. **a new study** by Penn evolutionary biologists

When occurring in the pattern “the study also found that”, the aim is primarily that of introducing additional inferences and generalizations based on data. This is also documented by the common terms appearing in the surrounding co-text, for example *significantly* or *associated*, as illustrated in example 10:

**Example 10. Introducing additional findings**

- 10a. **The study also found that** businesses’ social posts **significantly** strengthen the
- 10b. **The study also found that** a positive chat bot experience was **associated**

**Verbs: Publish, support, say**

Significant verbs in terms of frequency in the corpus and/or recurrence in our target 5-grams are *publish*, *support* and *say*. *Publish* occurs 64 times in the UNREP Corpus (93 occurrences per 100,000 words) and is found in 20 target 5-grams in its past participle form. All n-grams pertain to the Publication procedure/venue category and seem to have a primarily promotional function. Sample concordance lines for *publish* are shown in Figure 6.

----- In the study, recently **published** in the *J* [HYPERLINK: <http://alcalc.oxfordjournals.org/content/early/2013/11>]

I addresses, potentially endangering their right to vote. </s></s> The study, **published** in the journal *Science Advances* [HYPERLINK: <https://advances.sciencemag.org/>]

ne nonetheless renewed their voting registration before the poll books were **published**. </s></s> In addition, the study examined low-turnout elections, making it I

obal Public Health and NYU Grossman School of Medicine in a new study **published** in the journal *Pediatrics*. Food and beverage companies spend \$1.8 billion c

onal benefit that people feel when they do good. </s></s> The results were **published** in the *Journal of Cleaner Production*, an international scientific publication

unctionality of a given product, thus enhancing product-use creativity, says **published** research from [HYPERLINK: <https://business.illinois.edu/ba/>] Ravi Mehta (

la College of Journalism and Communications who led the study. </s></s> **Publishing** next month in the *Journal of Computer-Mediated Communication*, the stud

ation -- all common global development tactics. </s></s> The findings were **published** today in *Environmental Research Letters*. </s></s> Of the six regions exam

iversity and St Gallen University in Switzerland looked at advertisements **published** in *Australian Women's Weekly* and the UK's *Good Housekeeping* between

----- Their findings, **published** in the *Journal of Consumer Culture*, indicate that while the image of mothe

----- The paper was **published** in the journal *Marketing Letters*.

ia versions of the VR scenario, like text. </s></s> These findings are set to **publish** Oct. 17 in *PLOS ONE* [HYPERLINK: <https://www.frontiersin.org/articles/10.3389/frsus.2020.583090/abstract>],

recent [HYPERLINK: <https://www.pnas.org/content/118/7/e2013391118>] **published** [HYPERLINK: <https://www.pnas.org/content/118/7/e2013391118>] in *Proceedings of the National Academy of Sciences*

----- The study was **published** online recently in the *Journal of Marketing* < <https://journals.sagepub.com/> > [HYPERLINK: <https://www.pnas.org/content/early/2020/06/02/1915516117>],"

----- The researchers studied genealogical data on 40,000 scientists who **published** 1.2 million papers in biomedicine, chemistry, math or physics between 196

és studied under mentors who demonstrated a unique skill in ideating and **publishing** celebrated research and who displayed independent thought by breaking c

e. ----- The study, **published** in the *Journal of Computer Mediated-Communication*, challenges the notio

[HYPERLINK: <http://onlinelibrary.wiley.com/doi/10.1002/smj.2646/full>] is **published** in the current [HYPERLINK: <http://onlinelibrary.wiley.com/doi/10.1002/smj.2646/full>]

Figure 6. Sample concordance lines for publish

and decarbonization, not growth. </s></s> Universal Basic Services could **support** people in a world with less work, allowing people the time to undertake all

allowing people the time to undertake all the unpaid care work required to **support** children, the elderly, and all those in between. </s></s> There's a reluctanc

msterdam and The Pennsylvania State University. </s></s> This work was **supported** by the NASA NuSTAR and XMM-Newton Guest Observer programs, a Ka

alth at NYU Langone. </s></s> "We need a digital media environment that **supports** healthy eating instead of discouraging it." </s></s>

rshonsky, and Omni Cassidy of NYU Langone. </s></s> The research was **supported** by grants from the National Institutes of Health (DP5OD021373-05) and A

a comprehensive pediatric hospital, also in Manhattan, **supporting** a full array of children's health services; N [HYPERLINK: <https://doi.org/10.1002/ajcp.1234>]

the strong association with the common good of the fair-trade label, which **supports** farmers in developing countries. </s></s> "Other studies have shown that c

"Accent "channel support" disclosures, such as "Shop using these links to **support** the channel." </s></s> -----

ork on November 7 in Jersey City, New Jersey. </s></s> The research was **supported** by the National Science Foundation.

first study, 82 percent of participants in the VR condition signed a petition **supporting** affordable housing versus 67 percent of the people who read a narrative th

ly inspiring to see that positive, lasting impact." </s></s> The research was **supported** by a grant from the Robert Wood Johnson Foundation.

----- The research was **supported**, in part, by a grant from the Netherlands Organization for Scientific Resea

likely to tip them or tip them less. </s></s> They may also be less likely to **support** regulations that financially protect these workers.

atyam Mukherjee are co-authors on the paper. </s></s> The research was **supported** by the Northwestern Institute on Complex Systems (NICO), MURI-Defens

A Sheth Foundation Dissertation Award Minton received in 2013 **supported** the research. -----

les, choosing individuals who are loud and confident but lack the ability to **support** their followers' development and success." </s></s>

The project was partially **supported** through a grant from the University at Buffalo Gender Institute.

see a number of airlines not offering this as an option, which is somewhat **supported** by the anecdotal study of known cases on long-haul flights not leading to t

s who were seeing the messages, exposing those who might not explicitly **support** e-cigarettes, or potentially vulnerable populations such as youth." </s></s>

marketing retweets, ranging from those who publicly mentioned use of or **support** for e-cigarettes in their profiles, to others with a basic profile -- who made

Figure 7. Sample concordance lines for support

Common right modifiers include an explicit reference to a publication venue, such as in the strings “published <in>” or “published <in the Journal>”, whereas most frequent objects are “study”, “paper” and “findings”.

The verb *support* shows 41 raw occurrences (59 occurrences per 100,000 words) and is found in 11 of the target 5-grams as part of the passive construction “was supported”. The n-grams ascribe to the facilities category—more specifically, Sponsors. Figure 7 illustrates a set of sample concordances for *support*.

The embedded trigram “was supported by” (14 total occurrences per 100,000 words, 10 raw occurrences) is often followed by specific names of universities, programs, or institutions (“was supported by the National Science Foundation”) or by the unit “<by a grant from>” centering around the word “grant” (“was supported

ie out, so we shouldn't be able to see anything that's behind the black hole,"	said	Wilkins, who is a research scientist at the Kavli Institute for Particle Astrophysics
observe this directly and see Einstein's general theory of relativity in action,"	said	[HYPERLINK: https://profiles.stanford.edu/roger-blandford]Roger Blandford
roduces these high energy electrons that then go on to produce the X-rays,"	said	Wilkins. </s><s> As Wilkins took a closer look to investigate the origin of the
g theoretical predictions of how these echoes appear to us for a few years,"	said	Wilkins. </s><s> "I'd already seen them in the theory I've been developing, s
cing to let us get higher resolution looks in much shorter observation times,"	said	Wilkins. </s><s> "So, the picture we are starting to get from the data at the r
ou could still access the alcohol pages, whereas you couldn't on Facebook,"	said	Eleanor Winpenny, a co-author of the study from RAND Europe, who also pr
sults show that children are not protected from online marketing of alcohol,"	said	Professor Theresa Marteau, Director of Cambridge's Behaviour and Health I
----- Professor Marteau	said	that it is the interactive nature of these adverts which may make them more
hether individuals have moved before they are removed from the voter rolls,	said	Yale political scientist [HYPERLINK: https://politicalscience.yale.edu/people/
maintaining states' voter-registration files cries out for greater transparency,"	said	Huber, the Forst Family Professor of Political Science in the Faculty of Arts &
ntial fraud (although there is little evidence it causes unlawful voting, Huber	says	) and creates headaches for political campaigns, which rely on accurate regi
number of registered voters incorrectly listed as having moved, the researchers	said	, explaining that a significant number of people who did not respond to the p
poll books weren't covered in the analysis because they didn't vote, Huber	said	. </s><s> The researchers are not suggesting that ERIC intentionally targete
ERIC intentionally targeted minorities. </s><s> "There's no malice here," Huber	said	. </s><s> "ERIC wants to help states, but relying on administrative records in
vay to identify mistakes and prevent people from being disenfranchised," he	said	. </s><s> "The state also has same day voter registration, which is another s
"Bracket purchases increase shelving and shipping costs," Baytar	said	. </s><s> "We're hoping these technologies can reduce those costs, while al
fit - particularly at the bust, waist and hips. </s><s> One explanation, Baytar	said	, was that although the image of the dress was taken when the dress was or
s were relatively high, though they were higher for the real garment," Baytar	said	. </s><s> "This is still a promising technology, because AR can provide gooc
The study was completed in 2016 but has new significance, Baytar	said	, with brick and mortar stores being closed during the COVID-19 crisis.
and even four years later the technology still needs to be improved," Baytar	said	. </s><s> "We can expect that as these technologies evolve, people will trus

Figure 8. Sample concordance lines for say

by a grant from the Japan Society for the Promotion of Science"). To the left, we generally find the words "study" and "research" ("The research was supported by the Northwestern Institute on Complex Systems").

The longer stretches of text built around the form "supported" are aimed at stating financial aids and acknowledging the sponsors involved (example 11) or at introducing and discussing the object of the study (example 12).

**Example 11. Stating and acknowledging sponsors**

- 11a. The study **was supported by** the David and Lucile Packard Foundation
- 11b. The research **was supported by** the National Science Foundation
- 11c. The research **was supported by** grants from the National Institutes of Health

**Example 12. Discussing the object of the study**

- 12a. which is somewhat **supported by** the anecdotal study of known cases
- 12b. many such opinions are not **supported by** data
- 12c. The results are **supported by** celebrity endorsement data,

Although it was not part of our 5-grams, the verb say stands out in the UNREP corpus and deserves some attention. With a total of 676 hits per 100,000 words (466 raw occurrences), different forms of say typically collocate with scholars' last names and introduce direct quotations—a typical feature of PRs. "Said" and "says" are the prevailing tokens as in example 13 below (Figure 8).

**Example 13. Framing a direct quote**

- 13a. [...] social media is taking over our face-to-face time," Hall **said**
- 13b. and are expected to become even more common over the next few years," **says** Yang Cheng.

**CONCLUSIONS**

The findings of the mixed-methods analysis of five n-grams aim to contribute to the characterization of PRs as a hybrid genre at the crossroads between news and ads, and to add knowledge about university-issued PRs, which present some of the constitutive elements of scientific discourse while at the same time enhancing the personalization of the making of scientific research.

In response to RQ1, which inquired about the presence of typical n-grams characterizing PRs, in the UNREP Corpus we found a set of 5-grams and other related or embedded shorter units that typify university PRs. Most of these n-grams refer to the research group and its composition, the publication procedure and venue, the facilities involved in the research and its sponsors, and the novelty of the findings. These areas show an overlap between university PRs and the typical structure of scientific articles (RQ2). On the contrary, much fewer multi-word units point at methodological aspects and no 5-grams at all were identified in relation to the specific object of the study, its features, and any Implications derived from it. It should be noted, however, that methodology, object of study, and Implications are perhaps the most distinctive aspects of a scientific article, and therefore the way they are presented most presumably varies across different papers. It turns out that the type of information provided in these PRs seems mostly tailored around the individuals who carried out the research, the institutions that made it possible and the journals that disseminated it. In other words, PRs appear to be predominantly limited to areas connected to the *who* and the *where*, which, incidentally, are also the elements that serve promotional needs. This trend is also evident from a more detailed analysis of the n-grams' main purposes, which always combine an informative function with a more evident promotional intent (RQ3). In particular, what emerges from the data is that the mentions of constituting elements of the PRs fulfill multiple aims. When members of the research team are mentioned (e.g. professors, co-authors, lead authors), this is done not only to describe the composition of the research groups and add relevant information about them but also to highlight and elicit leading roles and to quote verbatim - especially prominent figures in the study. Similarly, the naming of the institutions, repeated multiple times throughout each PRs, is used to indicate the affiliation of the mentioned authors, but also aims to remark ownership of the research, to tone up the value of the sponsoring sources and the actual facilities where the study was carried out. References to the study are certainly functional to inform about its results and to trigger additional implications based on data, as is clearly demonstrated by the abundance of hyperlinks to the original study; in parallel, however, this is also a way to highlight the novelty of the publication and feature the publication venue. Naming supports and sponsors, finally, has a more explicit promotional aim, which is nonetheless strictly intertwined with the informative one, where any financial aid and sponsors involved in a research project are acknowledged.

While the hybridity of PRs as inherently both informative and promotional texts has been the first perspective level through which several studies have approached these types of texts, the present study's exploration of n-grams in university PRs has elicited a further level of ambivalence. On the one hand, this specific type of communication for science dissemination is certainly aimed at informing, but it provides information about-and draws from-one very specific, structured, and codified type of discourse, the academic one. It turns out that structural and content elements of academic discourse are inevitably absorbed into the texture of the language and the moves of the PRs and this is strongly reflected in the n-grams. On the other hand, promotion is bound and pinned by regimented communication: Universities are higher-education institutions and their communication strategies are to be consistent with expectations of authoritativeness, reputability, and trustworthiness. University-issued PRs are characterized by word clusters that reflect these aspects by bringing forth the actors who made the study possible: renowned professors, cutting-edge facilities, and prestigious venues all contribute to the construction and strengthening of the reputation of the institution. A sort of grammar of university press releases, or even a recipe for their drafting, seems to have emerged from the present research, framing these scientific, discourse-derived texts within instances of personalization and precise informational coordinates.

These results might raise concerns about how universities are balancing their ambivalent roles as science producers and funding seekers. Such balance is certainly not easy to pinpoint, as universities cannot afford to avoid promoting their research, sometimes even at the cost of overshadowing their competitors; at the same time, they have an overarching mission to disseminate scientific knowledge in a truthful and reliable way. Universities thus need to keep disseminating science results. To further contrast the spread of misleading information, press releases could include more systematically reader-friendly descriptions of the methodology and procedures used in order to make laypeople progressively more familiar with the scientific method. This would contribute to improving people's critical skills and their ability to discriminate between reliable and unsupported claims.



In this study, we identified some typical n-grams in university PRs and their main purposes. On the one hand, our results offer insights into the specific texture of university PRs that turn out to have internal homogeneity and intrinsic peculiarities. In these terms, further research could prove that university PRs can be considered a sub-genre of the macro-category of PRs. This might be due to two concurring factors: one is the fact that these PRs are very strongly affected by the language of the academic articles they stem from; the other is their very content, namely published scientific research, which is by its nature intangible and hardly marketable - both concretely and ethically. On the other hand, our identification of typical n-grams in academic PRs could be used as a starting point for applied research aimed at developing specific annotation systems for PR texts.

Clearly, these findings are not exhaustive of what is a subgenre of press releases and more work can be done to explore, through closer observation of longer stretches of texts, ways in which, for example, definitions of more complex concepts are provided, or methodology is recontextualized and/or simplified to be more digestible by a public of non-experts. Furthermore, it is possible that some of the n-grams that were discarded because they did not match any category could be retrieved in future studies as they might offer insights into recursive patterns, typical of university press releases, and might offer evidence for new syntactic, lexical, and broadly communicative options for the popularization and communication of science.

**Author contributions:** All authors were involved in concept, design, collection of data, interpretation, writing, and critically revising the article. Specifically, **EP** wrote the following sections: *Linguistic Patterns of Science Dissemination, Methodology – The Corpus, Nouns: Study*, **LDF** wrote the following sections: *Results: 5-grams Categories and Functions, Nouns: Professor, Conclusions*, and **EG** wrote the following sections: *Introduction, Methodology - Data analysis, Results: Frequent Items and Patterns in 5-grams, Nouns: University, Verbs: publish, support, say*. All authors approve final version of the article.

**Funding:** The authors received no financial support for the research and/or authorship of this article.

**Acknowledgements:** The authors would like to thank to Dr. Sergio Pizziconi for contributing to the initial stages of the project and for offering constructive comments and suggestions.

**Declaration of interest:** Authors declare no competing interest.

**Data availability:** Data generated or analyzed during this study are available from the authors on request.

## REFERENCES

- Barbieri, F. (2018). *I don't want to and don't get me wrong*: Lexical bundles as a window to subjectivity and intersubjectivity in American blogs. In J. Jopaczyk & J. Tyrkkö (Eds.), *Applications of pattern-driven methods in corpus linguistics* (pp. 251-275). Benjamins. <https://doi.org/10.1075/scl.82.10bar>
- Bhatia, V. K. (2004). *Worlds of written discourse: A genre-based view*. A&C Black.
- Biber, D. (2006). Stance in spoken and written university registers. *Journal of English for Academic Purposes*, 5(2), 97-116. <https://doi.org/10.1016/j.jeap.2006.05.001>
- Biber, D., & Barbieri, F. (2007). Lexical bundles in university spoken and written registers. *English for Specific Purposes*, 26(3), 263-286. <https://doi.org/10.1016/j.esp.2006.08.003>
- Biber, D., Conrad, S., & Cortes, V. (2004) 'If you look at ...: Lexical bundles in university teaching and textbooks', *Applied Linguistics* 25(3), 371-405. <https://doi.org/10.1093/applin/25.3.371>
- Biber, D., Johansson, S., Leech, G., Conrad, S. & Finegan, E. (1999). *Longman grammar of spoken and written English*. Pearson Education Limited.
- Bratton, L., Adams, R. C., Challenger, A., Boivin, J., Bott, L., Chambers, C. D., & Sumner, P. (2019). The association between exaggeration in health-related science news and academic press releases: A replication study. *Wellcome Open Research*, 4, 148. <https://doi.org/10.12688/wellcomeopenres.15486.2>
- Catenaccio, P. (2007). New(s) genre and discursive identity. The changing face of the press release in the age of the Internet. In G. Garzone, G. Poncini, & P. Catenaccio (Eds.), *Multimodality in corporate communication. Web genres and discursive identity* (pp. 55-72). Franco Angeli.
- 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>
- Clark, B. R. (1998). *Creating entrepreneurial universities: Organizational pathways of transformation. Issues in higher education*. Emerald Group Publishing Limited.

- Cortes, V. (2013). *The purpose of this study is to*: Connecting lexical bundles and moves in research article introductions. *Journal of English for Academic Purposes*, 12(1), 33-43. <https://doi.org/10.1016/j.jeap.2012.11.002>
- Csomay, E. (2013). Lexical bundles in discourse structure: A corpus-based study of classroom discourse. *Applied Linguistics*, 34(3), 369-388. <https://doi.org/10.1093/applin/ams045>
- De Cock, S. (1998). A recurrent word combination approach to the study of formulae in the speech of native and non-native speakers of English. *International Journal of Corpus Linguistics*, 3(1), 59-80. <https://doi.org/10.1075/ijcl.3.1.04dec>
- De Cock, S., & Granger, S. (2021a). Uncovering the typical phraseology of press releases and business news reports using a lexical bundle approach. *EUROPHRAS 2021*. <http://hdl.handle.net/2078.1/259638>
- De Cock, S., & Granger, S. (2021b). Stance in press releases versus business news: A lexical bundle approach. *Text & Talk*, 41(5-6), 691-713. <https://doi.org/10.1515/text-2020-0040>
- Di Ferrante, L., Lattanzi, G., & Petrocelli, E. (2021). Those who get lost in communication science. Promotion and downplay in university press releases. In G. Garzone, W. Giordano, & S. Pizziconi (Eds.), *Discursive perspectives on knowledge dissemination in corporate and professional communication: Focus on ethical and ideological aspects* (pp. 31-56). Editura Politehnica.
- Fairclough, N. (1992). *Discourse and social change*. Polity Press.
- Garzone, G. (2014). News production and scientific knowledge: Exploring popularization as a process. In G. Bongo, & G. Caliendo (Eds.), *Die Sprache der Popularisierung [The language of popularization]* (pp.73-107). Peter Lang.
- Garzone, G. (2021). *Specialized communication and popularization in English*. Carocci.
- Gotti, M. (2014). Reformulation and recontextualization in popularization discourse. *Ibérica, Revista de la Asociación Europea de Lenguas para Fines Específicos [Iberica, Journal of the European Association of Languages for Specific Purposes]*, 27, 15-34.
- Gray, B., & Biber, D. (2015). Phraseology. In D. Biber, & R. Reppen (Eds.), *The Cambridge handbook of English corpus linguistics* (pp. 125-145). Cambridge University Press. <https://doi.org/10.1017/CBO9781139764377.008>
- Hyland, K. (2008a). *As can be seen*: Lexical bundles and disciplinary variation. *English for Specific Purposes*, 27(1), 4-21. <https://doi.org/10.1016/j.esp.2007.06.001>
- Hyland, K. (2008b). Academic clusters: Text patterning in published and postgraduate writing. *International Journal of Applied Linguistics*, 18(1), 41-62. <https://doi.org/10.1111/j.1473-4192.2008.00178.x>
- Hyland, K. (2012). Bundles in academic discourse. *Annual Review of Applied Linguistics*, 32, 150-169. <https://doi.org/10.1017/S0267190512000037>
- Jacobs, G. (1999). *Preformulating the news: An analysis of the metapragmatics of press releases*. Benjamins. <https://doi.org/10.1075/pbns.60>
- Jacobs, G., & Van Hout, T. (2009). Towards a process view of preformulation. In J. Renkema (Ed.), *Discourse, of course: An overview of research in discourse studies* (pp. 239-251). Benjamins. <https://doi.org/10.1075/z.148.24jac>
- Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., & Suchomel, V. (2014). The sketch engine: Ten years on. *Lexicography*, 1, 7-36. <https://doi.org/10.1007/s40607-014-0009-9>
- Lasswell, H. (1948). The structure and function of communication in society. In L. Bryson (Ed.), *The communication of ideas, a series of addresses* (pp. 37-52). The Institute for Religious and Social Studies.
- Linell, P. (1998). Discourse across boundaries: On recontextualizations and the blending of voices in professional discourse. *Text & Talk*, 18(2), 143-158. <https://doi.org/10.1515/text.1.1998.18.2.143>
- Nesi, H., & Basturkmen, H. (2006). Lexical bundles and discourse signaling in academic lectures. *International Journal of Corpus Linguistics*, 11(3), 283-304. <https://doi.org/10.1075/ijcl.11.3.04nes>
- Partington, L. & Morley, J. (2004). From frequency to ideology: Investigating word and cluster/bundle frequency in political debate. In B. Lewandowska-Tomaszczyk (Ed.), *Practical applications in language and computers - PALC 2003* (pp. 179-192). Peter Lang.
- Pulido, C. M., Villarejo-Carballido, B., Redondo-Sama, G., & Gómez, A. (2020). COVID-19 infodemic: More retweets for science-based information on coronavirus than for false information. *International Sociology*, 35(4), 377-392. <https://doi.org/10.1177/0268580920914755>

- Santiago, R., Carvalho, T., & Relva, R. (2008). Research and the universities' image. *European Journal of Education*, 43(4), 495-512. <https://doi.org/10.1111/j.1465-3435.2008.00370.x>
- Sleurs, K., & Jacobs, G. (2005). Beyond preformulation: An ethnographic perspective on press releases. *Journal of Pragmatics*, 37(8), 1251-1273. <https://doi.org/10.1016/j.pragma.2004.11.007>
- Strobbe, I., & Jacobs, G. (2005). E-releases: A view from linguistic pragmatics. *Public Relations Review*, 31(2), 289-291. <https://doi.org/10.1016/j.pubrev.2005.02.009>
- Stubbs, M. (2002). Two quantitative methods of studying phraseology in English. *International Journal of Corpus Linguistics*, 7(2), 215-244. <https://doi.org/10.1075/ijcl.7.2.04stu>
- Sumner, P., Vivian-Griffiths, S., Boivin, J., Williams, A., Venetis, C. A., & Davies, A. (2014). The association between exaggeration in health related science news and academic press releases: Retrospective observational study. *BMJ*, 349. <https://doi.org/10.1136/bmj.g7015>
- Woloshin, S., Schwartz, L. M., Casella, S. L., Kennedy, A. T., & Larson, R. J. (2009). Press releases by academic medical centers: Not so academic? *Annals of Internal Medicine*, 150(9), 613-618. <https://doi.org/10.7326/0003-4819-150-9-200905050-00007>
- Yavchitz, A., Boutron, I., Bafeta, A., Marroun, I., Charles, P., Mantz, J., & Ravaud, P. (2012). Misrepresentation of randomized controlled trials in press releases and news coverage: A cohort study. *PLOS Medicine* 9(9), e1001308. <https://doi.org/10.1371/journal.pmed.1001308>

