Linguistic changes in the transition from summaries to abstracts: The case of the Journal of Experimental Medicine

Carlo Galli, Maria Teresa Colangelo, and Stefano Guizzardi

Department of Medicine and Surgery, Histology and Embryology Lab, University of Parma, Parma, Italy

ORCID:
C. Galli: 0000-0001-7476-7181
M.T. Colangelo: 0000-0002-4664-5766
S. Guizzardi: 0000-0003-2395-8767

*Corresponding author: Carlo Galli, Department of Medicine and Surgery, Via Volturno 39, 43126 Parma, Italy.
E-mail: carlo.galli@unipr.it

Key points

- The introduction of Abstracts to replace article summaries in 1990 recognized changes to linguistic reporting that have been apparent during the century.
- The 1970s showed a dramatic increase in the informal language used in article abstracts and summaries.
- The Journal of Experimental Medicine (JEM) demonstrates an increase in first-person pronouns within article abstracts and summaries, but moves from singular to plural to represent the increase in multi-authored research works.
- Linguistic changes during the century also include a greater focus on the future rather than the past, and an increase in language that indicates ‘clout’ which signifies author self-confidence.

Keywords: Abstract, academic writing, science writing, writing style

INTRODUCTION

Abstracts are considered as an essential part of academic publications and they often constitute the only part of the text that is actually read during a literature search—besides the title. It is common knowledge that the rationale of an abstract is to provide a short description of the content of the text, like a mini-version of the article itself, to provide readers with all the elements they need to decide whether that article is of interest to them, and whether it meets their needs (Cross & Oppenheim, 2006).

It would be unthinkable, in today’s information overload setting (Landhuis, 2016), to expect readers to go through a whole article—a process that may take hours, when accurately done—just to understand whether they can skip it and move on to the next paper in a list that may comprise thousands of items. Abstracts are a necessity, and thus often follow a standard structure, to improve their readability. In life sciences, they usually comprise an introduction/background section, materials and methods, a results section and a conclusion section, just like a small replica of a study report (Atanassova et al., 2016).

In medicine, which often has, for instance, the need to sort out desired clinical trials for metanalysis that can then drive guidelines and decision-making processes, abstracts can even comprise finer-grained sections, like study design or population of interest (Bahadoran et al., 2020).

The usefulness of an abstract in an article is exerted and exhausted mostly at the search level. Once a prospective reader has perused it and the article has been deemed of interest, the attention focus shifts to the main text, and the abstract has accomplished its function. It is like a tag that allows readers to choose in the vast amount of the scientific literature (Alspach, 2017). It can be considered part of the metadata of an article. The adoption of modern information technologies is even prompting researchers to devise novel approaches and algorithms to screen and skim through abstracts to make literature searches quicker and better (Afantenos et al., 2005; Hersh, 2021).

Abstracts are such a consistent feature of academic publications that is hard to imagine an article, or a journal, without them. The consistent presence of abstracts, however, is a relatively recent feature of scientific papers, as it was notably absent in
many journals published in the 19th and the first part of the 20th
centuries (Galli et al., 2020). As the function of abstracts is mainly
in literature searches, they evolved on par with the increase in
the number of scientific publications. The most likely precursor of
abstracts is the summary section, which could be frequently
found at the end of the article, to sum up what had just been
exposed in the body of the text, to distil its content into few
lines, with the presumable goal of creating an information struc-
ture that made remembering and memorizing the article easier
(Vaughan, 1991). This was reflected by a different structure: sum-
maries lacked a material and methods section, because they did
not need to anticipate the experimental details of a study, while
they often had a list structure, which included the main take-
home messages of the study. At a certain point in time, most
journals switched from a post-textual summary to a pre-textual
abstract, if they did not already possess one. This formal change
reflected a turning point in the readers’ relation to scientific liter-
ature, as journals adapted to a new format that appeared to
respond better to the new needs of their readership. Each journal
performed the transition at a different moment, but, by the end
of the 20th century, virtually all journals in the biomedical fields
had abstracts. Some journals, including the Journal of Experimental
Medicine (JEM), marked that transition quite visibly. JEM radically
changed its format in the first issue of Volume 172 in 1990 (Galli
et al., 2020), and a short piece by the then editor in chief,
M. McCarty, explained how the journal’s new look better
responded to contemporaneous aesthetics, including abstracts at
the beginning of each article, which was described as a ‘touch of
modernization’ (sic), although the editor admitted that this choice
was also supported by the fact that readers were ‘now accus-
tomed to look for it in most publications’ (McCarty, 1990). So, in
the case of this journal, we possess a safe terminus post quem to
date the appearance of abstracts.

The purpose of this study is to use quantitative methods to
assess stylistic changes in summaries/abstracts over the course
of time. The working hypothesis is that the actual transition from
a real summary, in the pristine sense of the word, to a new form
of abstract, for indexing and retrieving purposes, did not happen
for JEM overnight in July 1990, when the editor decided to repo-
sition it at the beginning of the article, but can be traced back to
the years preceding 1990. We believe it is important to investi-
gate whether the editor’s decision anticipated or followed a
change in the text that had already started, independently of his
intervention, due to cultural changes in the use of scientific liter-
ture that were already running through the scientific society, and
how far in the past such changes had emerged, because it can
serve as a case study for the changes in the way we interface
with scientific literature that are still occurring and that may
require journal structure to pivot again in the future.

Secondarily, as the function of summaries was primarily the
cognitive role of helping readers understand and memorize the
content of a study, it could be useful to investigate whether
certain characteristics of summaries that have been lost as they
morphed into abstracts may still be useful in abstracts to improve
article visibility, an ever-difficult endeavour in today’s scientific
world.

**MATERIALS AND METHODS**

This study analysed the corpus of abstracts of all the articles that
appeared in the JEM since its foundation. To generate this cor-
pus, the python littér-getter library was used, and a Medline sea-
Arch was carried out through PubMed API using the search term
‘The Journal of experimental medicine [Journal]’. This retrieved
an XML file for each indexed article, which was then used to cre-
ate a pandas DataFrame (Mckinney, 2010) by means of the
BeautifulSoup library. The extracted data for each article were
‘PMID’, ‘Title’, ‘Abstract’, ‘Year’, ‘Volume’, ‘Issue’, ‘Type’ (mean-
ing the type of item, i.e. article, review, commentary etc),
‘Authors’, ‘Affiliation’ and ‘Country’. For embeddings analysis,
the text of the abstracts was pre-processed by lowercasing the
text, removing stop-words using the Gensim library (Řehůřek
& Sojka, 2010), removing punctuations and numbers. The text was
then passed into the Spacy library (Honinal & Montani, 2017),
using the large English vocabulary and Principal component
analysis was used for dimensionality reduction (scikit-learn
implementation).

The lowercased text of the abstracts (but without further pre-processing) was also passed into the proprietary software
Linguistic Inquiry and Word Count (Pennebaker et al., 2015) for
further analysis. LIWC is based on a series of user-defined diction-
aries (Pennebaker et al., 2003), which are used to define scoring
variables associated with specific thematic spheres (Donohue
et al., 2014). Most of these scores are based on how well the
sample text matches a specific dictionary. LIWC 2015 also
includes four non-transparent summary variables, namely Analyti-
cal Thinking, Clout, Authenticity, and Emotional Tone, which are
based on Pennebaker’s previous published research on text cor-
pora (Pennebaker et al., 2015; Tausczik & Pennebaker, 2010). As
the name implies, Analytical Thinking refers to the use of a logical
and consistent language, while Clout is associated with a self-
confident and authoritative attitude (Kacewicz et al., 2014).
A high authenticity score indicates an unhedged, not detached
language, while the Emotional Tone score refers to the quality of
the emotions that permeate the text. LIWC also analyzes the
parts of speech of the texts and quantifies, among other, the use
of personal pronouns, adjectives, verbs, and so forth (as relative
frequencies). The software also infers the time focus of a text,
based on the use of temporal expressions.

Matplotlib (Hunter, 2007) and Seaborn (Waskom, 2021) librar-
ies were then used to plot the data. All analysis was conducted
on Jupyter notebooks (Kluyver et al., 2016).

**RESULTS**

Our PubMed search retrieved 24,012 articles published in the
JEM, from 1896, when the first issue was published at the John
Hopkins University, to the present day. The number of published items steadily increased over the decades and peaked around the 1990s (Fig. 1).

Noticably, fewer articles appeared in the following decades, and this may possibly reflect the fortunes of the journal on the market or its popularity in the scientific community. Its non-exponential growth, however, partially balances out the distribution of the articles over the years and may prove advantageous for further analysis.

Only 969 articles did not have a summary or an abstract (Fig. 2), and these were quite evenly distributed across the decades, with most articles, however, appearing either before the 1920s or after the 1990s. Though, however, in the case of early studies, the lack of a summary is most likely due to a lack of standardization in the format of reporting— as these articles do not otherwise visibly differ from the articles with summary—, in the case of the most recent publications, the ones without abstracts fall within the editorial genre or are simply corrections of previously published works (data not shown). These studies, however, were obviously not considered in the subsequent analysis.

Our first overview of the data aimed at assessing whether a semantic change in summaries/abstracts could be detected, using standard NLP approaches. To do that, we obtained embeddings, that is, dense vector representations of the texts, using the free Spacy library. Vectors are basically an ordered array of real numbers, and they are a convenient mathematical notation that is commonly used in several NLP technologies to represent words or even sentences. Based on the distributional semantics hypothesis, that is, that the meanings of words that co-occur frequently in a corpus of texts are likely to be associated, vector semantics attributes similar vectors to co-occurring words thanks to complex algorithms, which may include neural networks (Konstantinov et al., 2021). A whole sentence can be represented by vectors, most often by the mean of the vectors of the individual words making up the sentence.

The Spacy library generated a vector of length 300 (i.e. an array of 300 real numbers) for each summary/abstract, which we then plotted after dimensionality reduction by Principal Component Analysis (Fig. 3). This procedure ‘compressed’ each 300-dimensional vector into a 2-dimensional tuple, which was conveniently used as a set of x, y coordinates to plot them in a 2D graph. Appendix Table A1 summarizes an example of abstract before and after pre-processing and its corresponding embedding, together with the result of the PCA dimensionality reduction.

Figure 3 shows an elongated cloud of dots, where each point represents a summary/abstract and the colours express the publication decade. The figure clearly shows that the texts are not homogeneously distributed, but older texts (e.g. purple) seem to segregate at one pole of the cloud, and the more recent ones (e.g. khaki and orange) are mostly located at the opposite pole. Since the cloud almost looks as if it was formed by the coalescence of two smaller clouds (indicated by two circles in Fig. 3), we decided to run a k-means clustering algorithm, with k = 2. A clustering algorithm is a mathematical procedure that automatically assigns a set of items to a preset number of clusters, finding the best way to cluster the data into homogeneous groups. Unsurprisingly, the algorithm split the articles at the junction of
the two apparent clouds, and once plotted by publication year, it became apparent that the first cloud contained articles mostly published before the 1970s, whereas the second cloud was mostly composed by articles published after that decade (Fig. 3).

We then proceeded to consider the linguistic surface of the texts, and several characteristics could be observed changing over time. Our data suggest that the use of ‘I’ increased in the 1970s and 1980s, only to decrease again in articles published in the last 20 years. However, the use of ‘we’ has steadily increased since the 1970s. The choice of the first person plural over the singular pronoun might also have to do with the increasing number of authors of publications, which, in recent years, require larger teams because of their increasing complexity (Fig. 4) and their correlation, as measured by Fisher’s $r$, was 0.38. Other pronouns, such as third person or, even more so, second person pronouns, are understandably less frequent in scientific reports, which are usually focused on the results obtained by the narrator’s group (data not shown).

Somewhat consistently with a wider use of first-person pronouns (Fig. 5), a more informal language has become progressively acceptable (Fig. 6) and experienced a great increase in the 1970s as compared with the preceding decades. To identify informal language, LIWC relies on specific dictionaries, which include words ranging from swear words (e.g. damn), netspeak (e.g. btw, lol, thx), assent (e.g. agree, OK, yes), non-fluencies (e.g. er, hm, umm) and fillers (e.g. I mean, you know etc.) (Tausczik & Pennebaker, 2010). The score is then generated as the mean count of occurrences of words in the dictionary.
FIGURE 3  The left-hand-side chart shows the distribution of the embeddings for the summaries/abstracts in the corpus obtained by Spacy, after PCA dimensionality reduction. The cloud of dots seems formed by the juxtaposition of two partially overlapping smaller clouds, approximately indicated here by a red and blue dotted circle. We split the clouds through a k-means clustering algorithm and created two sub-groups. The right-hand-side plot is a bar chart that shows the number of summaries/abstracts (on the y-axis) by publication date (on the x-axis) of these two sub-groups, arbitrarily indicated by the orange or blue colour. The red dotted line indicates the introduction of abstracts.

FIGURE 4 Barchart representing the average number of authors over time in the articles corpus of the Journal of Experimental Medicine. Values are expressed as mean ± SD.
We then assessed the four LIWC non-transparent summary indices, Analytical thinking, Clout, Authenticity and Emotional tone (Fig. 7). Unsurprisingly, all abstracts, regardless of their publication date, scored very high on the Analytical score, which is an index associated with the use of a more rigorous, precise and technical language, as it is expected from scientific papers, and no substantial change over time was detected (Fig. 7).

The Clout score, which is associated to the use of confident, authoritative language, steadily increased from the late 1960s, consistently with reports by Hyland et al. (Hyland &
Jiang, 2017). Biomedical articles published on JEM did not experience a change in the emotional tone, whereas the Authenticity score, which reflects the degree of detachment from the text, peaked around the 1920s and then steadily decreased.

The results with the Clout index are corroborated by the data in Fig. 5, because personal pronouns are part of the words that make up that index.

We then assessed the temporal dimension in this corpus (Fig. 8). The temporal focus of the text is assessed by LIWC again through the use of specific dictionaries that comprise expressions associated to the past (e.g. ago, did, talked etc.), to the present (e.g. today, is, now etc.), or to the future (e.g. may, will, soon etc.) (Tausczik & Pennebaker, 2010). Once again, the score is the mean count of occurrences of words in the dictionary over the total number of summaries/abstracts. While a future focus was hardly detectable, as it may be anticipated in texts that center on a set of experiments conducted in a very recent past, it is, however, noteworthy that LIWC was capable to pick up a strong focus on the present, which peaked around the first 20 years of the 20th century only to then decrease and stabilize, and a focus on the past, which peaked around the 1950s, and then abruptly decreased after the 1960s.

**FIGURE 7** The line plot shows how the four summary scores (Analytical Thinking as a solid blue line, Clout as a dashed orange line, Authenticity as a dotted green line and Emotional Tone as dashed red line) change over time in the present corpus. The red dotted line indicates the introduction of abstracts.

The presence of abstracts is ubiquitous in scientific production. Abstracts almost invariably precede a scientific product and are often used as a sample to assess the value of the product itself (Ghasemi et al., 2019). Many journals at the beginning of the 20th century, however, did not possess abstracts but rather summaries, placed at the end of the main text. While both abstracts and summaries are, roughly speaking, a condensed version of the main text, they differ profoundly (Vaughan, 1991). The transition from summaries to abstracts, then, was more than just a displacement of a piece of text. It was rather a matter of reorganizing the text itself. The Journal of Experimental Medicine is a useful case study to investigate this passage.

Browsing through summaries published in JEM during the 1980s, it becomes apparent that these texts had already evolved from what they used to be almost a century earlier. The editor’s move was likely to formally reflect a change in the habits of readers that had presumably started to take place years, possibly decades, before. As similar changes may still be occurring in the literature, even in the absence of a formal re-organization of the text, we deem instructive to investigate how extensively articles had changed by the time the editorial team decided to adapt the structure of their articles accordingly.

**DISCUSSION**

The presence of abstracts is ubiquitous in scientific production. Abstracts almost invariably precede a scientific product and are often used as a sample to assess the value of the product itself (Ghasemi et al., 2019). Many journals at the beginning of the 20th century, however, did not possess abstracts but rather summaries, placed at the end of the main text. While both abstracts and summaries are, roughly speaking, a condensed version of the main text, they differ profoundly (Vaughan, 1991). The transition from summaries to abstracts, then, was more than just a displacement of a piece of text. It was rather a matter of reorganizing the text itself. The Journal of Experimental Medicine is a useful case study to investigate this passage.

Browsing through summaries published in JEM during the 1980s, it becomes apparent that these texts had already evolved from what they used to be almost a century earlier. The editor’s move was likely to formally reflect a change in the habits of readers that had presumably started to take place years, possibly decades, before. As similar changes may still be occurring in the literature, even in the absence of a formal re-organization of the text, we deem instructive to investigate how extensively articles had changed by the time the editorial team decided to adapt the structure of their articles accordingly.
We decided to investigate this change in a systematic way through a natural language processing (NLP) approach. To this purpose, we first obtained vector representations of the content of summaries/abstracts, using a very common NLP library, Spacy. Vectorial representations of a text allow to conveniently apply mathematical methods to texts, including the quantification of similarity and distance. The semantic distribution of the content of these texts was noticeably characterized by two distinct focal points (Fig. 3), which turned out to mostly include articles published up to the 1970s or after that date. These data would indicate that the 1970s represent a sort of pivotal moment, or a watershed, so to speak, in the evolution of summaries/abstracts.

This kind of high-level analysis does not provide details as to what distinguishes these two sets of texts, whether the vocabulary used, or their content, for example, research topics. To further clarify that, we had to dig deeper into the language features of the text, and we decided to take advantage of LIWC, a proprietary software that, based on the use of certain empirically validated keywords, assigns scores to specific psycholinguistic features of the text (Chung & Pennebaker, 2012; Donohue et al., 2014). As the focus of the present study is not psycholinguistics, our interest lays more on the textual characteristics that are detected by LIWC than on the actual categories that LIWC uses and how they can correlate to psychological traits and events. LIWC highlighted several features that had changed over time.

Noteworthy, the use of personal pronouns in JEM abstracts changed quite dramatically in the last quarter of the 20th century (Fig. 5). A vast amount of literature has confirmed the increase in the use of first-person pronouns, which are often a natural choice for the syntactic structure of these texts, in several science fields (Wheeler et al., 2021). A scientific article usually describes an experiment, or a series of experiments conducted by the authors, and the use of first personal pronouns can be traced back to the first examples of scientific journal in the 17th century (Kronick, 1984). Although their use was already explicitly encouraged in academic publications at the beginning of the 20th century (Kuo, 1999; Tang & John, 1999), common habits of scientific writing have long insisted that first personal pronouns were to be avoided (Harwood, 2005). Passive constructions, on the contrary, are often employed in scientific and technical texts, as they allow to avoid expressing the agent of an action, thus making the sentence more impersonal, official, authoritative, unaffected by emotional involvement (Hyland, 2002). Plus, according to some authors, this may even allow authors to avoid taking direct responsibility for what is described. This tendency, however, is
being progressively rejected, and the use of ‘I’ and ‘we’ pronouns is a more common encounter now, as they are being perceived as instrumental to make simpler and clearer sentences, and thus to be preferred over passive structures. The preference in the use of the plural versus the singular pronoun, then, may be dictated by several factors, including the cultural background of the authors (Vassileva, 1998).

Consistently with that, our analysis has also detected a decrease in language formality, since the 1970s, with contemporary abstracts deemed more informal than older texts (Fig. 6), in agreement with previous reports (Hyland & Jiang, 2017). While this is consistent with similar findings, in biology and engineering journals—though not in all areas of science—there can be several explanations, which range from a general trend of change in language use in the population, to an increased need of writers in the life science area to connect with readers, to make stronger cases, especially for grant application purposes (Hyland & Jiang, 2016; Hyland & Jiang, 2017). This progressive shift to a higher degree of informality should also be framed in a larger context, as English has established itself as the academic lingua franca of science, and increasing numbers of scientific contributions by non-native speakers are published, which may affect register distinctions and the overall quality of the language (Candlin & Hyland, 2014; Yuan et al., 2013; Zhao, 2017). While these are interesting linguistic features of the texts, LIWC allows for broader assessment of the language.

LIWC summary indices are among the main outputs of this software. These include Analytical thinking, Clout, Authenticity and Emotional tone (Fig. 6). Both Clout and Authenticity display changes over time. The way Clout is computed, an increase in the frequency of first-person pronouns can increase this score, which is about the self-confidence and charisma that are projected by the language use. As we showed in Fig. 5, the frequency of ‘I’ and ‘we’ pronouns has increased dramatically since the 1970s, so this may in part explain the increase in the Clout index. Unfortunately, as these indexes are non-transparent, it is difficult to clearly assess whether other factors beside a wider use in personal pronouns may have driven the change in Clout score.

Our data show that Authenticity score peaked in our corpus around the 1920s and then steadily decreased after the 1950s. To get a clearer idea of what kind of honest, authentic, unhedged language may be picked up by LIWC in the JEM corpus, we browsed through the corpus, and a quick look at summaries from that era does reveal strong stylistic differences:

...When a mouse or white rat was inoculated, spirochetes always appeared in the peripheral blood, but no other symptoms developed...Thus we found that rats and mice are media but not victims of the disease.. (Ishiwara et al., 1917)

In this passage, arbitrarily chosen because its exemplarity, readers are taken through the heuristic process, they are accompanied along the experimental procedure, the results obtained and the reasoning that allowed the authors to draw certain conclusions from them. We get to know that spirochetes were inoculated, and every time the authors did it, they would find a certain result, which creates a time perspective. This repeated observation, in turn, lead the authors to their scientific achievement. The authors are telling us the story of how they conducted the study, and, for a brief moment, we are standing at their side, at the bench, and we are living with them through that specific moment in time.

This stands in sharp contrast with a typical passage from a contemporary abstract:

...We show that placental trophoblasts constitutively secrete the inflammasome-associated cytokines IL-1β and IL-18, which is blocked by NLRP3 inflammasome inhibitors and occurs without detectable gasdermin D cleavage... (Megli et al., 2021)

The authors here situate their statement in an a-temporal context, they are showing in the paper that this is how this specific biological process happens in nature, physiologically. Our personal connection to the authors does not really play any role here; we are facing impersonal laws of nature that have been ascertained by the authors, but the whole process of discovery, the human tribulation that was weathered to achieve their results is not disclosed, is not relevant to the aseptic light of science. Could this be relevant to the Authenticity score? Authenticity has been described as higher when the speaker ‘has a personal stake in the issue, is similar to the receiver’ and is ‘willing to admit uncertainty’ (Renn & Levine, 1991). Indeed, narrative styles are perceived as more genuine and authentic (Saffran et al., 2020), so it is possible that this index, which has been empirically determined and tuned (Pennebaker et al., 2015), is at least partially recording a more narrative approach to science communication in the earlier summaries. This would be also confirmed by the stronger temporal focus of texts before the 1960s (Fig. 8). LIWC highlights a distinctive present and past time positioning that decreased—quite consistently in the case of the latter—in the second half of the 20th century. The temporal focus is not affected only by the use of specific verb tenses, but also by adverbs and time expressions. This result would be consistent with the more common use of narration earlier in the first half of the last century.

So, quite noteworthy, abstracts would have strayed from the use of narrative devices, which are often perceived as a more immediate way of communicating, but at the same time tried to establish a connection with the reader through a more vivid language and a wider use of first-person pronouns.

Again, these changes occur in JEM, for the most part, around the 1970s, 20 years before the official format switch, and testify how, around those years, summaries changed quite profoundly, both in content and in the language used. This is in agreement with other biomedical journals we surveyed in a recent publication (Galli et al., 2020), which adopted abstracts around that
decade. Several hypotheses can be put forth to explain why articles changed so visibly in those years, and a prime suspect is possibly the development of computational methods to handle literature searches, with the Index Medicus progressively transitioning to digital supports. When taken together, these data suggest that the editor’s move to introduce abstracts was reactive, rather than pro-active, and it responded to a change in the way summaries were being composed that was already well underway and had been for at least a couple of decades. These data also highlight the use of narration in older summaries, a textual device that could turn quite fruitful to help new writers achieve that connection with the readers that, despite the big changes occurred in science, appears to still be so relevant to gain visibility.

Future studies will have to address these complex issues and better investigate how literature fruition has changed the way scientific literature is written and how the changing ways we approach scientific texts are still operating on this important modality of knowledge transmission.

CONCLUSIONS

In conclusion, the present study has highlighted a series of changes in the features of summaries/abstracts over time using NLP embeddings and LIWC software, including a decrease in authenticity/temporal focus, a wider use of first-person personal pronouns, together with features of more informal language and Clout.

Interestingly, while many trends can be observed when we consider the scientific production of this journal since its beginnings at the end of the 19th century, most of the trends that were taking place at the time of the transition from summary to abstract had already started since the 1960s or 1970s. This suggests that summaries, although they had retained the name that they had got in earlier days, had already started to acquire a different nature by the 1970s, which then consolidated in the following decades and the editor’s decision was actually quite late, as compared to the underlying phenomena. Recent literature appears to be striving to gain a more direct approach to the readers, thanks to the use of a simpler, informal language and to this purpose might benefit from recovering the use of narrative devices that appear quite common in older summaries.

ACKNOWLEDGEMENT

Open Access Funding provided by Universita degli Studi di Parma within the CRUI-CARE Agreement.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

Conceptualization, Carlo Galli and Stefano Guizzardi; methodology, Carlo Galli; software, Carlo Galli; formal analysis, Carlo Galli; investigation, Maria Teresa Colangelo; resources, Stefano Guizzardi; data curation, Carlo Galli; writing—original draft preparation, Maria Teresa Colangelo; writing—review and editing, Carlo Galli; supervision, Stefano Guizzardi; Both authors have read and agreed to the published version of the manuscript.

DATA AVAILABILITY STATEMENT

Data is available on request.

REFERENCES

Afantenos, S., Karkaletsis, V., & Stamatopoulos, P. (2005). Summarization from medical documents: A survey. Artificial Intelligence in Medicine, 33(2), 157–177. https://doi.org/10.1016/j.artmed.2004.07.017

Alspach, J. G. (2017). Writing for publication 101: Why the abstract is so important. Critical Care Nurse, 37(4), 12–15. https://doi.org/10.4037/ccn2017466

Atanassova, I., Bertin, M., & Larivière, V. (2016). On the composition of scientific abstracts. Journal of Documentation, 72(4), 636–647. https://doi.org/10.1108/JDOC-09-2015-0111

Bahadoran, Z., Mirmiran, P., Kashfi, K., & Ghasemi, A. (2020). The principles of biomedical scientific writing: Abstract and keywords. International Journal of Endocrinology and Metabolism, 18(1), e100159. https://doi.org/10.5812/ijem.100159

Candlin, C. N., & Hyland, K. (2014). Writing: Texts, processes and practices. Routledge. https://doi.org/10.4324/9781315840390

Chung, C. K., & Pennebaker, J. W. (2012). Linguistic inquiry and word count (LIWC). In Applied natural language processing (pp. 206–229). IGI Global. https://doi.org/10.4018/978-1-60960-741-8.ch012

Cross, C., & Oppenheim, C. (2006). A genre analysis of scientific abstracts. Journal of Documentation, 62(4), 428–446. https://doi.org/10.1108/002204106100953

Donohue, W. A., Liang, Y., & Druckman, D. (2014). Validating LIWC dictionaries. Journal of Language and Social Psychology, 33(3), 282–301. https://doi.org/10.1177/0261927X13512485

Galli, C., Colangelo, M. T., & Guizzardi, S. (2020). Striving for modernity: Layout and abstracts in the biomedical literature. Publications, 8(3), 8. https://doi.org/10.3390/publications8030038

Ghasemi, A., Bahadoran, Z., Zadeh-Vakili, A., Montazeri, S. A., & Hosseinpahan, F. (2019). The principles of biomedical scientific writing: Materials and methods. International Journal of Endocrinology and Metabolism, 17(1), e88155. https://doi.org/10.5812/ijem.88155

Harwood, N. (2005). ‘We do not seem to have a theory ... the theory I present here attempts to fill this gap’: Inclusive and exclusive pronouns in academic writing. Applied Linguistics, 26(3), 343–375. https://doi.org/10.1093/applin/ami012

Hersh, W. (2021). Information retrieval. In J. J. Shortliffe, H. Edward, & Cimino (Eds.), Biomedical informatics: Computer applications in health care and biomedicine (pp. 755–794). Springer International Publishing. https://doi.org/10.1007/978-3-030-58721-5_23

Honnibal, M., & Montani, I. (2017). Spacy 2: Natural language understanding with bloom embeddings. Convolutional Neural Networks and Incremental Parsing, 7(1), 411–420.
This table compares the original text of one abstract in the corpus (left column), with the same text after pre-processing (which included lowercase letters and stopwords removal), its corresponding 300-dimensional embedding and the same embedding after Principal Component Analysis (right column), as plotted in Fig. 3a.

| Original text                                                                 | Text after pre-processing | 300-dimensional embedding | 2-dimensional embedding after PCA |
|-------------------------------------------------------------------------------|--------------------------|---------------------------|----------------------------------|
| Obesity-induced secretory disorder of adipose tissue-derived factors is important for cardiac damage. However, whether platelet-derived growth factor-D (PDGF-D), a newly identified adipokine, regulates cardiac remodelling in angiotensin II (AngII)-infused obese mice is unclear. Here, we found obesity induced PDGF-D expression in adipose tissue as well as more severe cardiac remodelling compared with control lean mice after AngII infusion. Adipocyte-specific PDGF-D knockdoun attenuated hypertensive cardiac remodelling in obese mice. Consistently, adipocyte-specific PDGF-D overexpression in transgenic mice (PA-Tg) showed exacerbated cardiac remodelling after AngII infusion without high-fat diet treatment. Mechanistic studies indicated angII-stimulated macrophages produce urokinase plasminogen activator (uPA) that activates PDGF-D by splicing full-length PDGF-D into the active PDGF-DD. Moreover, bone marrow-specific uPA knockdoun decreased active PDGF-DD levels in the heart and improved cardiac remodelling in HFD hypertensive mice. Together, our data provide for the first time a new interaction pattern between macrophage and adipocyte: that macrophage-derived uPA activates adipocyte-secreted PDGF-D, which finally accelerates AngII-induced cardiac remodelling in obese mice. | obesity induced secretory disorder of adipose tissue-derived factors is important for cardiac damage. However, whether platelet-derived growth factor-D (PDGF-D), a newly identified adipokine, regulates cardiac remodelling in angiotensin II (AngII)-infused obese mice is unclear. Here, we found obesity induced PDGF-D expression in adipose tissue as well as more severe cardiac remodelling compared with control lean mice after AngII infusion. Adipocyte-specific PDGF-D knockdoun attenuated hypertensive cardiac remodelling in obese mice. Consistently, adipocyte-specific PDGF-D overexpression in transgenic mice (PA-Tg) showed exacerbated cardiac remodelling after AngII infusion without high-fat diet treatment. Mechanistic studies indicated angII-stimulated macrophages produce urokinase plasminogen activator (uPA) that activates PDGF-D by splicing full-length PDGF-D into the active PDGF-DD. Moreover, bone marrow-specific uPA knockdoun decreased active PDGF-DD levels in the heart and improved cardiac remodelling in HFD hypertensive mice. Together, our data provide for the first time a new interaction pattern between macrophage and adipocyte: that macrophage-derived uPA activates adipocyte-secreted PDGF-D, which finally accelerates AngII-induced cardiac remodelling in obese mice. | -1.09628297e-01, 3.10655117e-01, -1.73154056e-01, 2.14143004e-02, -1.37138113e-01, 5.15988432e-02, 5.33828549e-02, -4.01112363e-02, 3.78710814e-02, 9.84602749e-01, -4.89784852e-02, 1.38292760e-01, 5.81617206e-02, 2.69455612e-02, 1.38558134e-01, -1.12352581e-02, 1.35304667e-02, 1.30443811e-00, -1.05171204e-01, -1.26201520e-02, 1.50410393e-02, -1.08450204e-01, -1.58816099e-01, -8.28903243e-02, 2.17002518e-02, 8.72200951e-02, -9.56048599e-02, 9.13890302e-02, 1.38770655e-01, -3.04621402e-02, -1.77034095e-01, 1.11999221e-01, 2.34778244e-02, -1.59812793e-02, 1.12175934e-01, 9.99671072e-02, 1.10849726e-03, 5.68818338e-02, -7.39412159e-02, 6.21862523e-02, -4.98342067e-02, -1.24395080e-02, -5.33521641e-04, -6.64939582e-02, -8.05498606e-02, -5.45682423e-02, 1.00334268e-02, 8.75305682e-02, 1.73831373e-01, -3.16813961e-01, -4.20636386e-02, 5.78566827e-03, -7.03993989e-03, -8.71704891e-02, 7.81382397e-02, -2.37850938e-02, 1.15370058e-01, -9.65528339e-02, -9.12376121e-02, -1.00953907e-01, 3.54635343e-02, 1.02376856e-01, 4.61444594e-02, 1.22385696e-02, 9.93708149e-03, -2.26789251e-01, 8.86724591e-02, -9.38326642e-02, -4.94446121e-02, 1.39532838e-02, 5.53036974e-02, 1.34734035e-01, 7.54000715e-02, -8.70676288e-02, 1.36612639e-01, -6.17816783e-02, -7.22368807e-02, -2.55082874e-03, -6.99451566e-02, -1.41603529e-01, 9.93651748e-02, 1.07244894e-01, -2.0026861e-01, -3.07355244e-02, 3.95396550e-02, -3.41428459e-01 | -0.1911665, -0.10896736 |
| Original text | Text after pre-processing | 300-dimensional embedding | 2-dimensional embedding after PCA |
|---------------|---------------------------|---------------------------|----------------------------------|
| 2.5809136e-01, -1.34478197e-01, -2.33781800e-01, | -9.64847654e-02, 3.28934006e-02, 2.69112915e-01, | -1.48267634e-01, -2.8077064e-03, -1.67740583e-02, -1.96831077e-02, -6.43410499e-02, -1.47768231e-02, -1.64861283e-01,-1.71612501e-02, -6.8850392e-02, -1.73004568e-01, -1.2368045e-02, 2.20317557e-01, -7.3666083e-01, -2.05554068e-02, 5.45083992e-02, 1.75938532e-01, 1.29945293e-01, -5.34612080e-03, 5.69500804e-02, -3.2081623e-02, -5.02687357e-02, -3.77172492e-02, 4.51370627e-02, -1.44507196e-02, 1.33940727e-01, -6.56415075e-02, -5.03909513e-02, 4.06290889e-02, 1.94194689e-01, 1.49084199e-02, -1.4074546e-01, -1.12105235e-02, 3.02558299e-02, 4.61486552e-02, -1.32024631e-01, 2.89729703e-02, 1.38557345e-01, -9.18836147e-02, -4.57511907e-02, 7.41106048e-02, 9.50100270e-02, 8.70765466e-03, -7.01327473e-02, -1.59695566e-01, 5.77928126e-02, 1.77104305e-02, -1.64799184e-01, 1.32056907e-01, -2.88504027e-02, 9.56825260e-03, -8.08174605e-04, 6.33797795e-02, -6.81762695e-02, 1.28372267e-01, -6.66773471e-04, 1.45212367e-01, -1.24567851e-01, -6.09490555e-03, 9.84947011e-02, 5.59167825e-02, -1.68591589e-01, -1.29247516e-01, -9.70109552e-03, -4.57916819e-02, -8.40330273e-02, -1.31896436e-01, 6.07744120e-02, -1.37553120e-02, -1.03001371e-01, -1.38771877e-01, -8.6905116e-02, -4.42322828e-02, 6.18407615e-02, -5.63463755e-02, -1.71888359e-02, 6.98950497e-02, 1.02965027e-01, 6.44710724e-02, -1.56784564e-01, 8.84224549e-02, 4.97249477e-02, -6.04418404e-02, 1.35822222e-02, -4.71370742e-02, 5.87075049e-01, 3.01273614e-02, -1.87309429e-01, 4.67379948e-03, -5.68087548e-02, 6.39680028e-02, -6.29417971e-02, -3.27462666e-02 |
| Original text | Text after pre-processing | 300-dimensional embedding | 2-dimensional embedding after PCA |
|---------------|---------------------------|--------------------------|----------------------------------|
| 3.56966220e-02, 1.72857955e-01, 2.86819395e-02, −9.71246809e-02, −6.57420754e-02, −1.96780875e-01, 7.17863590e-02, 2.74601784e-02, 2.57735271e-02, −2.8639204e-03, −5.69082979e-02, −1.11065567e-01, 4.92754914e-02, −1.3895617e-01, 3.51291313e-03, 1.19947545e-01, −1.56011209e-02, −8.92051160e-02, 9.95840281e-02, 5.1581770e-02, 3.30281183e-02, 4.3318548e-03, −1.58227875e-03, −1.76240817e-01, −1.15006603e-01, 9.17656869e-02, −2.23911107e01, 1.29857630e-01, 7.31468126e-02, 9.67805237e-02, −8.12668353e-02, 4.79814011e-02, −5.25187179e-02, −3.19090217e-01, 3.59347626e-03, −5.38732708e-02, 8.87015685e-02, 1.02832057e-01, −4.17356193e-02, 1.56539381e-01, 3.49708237e-02, −1.32309854e-01, 2.19790617e-02, 4.38613407e-02, 9.55976699e-03, 1.98860671e-02, 5.23035616e-02, −5.03931083e-02, −9.85947326e-02, 2.63783019e-02, 2.84470282e-02, −5.25804423e-02, −3.75145413e-02, 7.21476998e-02, 5.44574000e-02, 2.58052014e-02, −2.91445311e-02, −7.63968453e-02, −5.03793173e-02, −1.63502172e-01, 3.78855392e-02, −3.25040370e-02, 8.08597207e-02, 1.34754345e-01, 1.2394032e-02, 4.73713595e-03, −6.40352517e-02, −1.09094858e-01, −2.30703950e-01, 1.89133450e-01, −3.85574512e-01, −5.90342879e-02, 4.44382206e-02, 4.19860221e-02, 4.4476904e-02, −3.93575840e-02, 1.23226985e-01, 7.1689919e-02, −1.72445662e-02, 2.53674358e-01, 6.70982003e-02, 7.9088071e-02, 6.0120840e-02, −4.38295677e-03, −4.18853201e-02, −1.17082635e-01, −4.5236297e-02, 3.6312976e-03, 3.70284915e-02, −1.83529705e-01, 2.59355724e-01, 1.70830011e-01, 5.31571023e-02, 3.99252996e-02, −1.00967675e-01, 1.0442780e-01, 8.38779658e-02, 1.08027443e-01, −1.62372291e-01, 1.88282236e-01, 1.44686639e-01, 1.25996262e-01, −4.28354144e-02, −2.38855015e-02, −3.06700245e-02, −6.22372366e-02, 1.5171985e-01, −5.32211157e-02, −4.94633876e-02, 7.49334022e-02, −5.75652998e-03 |