You Won't Believe What's in this Paper! Clickbait, Relevance, and the Curiosity Gap.

Abstract

Drawing on a corpus of clickbait headlines (Chakraborty, 2016) and using ideas from the relevance-theoretic pragmatic framework (Sperber and Wilson, 1986/95), this paper examines some of the ways in which writers of clickbait headlines arouse the curiosity of their readers by creating an "information gap" (Loewenstein, 1994). Comparative corpus analysis is combined with close analysis of illustrative examples to explore the contribution that particular parts-of-speech make to the creation of successful clickbait. I focus on two main categories that are overrepresented in clickbait headlines to a statistically significant degree: (i) definite referring expressions and (ii) superlatives and intensifiers. The results and analysis reveal that these parts-of-speech contribute to an information gap by encouraging readers to construct new conceptual files based on the terms used in the headline, while providing little or no content for those files. This then drives the reader to click on the associated link with the expectation that the article will contain relevant information with which he can enhance his conceptual files, and that this, in turn, will reward him with cognitive effects.

Keywords: clickbait; curiosity; information gap; relevance theory; corpus analysis; online communication

1 Introduction

Online advertising is a hugely profitable and growing industry. In 2016 in the US, internet advertising revenues surpassed television advertising for the first time (Interactive Advertising Bureau, 2017). Since then online advertising has continued to grow year-on-year (Interactive Advertising Bureau, 2019). It is therefore perhaps not surprising that online marketers have developed new techniques to drive users to pages featuring online advertising. The more people who view a webpage, the more valuable the advertising space on that page becomes. One technique to encourage users to visit a page is the use of so-called clickbait. Clickbait links take readers to landing sites which feature advertisements alongside a usually low-quality content article. Publishers earn fees from the advertisers based on the number of views the pages receive (Potthast et al., 2018).

While all web links are, we assume, designed to be clicked on, clickbait links are distinct in that inducing a click is their sole purpose. Once a user has followed the link, the creator has achieved their goal. Any subsequent engagement with the content on the site itself is of little importance to the clickbait writer. For this reason, the content on the landing sites tends to be low quality and it rarely lives up to the promise of the headline itself. As a result, clickbait headlines are typically characterised as being deceptive, misleading, or disappointing in some way.

Existing work on clickbait has revealed a lot about what it tends to look like. Much of the existing research on the topic has been aimed at developing algorithms and software which can be used to detect likely clickbait content and to direct readers away from it. This detection work has identified various linguistic techniques which

tend to be employed in clickbait headlines. The other key strand of existing research into clickbait has focused on the role that curiosity plays in its success. Clickbait headlines, it has been argued, create an information gap which piques a reader's curiosity.

In this paper, I aim to bridge the gap between these two strands of research. I will consider how the language in clickbait headlines is used to create an information gap, and also how this information gap can be understood in cognitive pragmatic terms. To do this, I draw on insights from the relevance-theoretic pragmatic framework. Relevance theory (Sperber and Wilson, 1986/95) is a framework for understanding both communication and cognition. As such, it is ideally positioned to offer insights into why and how clickbait works, and into the role that the language of clickbait plays in arousing a reader's curiosity. Clickbait headlines are formulated, I will argue, so as to exploit our natural tendency to seek out inputs that are relevant to us. We find ourselves naturally drawn towards stimuli which seem likely to offer us cognitive rewards. The headlines create information gaps, and they imply that we will find cognitive rewards if we follow the links and fill in the gaps, and so we are enticed to click. In this article I suggest that insights from the relevance-theoretic pragmatic framework can reveal and explain some of the linguistic and pragmatic strategies which underlie the language of clickbait and which contribute to its success.

This paper has two main aims. First, I report on a comparative corpus analysis which was used to identify and confirm which categories of linguistic items are overrepresented in content drawn from media sources that are associated with clickbait use. The headlines from this dataset were compared with headlines taken from sources which are not traditionally associated with clickbait techniques. Existing work has suggested that certain parts of language (including pronouns, forward referencing devices, and superlatives) are used more often in clickbait headlines than in headlines from more traditional news sources. A comparative corpus analysis allows us to confirm that this is the case and to further interrogate the distribution of the categories and parts of language that are overused. We will consider, for example, whether all types of forward-referring expression are overused to the same extent. The second aim of this paper is to provide a relevance-based, pragmatic explanation for why we see these distributional patterns. What is it about these parts of language which make them a key part of the clickbait register and repertoire? What role do they play in creating an information gap and thus arousing curiosity? How can we understand their contribution in terms of pragmatics, and, more specifically, their contribution to relevance?

I begin in Section 2 by providing an overview of the limited existing work on clickbait and by discussing properties and characteristics usually associated with clickbait content. Then in Section 3 I introduce relevance theory as a framework for understanding utterance interpretation and I consider how we might reframe these information gaps and their links to curiosity in relevance-theoretic terms. I also briefly discuss relevance-based work on newspaper headlines, and suggest that clickbait headlines differ from traditional headlines in terms of how readers are expected to engage with them. In Section 4, details of the comparative corpus analysis methods used are presented. In Sections 5 and 6, I focus on two categories from the data: definite referring expressions (Section 5), and superlatives and intensifiers (Section

6). In each case the results of the comparative corpus analysis are presented. Then, using ideas from relevance theory, the distributional patterns are analysed in terms of how each category contributes to the creation of an information gap. Individual examples from the corpus are discussed as illustrations of how interpretation might proceed and of how the lexical items selected contribute to the creation of a promise of future relevance that entices the reader to click.

2 Defining Clickbait

The term “clickbait” is usually used to describe online content that is specifically designed to entice a reader to click on a link but which offers very little reward for doing so. As such, clickbait has a rather negative reputation, and much of the existing research has been motivated by the aim of developing automatic clickbait detection mechanisms (Chen et al., 2015; Chakraborty et al., 2016; Potthast et al., 2016; 2018). This work has sought to identify cues associated with clickbait material and develop software to automatically screen out the clickbait content. For example, Chen et al. (2015) examine both textual and non-textual clickbaiting strategies, and in doing so they identify a range of cues for recognising clickbait content. These include lexical/semantic cues (“unresolved pronouns, affective language & action words, suspenseful language, overuse of numerals”) and syntactic/pragmatic cues (“forward reference, reverse narrative”) (p.4).

Very little work has considered clickbait headlines from a linguistic or pragmatic perspective. A notable exception is the work by Blom and Hansen (2015) who focus on the linguistic resources which may be exploited in the creation of an information gap. They conduct an analysis of headlines on Danish news websites, and investigate the role played by forward-referring techniques in “creating anticipation and making readers click” (p.89). Blom and Hansen identify the following eight manifestations of forward reference: demonstrative pronouns, personal pronouns, adverbs, definite articles, ellipsis, imperatives, interrogatives, and general nouns with implicit discourse deictic reference. They compare the use of these features in both commercial and non-commercial media in their Danish corpus, and their analysis reveals “a strong tendency for using forward-referring headlines in commercial tabloid media and commercial media without paywalls” when compared with non-commercial and non-tabloid media. Thus they establish a link between the revenue generating focus of the sites and the language that is used on those sites.

Content that is classified as clickbait tends to share certain characteristics that separate it from other online content. Likewise, headlines that link to clickbait sites share characteristics that separate them from other headlines. First, clickbait content tends to focus on certain topics. It tends to either be light-hearted, popularist and sensationalist, or it relates directly to the reader by, for example, promising to reveal something about their personality via a quiz or test of some sort. Examples of typical clickbait headlines are given in (1) to (4). These examples are all taken from a corpus compiled by Chakroborty et al. (2016). This corpus consists of article

headlines taken from media websites associated with the publication of clickbait content: Buzzfeed, Upworthy, ViralNova, Thatscoop, Scoopwhoop and ViralStories.[1]

(1) 12 Mind-blowing Ways to Eat Polenta.

(2) Stop Everything and Look at These Adorably Stylish Dogs.

(3) Someone Calculated How Rich Harry Potter Was And The Answer is Surprising.

(4) Are You More House Stark Or House Targaryen?

As these examples illustrate, clickbait headlines vary in terms of what specific information or content we might expect to find on the landing site. The headlines in (1) to (3) promise us information about polenta, pictures of dogs and information about Harry Potter, respectively. The headline in (4) prompts us to find out which characters from a television show we are most like. Writing coach and journalism teacher Roy Peter Clark (2014) discusses the focus of clickbait content in a blog post for The Poynter Institute for journalism (https://www.poynter.org/about/). He notes that clickbait content is often focused on pop culture topics such as “sex, celebrity and miracle cures”, and the reader is often presented with the promise that he will feel “outraged”, “amazed” or “inspired” by what he will see. This sort of content is also common in tabloid journalism, and so popularist, sensational content alone, does not necessarily make something clickbait. We must look to other characteristics associated with clickbait to distinguish it from other popularist content.

The second key property associated with clickbait is that it arouses the curiosity of the reader by making them aware of an apparent gap in their knowledge. Loewenstein (1994:75) proposes that curiosity is a “cognitively induced deprivation that arises from the perception of a gap in knowledge or understanding”. The symptoms of curiosity are, according to Loewenstein (1994:93) “intensity of motivation, transience, association with impulsivity, and disappointment when information is successfully assimilated”. This characterisation of curiosity helps us to understand why clickbait works. Clickbait, as we shall see, is designed to create an information gap. This gap arouses curiosity in the reader, which increases the chances that he will feel an intense and impulsive, albeit transient, motivation to click on the link.

Creating an information gap is key to the success of clickbait, and Upworthy cofounder Peter Koechley (2012) describes how successful headlines are carefully constructed and tested to maximise reader engagement and clicks. As he explains:

> Upworthy curators come up with 25 headlines for every single nugget they want to post. Then the team narrows the list down to a few finalists, and finally we conduct a bunch of geeky experiments to determine the winner.

[1] Editor-in-Chief Ben Smith (2014) has argued that Buzzfeed does not contain clickbait. He defines clickbait narrowly as content that tricks the reader or which breaks the promises that it makes. Smith claims that Buzzfeed articles do not do this and that they instead provide “entertaining web culture content”.

> We obsess over headlines because we want our content to go viral - and writing a brilliant headline is the easiest way to make that happen (Koechley, 2012).

As he goes on to explain, “a good social media headline seduces people to click through by telling them enough to whet their curiosity but not enough to fulfil it” and “social headlines need to create a curiosity gap. Too vague, and nobody cares. Too specific, and nobody needs to click”. In each example in (1) to (4) we are told that apparently highly relevant information is to be found on the landing site, but we are not provided with the information itself until we click. The polenta information promised in (1) is described as “mind-blowing”, the dogs in (2) are so adorably stylish that it is worth stopping everything else to look at them, and the information about Harry Potter in (3) will, we are promised, be surprising. Finally, clicking on the link in (4) will tell you something about yourself by revealing which of the Game of Thrones families you most resemble. In each case, we are promised relevant information. According to Loewenstein (1994:92-93) “curiosity is driven by the pain of not having information” and “is the feeling of deprivation that results from an awareness of the [information] gap”. Clickbait headlines arouse curiosity by telling us that highly relevant information exists while not actually providing us with the information itself. Instead, we are promised that it is just a click away. This is typical of clickbait headlines, and is one way in which they differ from other headlines that we might find online or offline. Compare the headlines in (1) to (4) with the headlines in (5) and (6) which are taken from the Daily Mail Online website.

(5) JK Rowling WINS legal battle with “utterly dishonest” former PA who used her credit card to splash out on coffee and toiletries - as Harry Potter author reveals she sued her to “protect reputation of her staff” (Keay & McManus, 2019).

(6) Duncan Bannatyne, 70, and his glamorous wife Nigora Whitehorn, 39, pack on the PDA after enjoying a date night at London hotspot (Phillips, 2019).

The Daily Mail is a middle-market, tabloid newspaper which offers its readers both entertainment articles and coverage of politics and news events. The headlines in (5) and (6) are taken from entertainment-focused articles. However, even though these articles deal with popular culture topics and celebrity gossip, the headlines that promote them to the reader are not clickbait. One key different between standard tabloid headlines and clickbait headlines is the completeness of the information that is contained in the headline itself. Unlike clickbait, the non-clickbait headlines in (5) and (6) provide a complete overview of the article to which they link, and they cover all of the key details of the story. By the time we have read the headline in (5) we know that JK Rowling has won her court case and we know all of the key details of the case itself. Compare this to the clickbait in (3). We are not told how rich Harry Potter is. We are just told that this information exists and that it is surprising.

Roy Peter Clark (2014) considers clickbait to be a “mini-genre”, and he identifies various moves that are often involved in creating a successful clickbait headline. These include provoking outrage, putting “odd and interesting things next to each other”, and building an engine for the story. According to journalist Tom French (2002:50), the engine of a story is “an unanswered question that the reader wants to

know the answer to". It is a gap in the reader's knowledge. According to Clark himself, inclusion of an engine is one narrative-based way to inspire curiosity in the headline reader in this way.

Finally, Loewenstein associates curiosity with disappointment. The pleasure derived from the satisfaction of curiosity rarely aligns with the intensity of feeling associated with the curiosity itself. We see this reflected in readers' reactions to clickbait content. A key feature of clickbait is the gap in quality between what is promised by the headlines and what is delivered. The information found on the landing site is generally disappointing and misleading. As Biyani et al. (2016:95) explain, the reader finds himself taken to "genuine pages delivering low quality content with misleading titles". For example, clicking on the headline in (1) takes the reader to a page containing a list of links to 12 fairly ordinary recipes for cooking with polenta. The reader would be forgiven for feeling a little misled by the use of the term *mind-blowing* to describe this fairly unremarkable content. This quality gap has contributed to clickbait being characterised as "one of the pests of social media", and as a form of "false or misleading news" (Potthast et al., 2018:1506). It is considered to be deceptive in nature and "create[s] and exploit[s]…knowledge gaps to entice readers to click through" (Chen et al., 2015:1-2).

However, even when we, as readers, know that what we will find on the landing site is unlikely to live up to our expectations, clickbait still appears to work. In a study into emotional arousal when reading clickbait, Pengnate (2016:7) found that "while online users express negative perceptions toward clickbait, they are still interested in clicking through the headlines". In short, we know that we are being clickbaited, but we click nonetheless. According to Potthast et al. (2018:1506), "its [clickbait's] working mechanisms are still barely understood", and "it remains an open question how the relatively short teaser message can have such a strong effect" (p.1501). As a theory of human cognition and communication, relevance theory can offer an explanation for this effect and for the mechanisms that underlie it. To explore this, I use empirical findings from corpus analysis and analyse both distributional patterns and individual examples using theoretical notions from relevance theory. By deliberately creating information gaps and pointing the reader to a source where that gap will, apparently, be filled, clickbait writers exploit our natural tendency to seek out relevant information. In the rest of this paper, I explore the language of clickbait and the role that certain parts-of-speech play in creating information gaps that many readers find irresistible. I use relevance theory to explain not only how an information gap is created, but also how such gaps interact with cognitive processes and with the search for relevance.

3 Relevance Theory, Headlines, and Information Gaps

Relevance theory (Sperber and Wilson, 1986/95; Carston, 2002; Wilson and Sperber, 2012; Clark, 2013) is a framework for understanding utterance interpretation based on two main principles: the cognitive principle of relevance and the communicative principle of relevance. According to the cognitive principle of relevance, we, as humans, are geared to the maximisation of relevance. That is, we are geared to seek out potentially relevant inputs in our environment and to process them in a relevance-maximising way. An input will be relevant to an individual if it

leads him to update his assumptions in one of three ways. It might lead him to strengthen an existing assumption, it might lead to the contradiction and elimination of an existing assumption or it might combine with an existing assumption to yield a new contextual assumption not previously available. The relevance of an input is also affected by the effort that is required to process it. The more processing effort that is required, the less relevant an input will be.

Ostensive stimuli, including utterances, create, not just a hope, but an expectation of relevance. This is captured in the second, communicative principle. Utterances, as ostensive acts of communication, carry with them a presumption of their own optimal relevance. An input will be optimally relevant if it is (a) worth the addressee's effort to process and (b) the most relevant input allowing for the hearer's abilities and preferences.[2] The fact that the addressee of an ostensive act is in the privileged position of being able to presume optimal relevance, opens up a number of interesting issues and questions when it comes to clickbait headlines and, indeed, to online media and communication more generally.

Online communication often takes place in a so-called collapsed context (Wesch, 2009; Marwick and boyd, 2010). When we communicate offline we generally speak to one group of people in one context and another group in another context. When we move online, however, we may find ourselves communicating across groups of people and across contexts. A post on a personal Twitter account might, for example, be viewed by the user's family, friends and work colleagues as well as by total strangers. To navigate these collapsed contexts, users often address their utterances to a so-called imagined audience (Marwick and boyd, 2010; Brake, 2012; Litt, 2012). We may not know who will be in our actual audience, but when we construct our utterances we imagine who that audience might be. The writer of a clickbait headline cannot know who exactly will see and read their work, or indeed, when their work might be read. They must, however, imagine some sort of audience for whom their utterance will be relevant. This kind of broadcast communication is, of course, nothing new. Producers of print media, as well as radio and television broadcasters have been communicating with an unseen and imagined audience for as long as they have produced content. Sperber and Wilson (1986/95:158) briefly discuss how we might understand this type of communication from a relevance perspective as they consider discourse contexts in which there is no definite addressee. As they explain:

> In broadcast communication, a stimulus can even be addressed to whoever finds it relevant. The communicator is then communicating her presumption of relevance to whoever is willing to entertain it (Sperber and Wilson, 1986/95:158).

While Sperber and Wilson were writing before the development of social media and mass online communication, the same basic idea applies in online contexts. An online utterance will be relevant to whoever is willing to entertain it. Therefore, to be

[2] A speaker cannot be expected to provide information that they either do not have or are not willing share, even if that information would increase the overall relevance of the input.

successful, a clickbait headline must draw the reader's attention and convince him to entertain the presumption that he will find it optimally relevant. Of course we see the same requirement in offline newspaper headlines. A headline, whether online or offline, clickbait or non-clickbait, must persuade the reader that he should position himself as part of the imagined audience. That is, a successful headline will encourage the reader to entertain the presumption that the information will be relevant to him. Understanding the difference between clickbait and non-clickbait headlines, is a case of understanding the motivations and intentions of the writer and the discourse context in which the headline occurs. Next, I begin to do this by considering existing relevance-based analyses of newspaper headlines and by using ideas from relevance theory to illustrate the difference between these more traditional headlines and the headlines that we find on clickbait sites.

Ifantidou (2009) and Dor (2003) have both used ideas from relevance theory to analyse newspaper headlines. Ifantidou conducted a reader reaction study to examine the interpretation of newspaper headlines. She concludes that headlines should be treated as autonomous texts which can and should be interpreted in their own right. She explains that:

> Headlines are purposefully read for the sake of a quick and loose news update…headlines are intended as autonomous meaningful constructions and are (or should be) designed to be interpreted as such (p.702).

Dor (2003) reports on the headline development process at the news-desk of an Israeli national newspaper. He identifies a series of strategies that the copyeditors use when writing headlines. This leads him to an analysis of headlines as "relevance optimizers… [which] are designed to optimize the relevance of their stories for their readers" (p.696). Headlines are, he suggests, "negotiators between stories and readers" (p.720), and they "guide individual readers to those specific stories which would be worth their while to read in the full version". Crucially, he concludes that a key role of a headline is to function as "a relevance-based selection mechanism" and that by the time a reader has read a headline they have already received an optimally relevant summary of the story. Only readers who are particularly interested in the topic, or who enjoy reading news for its own sake, will, he suggests, go on to read the whole article.

Thus, for Dor headlines provide a label and/or summary so that readers can decide whether the rest of the content is likely to be relevant to them, and according to Ifantidou, readers may use headlines to get a quick and rough sense of what is going on in the news. For both, newspaper headlines are self-contained, autonomous texts. As illustrated by the examples in (1) to (4), clickbait headlines are not usually self-contained. Rather they raise questions and provide hints, but rarely tell the whole story. Writers of clickbait headlines are aiming to arouse curiosity in as many readers as possible so as to maximise clicks. They are less concerned with whether the content on the landing site is actually relevant to each reader than a journalist writing a more traditional headline would be. The clickbait headline need only appear to be relevant long enough to induce the reader to click, and creating an information gap is a key strategy in achieving this.

To illustrate how the difference between clickbait and non-clickbait headlines might be understood in relevance-theoretic terms, compare the headlines in (7) and (8). These were discussed by Upworthy founder Peter Koechley (2012) as examples of a non-clickbait (7) and a clickbait headline (8) representing the same news story.

(7) Obama says gay marriage should be legal.

(8) Now THIS is why I voted for Barack Obama.

The headline in (7) was written for the search engine Google and contains strategic key words which are likely to match the search terms of users. The headline in (8), on the other hand, was written to appear on social media platforms. Keywords and search engine optimisation are not as important on social media. Users of sites such as Facebook or Twitter are unlikely to enter search terms, but will instead be exposed to content via promoted posts and ad links whilst browsing feeds. While Google headlines are optimised to appear in search results, the success of a social media headline depends on whether it (a) attracts a user's attention as they are browsing, and (b) arouses curiosity and induces them to click on the link. We can start to understand how the headline writers achieve their aims by considering how interpretation is likely to proceed in each case.

When we process an utterance we are aiming to derive the speaker's overall intended meaning, and that means working out what the speaker intends to explicitly communicate and what she intends to imply or implicate. To derive an explicature for the headline in (7) the reader need only decode the linguistic forms and assign reference to the name *Obama*. For the purposes of this discussion, I assume the relevance-based approach to reference outlined in Scott (2020). Reference resolution is an inferential process which contributes to the derivation of the proposition expressed. To resolve reference, the reader must map a conceptual file representing what he takes to be the intended referent onto the corresponding slot in the logical form of the utterance. The referring expression used is a means by which the writer can provide guidance to the reader in this process. Different referring expression forms encode different content.

There is, of course, likely to be only one highly accessible referent for the name *Obama* in example (7). The reader is likely to resolve reference by mapping the argument slot onto her existing conceptual file for the Barak Obama who was, at the time of the headline, the President of The United States of America. This provides the reader with a truth-evaluable proposition, and if this interpretation achieves optimal relevance, the reader will accept this as the intended interpretation. There are various ways in which this interpretation could lead to a range of cognitive effects that justify the effort that the reader has put in to processing it. Processing this utterance may, for example, strengthen an assumption that the reader already held about Barak Obama. Alternatively, it may interact with other assumptions that the individual reader holds about him to yield further assumptions about Obama and/or gay marriage. The reader will then add these assumptions to his Barak Obama conceptual file. Thus, the headline in (7) functions in the way described by Dor (2003). It optimizes the relevance of the story by minimizing processing effort while making sure a sufficient number of contextual effects are deducible. A reader who is sufficiently interested may then decide to invest more effort in reading the full article

and deriving more effects from the details of the story. The headline is also an autonomous text in its own right, and provides the reader with a quick summary of the key information, in line with Ifantidou's conclusions.

Next consider the clickbait version of the headline in (8) ("Now THIS is why I voted for Barack Obama"). In this case, to derive an explicature, the reader must resolve reference on three referring expressions: *Barak Obama*, *THIS* and *I*. While resolution of the name is likely to follow the same process as it did in (7), there are no obvious candidate referents for the other two expressions. Instead, the reader is left with the assumptions in (9) and (10).

(9) There is a person (*I*') who voted for Barak Obama.

(10) That person did so for a particular reason.

These assumptions are unlikely to achieve many cognitive effects in their own right. Given easily accessible general knowledge about the world, including how people become presidents, and how and why people vote in elections, the reader is likely to already hold the assumptions in (9) and (10). He is likely to already assume that there are people who voted for Barak Obama and that they did so for particular reasons. Therefore, the headline in (8) does not lead to any cognitive effects in its own right. It is not an autonomous text in the sense Ifantidou discusses. Remember, however, that according to relevance theory all ostensive acts of communication carry a presumption that the speaker, or in this case, writer, was aiming at optimal relevance. Therefore, the fact that the reader has been led to entertain these assumptions means that he will assume that the writer thought that they would be optimally relevant, and the reader will try to work out how that could be.

Now consider the role that the demonstrative pronoun *this* plays in the interpretation of the headline in (8). We know that whatever *THIS* is intended to refer to, it is the reason that the writer voted for Obama. Demonstratives point something or someone out to the addressee. By drawing the reader's attention to whatever *THIS* is taken to refer to, the writer communicates that she thinks it will be relevant. Directing someone's attention towards something is an ostensive act that raises expectations of relevance whether the directing is performed using a pointing gesture or a linguistic demonstrative expression. The use of *THIS* in the headline is the equivalent of pointing towards the landing site, and it is reasonable for the reader to assume that they will find out what *this* is if they follow the link. In this way the writer overtly draws the reader's attention to the (as yet unrevealed) reason why she is voting for Obama. The reader can therefore derive the further assumption in (11).

(11) [The writer thinks] the reason [the writer] voted for Barak Obama is relevant [to the reader].

When someone points something out to us with a physical gesture we can expect to be rewarded with cognitive effects that satisfy our expectations of optimal relevance if we look in the direction of the gesture. Similarly, the use of the demonstrative in (8) raises our expectations that following the link will be optimally relevant. In each case, the writer has drawn the reader's attention to something, thus raising expectations of relevance. However, these expectations have not been satisfied. Instead, we have

an information gap, and a promise that the information to fill that gap and to satisfy our expectations will be found on the landing site.

Thus clickbait and non-clickbait headlines achieve their aims by interacting with the reader's expectations of relevance in different ways. Non-clickbait headlines provide a summary of the main article. Based on this summary readers will be able to update assumptions they hold about the topic and also decide whether to invest extra energy in reading the whole article. Clickbait headlines, on the other hand, function by creating an information gap. They leave readers with unanswered questions and unsatisfied expectations, and by doing so maximise the chances that a large number of readers will click through to the main site. The contrast in strategies is clear when we compare two directly parallel headlines such as (7) and (8). However, given that media sites which use the clickbait model and sites which follow a more traditional journalistic style often focus on different topics, stories and issues, this sort of direct analysis is not always possible. If, however, the different strategies result in different patterns of language use, then we should see this reflected over a larger data set. In the rest of this paper, I use comparative corpus analysis to investigate whether there is evidence of different linguistic strategies being used on sites that are associated with clickbait when compared with more traditional news providers.

4 Comparative Corpus Analysis

The comparative analysis was carried out on corpora of clickbait and non-clickbait headlines compiled by Chakraborty et al. (2016). The full version of each corpus contains 16,000 headlines from online news articles. Chakraborty et al. assigned the data to the clickbait or non-clickbait dataset based on the nature and reputation of the media source from which they were drawn. The clickbait corpus draws on headlines from online media sites which, according to Chakraborty et al. (2016), are associated with clickbait type content: Buzzfeed, Upworthy, ViralNova, Thatscoop, Scoopwhoop and ViralStories. These are sites which have a reputation for low quality content and misleading or sensationalist headlines, and which have extensive space given over to advertising content. The headlines in the non-clickbait corpus are drawn from media sources which have a reputation for quality journalism: WikiNews, The New York Times, The Guardian, and The Hindu. Articles published on WikiNews, for example, must comply with strict editorial guidelines. They must be specific, balanced and written in a neutral tone (Wikinews). For this reason Chakraborty et al. (2016:2) consider these to be the "gold standard for non-clickbaits".

As Chakraborty et al. note, the clickbait corpus contains headlines which would not normally be considered as clickbait. Buzzfeed, for example, has more traditional news content alongside more typical clickbait content. Similarly, the sources that were used for the non-clickbait corpus might also make use of techniques more commonly associated with clickbait, particularly, as Chakraborty et al. note, in sections promoting other sites or directing readers to other articles that they might like to read. As such, there is no claim that everything in the clickbait corpus is clickbait or that everything in the non-clickbait corpus is free from clickbait techniques. The two corpora represent two categories of online media sources. One of these is closely associated with clickbait content and the other is not. In the work

described in Chakraborty et al. (2016), the focus is on developing a clickbait detection browser extension which allows users to identify and block clickbait content. To do this, they require a close comparison of clickbait verses non-clickbait headlines, and for this reason, they work on a subset of the data from the corpora. They used only the “gold standard” Wikinews headlines as representative of non-clickbait headlines, and they recruited six volunteers to manually classify content from the clickbait corpus as clickbait or non-clickbait. This left them with a smaller corpus of headlines which they could be confident would be viewed as clickbait by readers. In this paper, the aim is to identify parts-of-speech that are overrepresented in those media associated with clickbait content. Therefore, the full versions of the corpora were used with a view to identifying how content from these sources differs from online content from more traditional news providers.

This study focuses on two categories of parts-of-speech, and demonstrates how theoretically underpinned analyses can be applied to empirical findings. In the analysis of definite referring expressions, I take a category that has already been identified in related literature as associated with clickbait. The analysis both provides extra support for the existing findings and offers an explanation for why we might see overuse in this category. In the analysis of superlatives and intensifiers, the data is approached from another perspective. Although not previously discussed in the literature, an overuse of superlatives and intensifiers is predicted by a relevance-theoretic analysis of information gaps. Our curiosity will be piqued by gaps which relate to highly relevant information, that is, information that promises to lead to cognitive effects. Promising to show us information that is the best, the biggest or the most intense that it can be would seem to be an effective strategy in creating an information gap and thus arousing curiosity. Comparative corpus analysis allows us to test whether this prediction is born out in the data.

The corpora were compared using the wmatrix corpus analysis and comparison tool (Rayson, 2008) which allows word level comparison, and which also tags each word within the data for part-of-speech and for semantic field. Wmatrix tags each word for grammatical part-of-speech using the UCREL Constituent Likelihood Automatic Word-tagger System (CLAWS). In this system, for example, the word *better* is tagged as a general comparative adjective (JJR) and the word *best* is tagged as a general superlative adjective (JJT). The UCREL USAS semantic codes are used for semantic tagging. In each case, a letter is used to indicate a particular semantic field and then numbers indicate sub-divisions within that field. Up to three pluses or minuses may be added to indicate positive or negative stance and the relative intensity of this. For example, comparative terms are tagged as A6. The subdivision A6.1 is then used to categorise terms which denote similarity or difference. The word *alike* is coded as A6.1(+), while the word *asymmetric* is coded as A6.1(-).

Once the data has been tagged, the software compares relative frequencies within the data and calculates a log-likelihood value to indicate overuse or underuse within one corpus relative to the other. Log-likelihood is a measure of statistical significance and a value of 3.84 or higher indicates a difference between the corpora which is significant at $p < 0.05$. Wmatrix also calculates a log ratio value as an indication of effect-size, where “every extra point of Log Ratio score represents a doubling in size

of the difference between the two corpora, for the keyword under consideration" (Hardie, 2014). So, for example, a log ratio of 3 indicates an effect size that is twice that of a log ratio of 2.

An initial analysis of the corpora revealed 79 parts-of-speech which are significantly ($p < 0.05$) overrepresented in the clickbait corpus. The analysis in this paper will, however, be restricted to the two categories discussed above: definite referring expressions and superlatives and intensifiers. The approach and analytical methods adopted here could, of course, be repeated for other categories which are significantly over or underrepresented in the clickbait corpus.

5 Definite Referring Expressions

5.1 Distribution in the Corpora

As discussed in Section 2, the use of forward referring expressions is considered a characteristic of clickbait headlines. Tables 1 and 2 provide details of the comparative analysis of definite/demonstrative determiners and pronouns from the clickbait and non-clickbait corpora. These results both lend support to the previous work on referring expressions in clickbait and they also allow us to more closely examine different parts-of-speech which fall within these general categories. All categories of pronouns were found to be overrepresented in the clickbait dataset, and when the overrepresented parts-of-speech from the clickbait corpus were ranked by effect size, 3 of the top 5 were definite referring expressions (*you, we,* and *these/those*).

Table 1: Distribution of demonstrative pronouns in the clickbait corpus, relative to the non-clickbait corpus.

Part of Speech	Log-Likelihood	Log Ratio	
These	+736.30	9.97	Significant overuse
This	+1752.78	4.96	Significant overuse
That*	+10.74	1.44	Significant overuse
Those	-0.01	Not significant	

*Calculated manually to ensure only demonstrative uses of that were included in the count.

Table 2: Distribution of personal pronouns in the clickbait corpus, relative to the non-clickbait corpus.

Part of Speech	Log-Likelihood	Log Ratio	
1st person singular subject (I)	+295.68	4.65	Significant overuse
1st person singular object (me)	+44.94	3.30	Significant overuse
1st person plural subject (we)	+932.55	6.42	Significant overuse
1st person plural object (us)	+57.94	1.62	Significant

			overuse
2nd person (you)	+6189.74	7.51	Significant overuse
3rd person singular subject (s/he)	+95.46	2.06	Significant overuse
3rd person singular object (her/him)	+32.15	2.19	Significant overuse
3rd person singular neutral (it)	+332.91	2.39	Significant overuse
3rd person plural subject (they)	+177.48	3.45	Significant overuse
3rd person plural object (them)	+41.27	2.64	Significant overuse

All definite determiners except for the plural distal form *those* were overrepresented to a significant degree, and plural proximal demonstrative *these* had the highest log ratio of any part-of-speech. All pronouns were also overrepresented in the clickbait corpus to a statistically significant degree, and the second person singular pronoun *you* had the highest log-likelihood score of any part-of-speech.

5.2 Relevance-theoretic analysis

The results from the comparative analysis of the Chakraborty et al. (2016) corpora suggest that clickbait headlines make use of personal and demonstrative pronouns significantly more often than non-clickbait headlines. How might we explain these patterns and what role do demonstratives and personal pronouns play in creating an information gap?

According to the comparative analysis in Table 1, the most overused demonstrative in the clickbait corpus is the plural proximal determiner *these*, followed by the singular proximal determiner *this*. Meanwhile, the singular distal determiner *that* shows the smallest effect with statistical significance, and the plural distal demonstrative *those* shows no significant difference, with very low frequency across the two corpora. Examples (12) to (15) illustrate these uses respectively.

(12) These Gadgets Will Make You Believe In The Future Of Food.

(13) This Goat Has Been Bullying His Tiger Friend.

(14) Anne Hathaway Comes To Jennifer Lawrence's Defense About That Phone Scolding Incident.

(15) How Do You Get Rid Of Those Annoying Gray Hairs.

First, let us consider why proximal demonstratives seem to be preferred over distal demonstratives. As the examples in (12) to (15) illustrate, the roles played by the proximal and distal demonstratives are different. In both (12) and (13) the proximal determiners in the headlines refer to something that the reader can expect to find on the landing site. We expect to find a list of gadgets when we follow the link in (12), and we expect to find a story about a goat when we click on the link in (13). The

reader will only have a skeletal conceptual file until he does so. The reader must click on the link in order to enhance his conceptual file for THESE GADGETS and THIS GOAT beyond a mere placeholder. Unless he does this, he will only be able to derive the basic existential propositions in (16) and (17).

(16) There exist gadgets that will make you believe in the future of food.

(17) There exists a goat that has been bullying his tiger friend.

At this point in the interpretation, the reader will have conceptual files representing GADGETS THAT WILL MAKE YOU BELIEVE IN THE FUTURE OF FOOD and A GOAT THAT HAS BEEN BULLYING HIS TIGER FRIEND. While the reader is likely to have entertained concepts, thoughts and propositions relating to gadgets, food, goats, and tigers before, it is less likely that they will have combined in the way suggested in these headlines. Thus these new conceptual files will be created but will remain little more than placeholders. The readers will be prompted to ask themselves about the gadgets and how they might make you believe in the future of food, and about the goat and how he has been bullying his tiger friend. The information gap in these cases is a result of the creation of new and perhaps unexpected conceptual files which comprise little more than a label. The label simultaneously creates an information gap by suggesting that more information exists and promising to fill that gap by pointing the reader to the landing site where his curiosity will be satisfied. The proximal demonstratives play an important part in this promise, as they suggest that the relevant information is to be found nearby on the associated landing site. The use of *this* and *these* overtly draws the hearer's attention to the information that will enrich the files and answer the reader's questions. The reader will assume that the writer is drawing the reader's attention to the goat and the gadgets because they are relevant. The use of the proximal demonstratives is the equivalent of pointing towards the landing site. As we saw in Section 3, pointing is an ostensive act, and by drawing our attention to something, the writer is communicating that she thinks it will be relevant. When someone points something out to us with a physical gesture we can expect to be rewarded with cognitive effects that satisfy our expectation of optimal relevance if we look in the direction of the physical pointing gesture. Similarly, the use of the demonstratives in (12) and (13) raises our expectations that following the link will provide us with cognitive effects.

The distal demonstratives in (14) and (15), however, have a different relation to the information gap. Notice that the proximal forms refer to the content on the landing site, and so, if we view the headline and article as one text, they are cataphoric. The distal forms, on the other hand, refer exophorically to information outside of the text, and the use of a distal demonstrative communicates that this is assumed to be shared, common knowledge. It is assumed that the reader has already heard about that phone scolding incident and has already thought about those annoying grey hairs. There is no need, it is assumed, for the reader to click through to the landing site to resolve these references. In part this helps to construct the imagined audience for the article. The intended audience for (14) is anyone who is aware of an incident relating to phones and Jennifer Lawrence, and for (15) it is anyone who has ever thought about annoying grey hairs. The key to the creation of an information gap lies elsewhere in these examples. The promise of relevance lies, not in finding out who

or what is being referred to, but rather in finding out how Anne Hathaway has come to Jennifer Lawrence's defense and how to get rid of the grey hairs.

We therefore see a difference in how the distal versus proximal demonstratives contribute to the information gap and thus to the success of the clickbait headlines. Proximals play a more direct role in creating an information gap. Use of distal demonstratives depends on reference to shared context, and while this can draw a reader in, it may also exclude anyone who does not have access to those shared assumptions. This might suggest a reason for why we do not see the same degree of overuse in the distal demonstrative determiners. Everyone is curious to find out how a goat could bully a tiger, but not everyone can relate to finding annoying grey hairs.

While both singular and plural proximal determiners are significantly more frequent in the clickbait corpus than in the non-clickbait data, there is a difference in effect size between the two. The plural proximal demonstrative *these* is the single word with highest log ratio in the entire corpus. Again, we can understand this difference in terms of relevance and cognitive effects. An article that features various gadgets has more potential in terms of cognitive effects than an article that only discusses one.

The part-of-speech category with the highest log-likelihood in the data set was pronouns. All categories of (non-possessive) pronouns were overrepresented in the clickbait corpus to a statistically significant degree (see Table 2). The first-person pronouns *I* and *me* were both overrepresented to a significant degree, and the second-person pronoun *you* had the highest log ratio of all the identified parts-of-speech. Examples illustrating typical uses of these pronouns are given in (18) to (22).[3]

(18) Which TV Female Friend Group Do You Belong In?

(19) How Intuitive Are You Really?

(20) 18 Animals Who Are Very Impressed With You And Your Life

(21) Leaving Home At 14 Was The Best Thing I Ever Did

(22) Running Helped Me Cope With Depression, But Then I Got Injured

Use of the second person pronoun *you* addresses the reader or readers directly. This contributes to the success of the clickbait by encouraging the reader to position himself as part of the intended audience. When we interpret an utterance as addressed to us, we process it with expectations of optimal relevance. Presenting the information as directly related to the reader, gives the writer another opportunity to create an information gap. There is an implicit promise that we will find out more about ourselves if we click on the link and engage with the content. The headlines in (18) and (19) ask the reader a question directly and imply that the answer can be found on the landing site. The headline in (20) takes a less direct approach, but

[3] The use of personal pronouns is not the only strategy used in each of these examples. However, for the purposes of this short discussion, I focus only on the role played by the pronouns.

nevertheless prompts us to ask questions about who these animals are and how they might be impressed with our lives.

In (21) and (22) first-person pronouns are used. This frames the content as a personal narrative. To find something relevant, we need to accept it as true. Use of first-person pronouns positions the writer as a first-person narrator who is drawing on personal experience. Again, this encourages a sense of the writer speaking directly to the reader, and we are led to believe that we will find a personal, first-hand account on the landing site which will provide us with relevant insights into the writer's experiences.

6 Superlatives and Intensifiers: The More the Better

6.1 Distribution in the Corpora

Table 3 shows the log-likelihood and log ratio for the adjectival and adverbial part-of-speech categories, including comparatives and superlatives.[4]

Table 3: Comparative distribution of adjective and adverb CLAWS categories in the clickbait corpus, relative to the non-clickbait corpus.

Part-of-Speech	Log-Likelihood	Use in the clickbait corpus, compared to the non-clickbait corpus	Log Ratio
General adjective (*old*, *good*)	-348.21	Significant underuse	-0.36
General comparative adjective (*older*, *better*)	-7.15	Significant underuse	-0.42
General superlative adjective (*oldest*, *best*)	+495.99	Significant overuse	2.33
General adverb	+2322.29	Significant overuse	2.33
Comparative general adverb (*better*, *longer*)	+95.60	Significant overuse	1.73
Superlative general adverb (*best*, *longest*)	+19.61	Significant overuse	1.46
Degree adverb (*very*, *so*, *too*)	+176.3	Significant overuse	1.49
Comparative degree adverb (*more*, *less*)	+0.22	No significance	0.15

[4] The CLAWS tagging system was developed by UCREL at Lancaster University, and they claim 96-86% accuracy. However, it has been noted that adverb categories are subject to tagging errors as they include a variety of ambiguous expressions (UCREL, 1996). The hope is that the comparative approach adopted here minimises the effect of any such errors.

Superlative degree adverb (*most*, *least*)	+346.9	Significant overuse	4.35

First consider the general adjective category. General and comparative adjectives were significantly underused in the clickbait corpus, relative to the non-clickbait corpus. However, superlative adjectives showed significant overuse. It is, therefore, not so simple as to say that clickbait headlines contain more adjectives than non-clickbait headlines. Adjectives occur across the headline data, clickbait or not. The pattern that does emerge, however, is that when adjectives are used in clickbait, superlative versions are favoured. Why describe something as *old* or *older*, when you can describe it as *the oldest*? Why describe something as *good* or *better*, when you could present it to your reader as *the best*?

We see a slightly different pattern in the adverb categories. The CLAWS part-of-speech tagging system tags for both general adverbs, which modify a verb, and a sub-set which it calls degree adverbs. Degree adverbs, also known as intensifiers, modify an adjective, adverb or determiner and indicate intensity. Typical examples include *so*, *as*, *very*, and *too.*

Significant overuse is seen across the categories of general adverb. However, in the category of degree adverb, we find overuse in the main category and in the superlative category, but no significant difference in comparative degree adverb use. The examples in (23) to (25) are illustrative examples from the clickbait corpus of headlines that are coded as containing degree adverbs, comparative degree adverbs and superlative degree adverbs respectively.

(23) 21 times Chris Pratt was too good for this world.

(24) This dog loves leaves more than you love anything.

(25) The most OMG movie scenes of 2015.

The overuse in degree adverbs as a general category suggests that writers of clickbait headlines are more likely to want to intensify a description than writers of non-clickbait headlines. When the headline involves a comparison of some sort, writers of clickbait headlines opt for presenting the content as superlative, rather than as simply comparative. Again, it seems, that extremes are preferred where possible.

Across the three groups, the most clear pattern that emerges is a significant overuse of superlative terms. There is a less clear pattern when it comes to the use of comparatives. They are overrepresented when tagged as general adverbs, underrepresented when tagged as adjectives and there is no significant difference when they are tagged as degree adverbs. Therefore, a closer analysis of comparative terms was carried out using the USAS semantic category A6.2, which codes comparative terms for the level of anomaly that they represent. A6.2+ denotes terms which compare degree of normality including *natural*, *common*, *ordinary,* and *familiar*, and A6.2- denotes terms which compare the degree of difference of unusualness including *unusual*, *strange,* and *odd*. The number of +/- symbols indicates strength of positivity/negativity. The corpus results for these categories are given in Table 4.

Table 4: Comparative distribution of USAS semantic categories of comparative terms in the clickbait corpus, relative to the non-clickbait corpus.

USAS Category	Description	Example	Log-Likelihood	Log Ratio	
A.6.1	Similar/Different	Compare	+4.12	1.99	No significant difference
A6.2+*	Comparing: Usual	natural, common, ordinary, familiar	+2.28	0.34	No significant difference
A6.2+++	Comparing: Usual	cliché	+2.33	1.66	No significant difference
A6.2-	Comparing: Unusual	weird, incredible, bizarre, mind-blowing, freaky, strange	+100.37	1.88	Significant overuse
A6.2--	Comparing: Unusual	strange, strangers	+12.84	4.12	Significant overuse
A6.2---	Comparing: Unusual	weirdest	+24.50	5.06	Significant overuse

*Manually adjusted to remove occurrences of representatives as part of the phrase House of Representatives.

When the comparative term is used to indicate that two things are similar or usual, we find no significant difference between the two corpora. However, when the terms are used to compare how unusual or strange things are, we find significant overuse in the clickbait corpus. Furthermore, the more unusual the relationship is, the more the associated term will be overused in clickbait headlines. Again, we find evidence that clickbait deals in extremes, and that when making comparisons, the more unusual things are, the better. We find a similar pattern when we carry out an analysis looking at the distribution of terms in semantic categories A13 and A14. Category A13.x covers various degree terms including boosters such as *really*, *so*, *very,* and *incredibly*, and minimizers such as *little* and *least*. Category A14 is for particularizers and exclusivizers such as *just* and *only*. The results from this analysis are shown in Table 5.

Table 5: Comparative distribution of USAS semantic categories A13 and A14 in the clickbait corpus, relative to the non-clickbait corpus.

USAS Category	Description	Example	Log-Likelihood	Log Ratio	
A13.1	Degree: Non-	even, as,	+5.23	0.88	No

	Specific	relatively			significant difference
A13.2	Degree: Maximizers	most, totally, perfectly	+865.59	4.84	Significant overuse
A13.3	Degree: Boosters	extremely, enormously, as hell	+459.38	2.94	Significant overuse
A13.4	Degree: Approximators	about, really, almost	-0.12	0.88	No significant difference
A13.5*	Degree: Compromisers	rather, some, pretty	+14.80	+3.83	Significant overuse
A13.6	Degree: Diminishers	under, slightly	+21.10	1.21	Significant overuse
A13.7	Degree: Minimizers	hardly, barely, little	-170.22	-3.41	Significant underuse
A14	Exclusivizers/ Particularizers	totally, perfectly, most	+486.92	+3.52	Significant overuse

*Manually adjusted to remove occurrences of pretty that are part of the name Pretty Little Liars.

Again we see a pattern of overuse for terms associated with extremes and exaggeration (boosters, maximizers and particularizers), while the category of minimizers is underused. The distribution data for compromisers and diminishers seems to run against this trend. This category includes terms that are typically associated with reduced intensity, and yet they are overrepresented in the corpus. A closer inspection of the data reveals an explanation for this apparent anomaly. While terms classified as diminishers were overrepresented, we find a very different distribution and use of these terms when we compare the two corpora and when we look at collocations and the context in which the terms appear. Tables 6 and 7 show the number of times each term in these categories appears in each corpora. The clickbait corpus contained a total of 152,403 words and the non-clickbait corpus contained 120,732 words. While the difference in corpus size means that we cannot draw conclusions from a direct comparison of the absolute frequencies of each word, the details are given here for completeness. What we can see, however, is that there is a difference in the individual words that have been categorised as diminishers in the two corpora. A closer look at the collocations for these words also reveals differences between the two datasets.

Table 6: Number of Diminishers in the Clickbait and Non-Clickbait Corpora

Clickbait Corpus		Non-Clickbait Corpus	
Slightly	40	Less	11
Mildly	18	Partially	10

A_little	17	Slightly	5
Less	11	A_little	3
Simply	10	A_bit	3
Under	7	Under	2
A_bit	1	But	1
A_little_bit	1	Simply	1

Table 7: Number of Compromisers in the Clickbait and Non-Clickbait Corpora

Clickbait Corpus		Non-Clickbait Corpus	
Pretty	23	Quite	1
Quite	5		
Rather	3		
Slightly	1		

The most frequently occurring diminisher was the word *slightly* which modified the following terms in the clickbait data: *obsessed*, *guilty*, *disturbing*, *horrifying*, *gross*, *infuriate*, *terrible*, *unexpected*, *incorrect*, *wrong*, *deranged*. In each case, it is used to mitigate an extreme and negative descriptor. Some examples are given in (26) and (27).

(26) 9 Slightly Disturbing Facts That Will Stop You From Cleaning Your Ears Ever Again.

(27) 17 Things You'll Understand If You're Slightly Obsessed With Singing In The Car.

In the non-clickbait data, however, *slightly* was used exclusively to modify verbs, and it occurred in headlines describing falls or rises in the economy or in political popularity measures, as illustrated in (28).

(28) US presidential candidate John McCain now leads slightly in the polls.

By far the most frequently used compromiser in the clickbait corpus was the word *pretty,* and it was found modifying terms including *cute*, *whimsical*, *crazy* and *gross* in examples such as (29).

(29) This Lion Got Into A Wheelbarrow At A Zoo And It Was Pretty Whimsical.

The USAS tagging guide (Archer et al. 2002) defines compromisers as terms “that express an assumed norm, or call into question the appropriacy of X”. However, the examples of pretty used to modify an adjective in the clickbait corpus appeared to rather be cases of informal understatement for comic effect.

In the next sub-section, I consider how we can understand these patterns and their contribution to the information gap in terms of their contributions to relevance and to the interpretation process.

6.2 Relevance-theoretic analysis

The corpus-analysis in Section 6.1 reveals a pattern in which superlatives and linguistic terms associated with extremes are overused in clickbait headlines relative to the non-clickbait corpus. In this section, I conduct a close analysis of typical examples from these data. Again, I use ideas from the relevance-theoretic framework to consider why these patterns may arise. How, that is, do these terms contribute to an information gap and how does interpretation proceed?

Consider the clickbait headline in (30), which was also briefly discussed in Section 2. The expression *mind-blowing* falls under USAS semantic tag A6.2- (comparing unusual).

(30) 12 Mind-blowing Ways to Eat Polenta.

As we saw in Section 2, there seems to be very little that might reasonably be considered mind-blowing on the landing site for this link. The reader is taken to a page which contains pictures of polenta dishes alongside links to external pages featuring recipes using polenta. Given this content, an alternative non-clickbait headline might be that given in (31).

(31) 12 recipes for cooking with polenta.

Both (31) and its clickbait equivalent are grammatically parallel. They are both noun phrases, rather than full sentences, and so they appear to be labels for the webpages to which they link. To interpret them and to form an expectation about what he will find on the landing site, the hearer must construct a concept onto which the noun phrase maps. This concept will take the form of a conceptual file containing all of the available linguistic, semantic, logical, and encyclopaedic information about the concept (Scott, 2020). Consider how this act of interpretation is likely to proceed in each case. Let us begin by assuming that the reader of (30) and (31) already has a conceptual file for polenta. Processing either utterance will activate this file. A conceptual file for polenta may contain some basic information and facts about polenta, including, for example, that it is a foodstuff and perhaps that it originally comes from Italy. General assumptions about foodstuffs and cooking make it likely that the assumption in (32) will also be manifest to the reader.[5] The non-clickbait headline thus makes this assumption accessible and ready for onward inference.

(32) There exist at least 12 recipes for cooking with polenta.

What happens next will depend on the interests of the reader and what he finds relevant. If he is interested in cooking and/or has a particular liking for polenta, he may infer that clicking on the link will take him to information that he will find relevant. He makes an evaluation about whether the content on the landing site will be relevant for him, and in doing so, he self-selects into the audience. Alternatively, he might decide that the content is not relevant for him, and, having thus positioned

[5] According to Sperber and Wilson (1986/95:39), an assumption "is manifest to an individual at a given time if and only if he is capable at that time of representing it mentally and accepting its representation as true or probably true".

himself outside of the intended audience, he will no longer presume optimal relevance for himself, and is likely to discontinue processing. He has, in effect, eavesdropped on a headline that was really intended for other people.

Now consider how the interpretation of the clickbait version of the headline is likely to proceed. Again, given general, encyclopaedic assumptions about food and cooking, the assumption in (33) is likely to already be manifest to the reader.

(33) There exist ways to eat polenta.

Processing the headline will activate this existing assumption. However, the headline also makes the more specific assumption in (34) manifest.

(34) There exist at least 12 mind-blowing ways to eat polenta.

It is less likely that the assumption in (34) will have previously been manifest to the reader. Whereas he is already likely to associate foodstuffs with recipes generally, it is less likely that any existing conceptual file for polenta will include associations with ways of eating it that would normally and accurately be described as *mind-blowing*. The use of the phrase *mind-blowing* will activate in the reader a range of associations and impressions linked to the sort of things or events that generally warrant the description *mind-blowing*. Given the meaning of this phrase, it is reasonable to assume that these will be amazing, incredible, and unexpected things. It is also unlikely that ways of eating polenta will previously have been considered to be amongst this set of things. Thus while previously held assumptions about the world might have led the reader to deduce the assumption in (32), this is not the case with the assumption in (34). Reading the headline in (34) leads the reader to entertain a completely new, and perhaps surprising assumption.

A label on a product must communicate what that product contains so that the reader can decide whether it is useful for him or not. Similarly, the label for a webpage must allow the reader to decide whether he wants to visit the site, and whether doing so is likely to be relevant for him. As we have seen, the success of clickbait sites depends on convincing as many readers as possible that the content will be relevant for them. How does this work in the case of (30)? In the process of accessing the assumption in (34) the reader will have constructed a conceptual file representing the concept MIND-BLOWING WAYS TO EAT POLENTA. As we have seen, he is unlikely to have previously entertained such a concept. If he goes no further and does not click on the link, then this conceptual file will remain fairly empty and is unlikely to be useful in future inferences. What, the reader will ask himself, could these mind-blowing ways to eat polenta be? How might the process of eating polenta be appropriately described as mind-blowing? The reader is unlikely to be able to infer answers to these questions, but the headline promises that the answers are just a click away. In relevance terms, this seems like a good investment of effort. If he clicks on the link and reads the article, his relatively bare conceptual file will be enhanced with new information, and what is more, that information will be mind-blowing. Whatever assumptions the reader might previously have held about polenta and how to eat it, the information on the landing site will, it is claimed, blow the reader's mind and force him to update these assumptions. The information gap has been created by encouraging the reader to construct a new conceptual file (in this

case for MIND-BLOWING WAYS TO EAT POLENTA) but not providing any meaningful content for this file other than the description or label.

Clark (2014) included the advice to "[p]ut odd and interesting things next to each other" as one of his "Top 8 Secrets of How to Write an Upworthy Headline". The contrast between (30) and (31) reveals how this contributes to the success of a clickbait headline in relevance-theoretic terms. By presenting "odd and interesting" things together, the reader is encouraged to make new associations and new connections between concepts. These new associations are likely to be relevant in their own right. If, indeed, the content on the landing site does turn out to be mind-blowing it will be highly relevant and will provide the reader with a wide range of cognitive effects.

Given this interpretation strategy, the overuse of superlatives, particularly when the descriptive element is unusual, and the overuse of terms which function as boosters and maximisers is to be expected.

7 Conclusion

In this paper I have examined features of clickbait headlines using comparative corpus analysis and I have used insights from relevance theory to suggest some pragmatic analyses of the distributional patterns that we observe. I have argued that we can understand the information gap in terms of a promise of future relevance and that the power of clickbait headlines to make us click can be understood as an exploitation of our natural tendency to seek out relevance. I have also argued that one technique is to encourage readers to construct a new conceptual file representing something mentioned in the headline, but providing little or no content for that file. The reader is then driven to click through to the landing site in the hope that the article will contain information that he can add to the file and which will yield cognitive effects.

I have discussed individual examples throughout this analysis. However, as revealed by the media insiders discussed in Section 2, one headline does not simply rely on one technique. Rather each headline is a carefully crafted piece of work, designed to create an information gap that many find impossible to resist. Comparative corpus analysis provides us with a means of identifying general patterns and tendencies in the data, and then close examination of individual examples allows us to analyse these strategies within a theoretical framework of general language use, communication, and cognition.

I am not, of course, claiming to have identified all of the linguistic or pragmatic techniques used to entice readers to click on a link. The comparative analysis conducted here revealed a large number of significant differences between the clickbait and non-clickbait headlines in terms of part-of-speech and semantic tags. Exploring more of these in detail against a framework of utterance interpretation will not only provide insight into clickbait as a phenomenon, but will also tell us more about how these elements function in non-clickbait contexts.

References

Archer, Dawn; Wilson, Andrew & Rayson, Paul, 2002. *Introduction to the USAS Category System.* [Online] Available at: http://ucrel.lancs.ac.uk/usas/usas_guide.pdf [Accessed 28 September 2020].

Biyani, Prakhar; Tsioutsiouliklis, Kostas & Blackmer, John, 2016. "8 amazing secrets for getting more clicks": Detecting clickbaits in news streams using article informality. *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, 94-100.

Blom, Jonas Nygaard & Hansen, Kenneth Reinecke, 2015. Click bait: Forward-reference as lure in online news. *Journal of Pragmatics,* 76, 87-100.

Brake, David Russell, 2012. Who do they think they're talking to? Framings of the audience by social media users. *International Journal of Communication,* 6, 1056-1076.

Carston, Robyn, 2002. Thoughts and Utterances: The Pragmatics of Explicit Communication. Blackwell, Oxford.

Chakraborty, Abhijnan; Paranjape, Bhargari; Kakarla, Sourya & Ganguly, Niloy, 2016. *Stop Clickbait: Detecting and Preventing Clickbaits in Online News Media* IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), San Fransisco.

Chen, Yimin; Conroy, Niall J. & Rubin, Victoria J., 2015. *Misleading online content: Recognizing clickbait as false news.* Proceedings of the 2015 ACM on Workshop on Multimodal Deception Detection,15-19.

Clark, Billy, 2013. *Relevance Theory.* Cambridge University Press, Cambridge.

Clark, Peter Roy, 2014. *Top 8 Secrets of How to Write an Upworthy Headline.* [Online] Available at: https://www.poynter.org/reporting-editing/2014/top-8-secrets-of-how-to-write-an-upworthy-headline/ [Accessed 23 April 2019].

Dor, Daniel, 2003. On newspaper headlines as relevance optimizers. *Journal of Pragmatics,* 35, 695-721.

French, Tom, 2002. Serial narratives: Their power comes from “that delicious sense of enforced waiting”. *Nieman Reports,* 56(1), 48-50.

Hardie, Andrew, 2014. *Log Ratio: An Informal Introduction.* [Online] Available at: http://cass.lancs.ac.uk/log-ratio-an-informal-introduction/ [Accessed 28 September 2020].

Ifantidou, Elly, 2009. Newspaper headlines and relevance: Ad hoc concepts in ad hoc contexts. *Journal of Pragmatics,* 41(4), 699-720.

Interactive Advertising Bureau, 2017. *IAB Internet Advertising Revenue Report: 2016 Full year Results,* PricewaterhouseCoopers Ltd.

Interactive Advertising Bureau, 2019. *IAB Internet Advertising Revenue Report: 2018 Full Year Results,* PricewaterhouseCoopers Ltd.

Keay, Lara & McManus, Leigh, 2019. JK Rowling WINS legal battle with “utterly dishonest” former PA who used her credit card to splash out on coffee and toiletries - as Harry Potter author reveals she sued her to “protect reputation of her staff”. [Online] Available at: https://www.dailymail.co.uk/news/article-6886399/JK-Rowlings-personal-assistant-PA-took-18-000-Harry-Potter-fraud-case.html [Accessed 08 November 2019].

Koechley, Peter, 2012. *Why The Title Matters More Than The Talk.* [Online] Available at: https://blog.upworthy.com/why-the-title-matters-more-than-the-talk-867d08b75c3b [Accessed 08 November 2019]
Litt, Eden, 2012. Knock, knock. Who's there? The imagined audience. *Journal of Broadcasting & Electronic Media,* 56(3), 330-345.
Loewenstein, George, 1994. The psychology of curiosity: A review and reinterpretation. *Psychological Bulletin,* 116(1), 75-98.
Marwick, Alice E. & boyd, danah, 2010. I tweet honestly, I tweet passionately: Twitter users, context collapse, and the imagined audience. *New Media and Society,* 13, 96-113.
Pengnate, Supavich, 2016. *Measuring emotional arousal in clickbait: Eye-tracking approach.* Twenty-second Americas Conference on Information Systems San Diego, 1-9.
Phillips, Ellie, 2019. *MailOnline.* [Online] Available at: https://www.dailymail.co.uk/tvshowbiz/article-6889875/Duncan-Bannatyne-70-wife-Nigora-Whitehorn-39-pack-PDA-dining-London-hotspot.html [Accessed 08 November 2019]
Potthast, Martin; Gollub, Tim; Komlossy, Kristof; Schuster, Sebastian; Wiegmann, Matti; Garces Fernandez, Erika Patricia; Hagen, Matthias & Stein, Benno, 2018. Crowdsourcing a Large Corpus of Clickbait on Twitter. *Proceedings of the 27th International Conference on Computational Linguistics,* Association for Computational Linguistics, Santa Fe. New Mexico, 1498-1507.
Potthast, Martin; Köpsel, Sebastian; Stein, Benno & Hagen, Matthias, 2016. Clickbait detection. In: Ferro, N. et al. (eds) *Advances in Information Retrieval. ECIR 2016. Lecture Notes in Computer Science, 9626,* Springer, Cham, 810-817.
Rayson, Paul, 2008. From key words to key semantic domains. *International Journal of Corpus Linguistics,* 13(4), 519-549.
Scott, Kate, 2020. Referring Expressions, Pragmatics, and Style: Reference and Beyond. Cambridge University Press, Cambridge.
Smith, Ben, 2014. *Why Buzzfeed doesn't do clickbait: You won't believe this one weird trick..* [Online] Available at: https://www.buzzfeed.com/bensmith/why-buzzfeed-doesnt-do-clickbait [Accessed 08 November 2019]
Sperber, Dan & Wilson, Deirdre, 1986/95. *Relevance: Communication and Cognition.* 2nd edition with postface ed. Blackwell, Oxford.
UCREL, 1996. *A Post-editor's guide to CLAWS7 tagging.* [Online] Available at: http://www.natcorp.ox.ac.uk/docs/claws7.html [Accessed 08 November 2019]
Wesch, Michael, 2009. Youtube and you: Experiences of self-awareness in the context collapse of the recording webcam. *Explorations in Media Ecology,* 8(2), 19-34.
Wilson, Deirdre & Sperber, Dan, 2012. *Meaning and Relevance.* Cambridge University Press, Cambridge.