Article

Discourse & Communication
2015, Vol. 9(2) 149–172
© The Author(s) 2015
Reprints and permissions:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/1750481314568548
dcm.sagepub.com
SAGE

How can computer-based methods help researchers to investigate news values in large datasets? A corpus linguistic study of the construction of newsworthiness in the reporting on Hurricane Katrina

Amanda Potts
Lancaster University, UK

Monika Bednarek
University of Sydney, Australia

Helen Caple
University of New South Wales, Australia

Abstract

This article uses a 36-million word corpus of news reporting on Hurricane Katrina in the United States to explore how computer-based methods can help researchers to investigate the construction of newsworthiness. It makes use of Bednarek and Caple's discursive approach to the analysis of news values, and is both exploratory and evaluative in nature. One aim is to test and evaluate the integration of corpus techniques in applying discursive news values analysis (DNVA). We employ and evaluate corpus techniques that have not been tested previously in relation to the large-scale analysis of news values. These techniques include tagged lemma frequencies, collocation, key part-of-speech tags (POStags) and key semantic tags. A secondary aim is to gain insights into how a specific happening – Hurricane Katrina – was linguistically constructed as newsworthy in major American news media outlets, thus also making a contribution to ecolinguistics.

Keywords

News values, newsworthiness, news discourse, corpus linguistics, tagging

Corresponding author:
Amanda Potts, Department of Linguistics and English Language, Lancaster University, Lancaster LA1 4YL, UK.
Email: a.potts@lancaster.ac.uk

Introduction

This article is concerned with analysing news values (Galtung and Ruge, 1965) with corpus linguistic techniques. It does so through a case study on one culturally important event: Hurricane Katrina, one of the costliest and deadliest storms in American history. Instrumental in how Hurricane Katrina has been conceptualised as a natural disaster is the vast amount of news reporting that has been published in the American news media. This is because the news media have traditionally been viewed as providing a 'window on the world' (Tuchman, 1978: 1), and as a result, we may say that news 'defines and shapes' events (Tuchman, 1978: 184). In addition, Bednarek and Caple (2014) suggest that linguistic analysis can reveal how an event is established as newsworthy.

This article therefore aims to make two contributions: first and foremost, it contributes to the study of newsworthiness/news values and the evaluation of corpus linguistic techniques. Second, by using a case study that focuses on an important environmental event, this article also makes a contribution to ecolinguistics, which is concerned with the relation between language and the environment (see Bednarek and Caple, 2010; Fill and Mühlhäusler, 2001, for overviews). From an ecolinguistic perspective, there is not much work on the reporting of environmental matters in the media (Mühlhäusler, 2003), although climate change news has seen some attention (e.g. Bell, 1991; Carvalho, 2007; Grundmann and Krishnamurthy, 2010), and Johnson et al. (2010) have analysed American television news on Hurricane Katrina from a critical discourse analysis (CDA) perspective. With respect to environmental disaster reporting, Cottle (2009) argues that we 'need to better understand how the news media variously enter into their constitution and forms of response' (p. 70).

A discursive approach to news values

The body of research on news values is vast and diverse, existing mostly within the discipline of journalism and communications studies (see Caple and Bednarek, 2013, for a review). By and large, news values are said to drive what makes the news, and the focus of this discipline is on answering the question *why* events make it into the news media. Within linguistics, news values are mainly discussed in passing, with only some exceptions, such as Bell (1991), Bednarek (2006) and Cotter (2010). Bell (1991: 76) proposes that language can be used to *maximise* news value, and Bednarek (2006) takes up Bell's suggestion with a focus on evaluative language, whereas Cotter (2010) is concerned with ethnographic newsroom analysis. Caple and Bednarek (2013) and Bednarek and Caple (2014) provide further discussion of linguistic approaches to news values.

From the 'discursive' perspective that we are taking here, we are concerned with *how* an event is constructed as newsworthy through semiotic resources such as language, as summarised in Table 1 (for comprehensive discussion see Bednarek and Caple, 2012a, 2012b, 2014).[1]

One aim of such an approach is to analyse news reporting around an event to identify what values are emphasised and how each news value is constructed linguistically. To clarify, we will undertake a brief example analysis. The following text is from the front-page story of the *Times-Picayune* (New Orleans), published on Tuesday, 30 August 2005, the day after Katrina made landfall in Louisiana:

Table 1. News values (alphabetical order).

News value	Definition	Linguistic resources
Consonance	The event/issue (including but not limited to the people, countries or institutions involved) is discursively constructed as (stereo)typical in the view of the target audience	Evaluations of expectedness (*notorious*, *routine for*, *famed for…*); similarity with past (*yet another*, *markedly similar to*, *once again…*); constructions of stereotypes and archetypal stories
Eliteness	The event/issue (including but not limited to the people, countries or institutions involved) is discursively constructed as of high status or fame in the eyes of the target audience	Various status markers, including labels, recognised names, evaluations of importance, descriptions of achievement (*experts at Harvard university*, *a high-profile arrest*, *Barack Obama*, *the Oscars*, *a key federal government minister*, *top diplomats*, *were selling millions of records a year…*)
Impact	The event/issue is discursively constructed as having significant effects or consequences (not necessarily limited to impact on the target audience)	Evaluations of significance (*historic*, *crucial…*); reference to important or relevant consequences (*note that will stun the world*, *Australia could be left with no policy*, *leaving scenes of destruction…*)
Negativity	The event/issue is discursively constructed as negative for the target audience	Negative evaluative language (*terrible*, *slaughter…*); reference to negative emotion/attitude (*distraught*, *worried*, *condemn*, *criticise…*); negative lexis (*conflict*, *damage*, *death*, *crisis*, *row…*)
Novelty	The event/issue is discursively constructed as new and/or unexpected for the target audience	Indications of newness (*fresh*, *new…*); evaluations of unexpectedness (*astonishing*, *strange…*), comparisons that indicate unusuality (*the first time since 1958…*); references to surprise (***shock** at North Cottesloe quiz night…*); references to unusual happenings (*British man survives 15-storey plummet…*)
Personalisation	The event/issue is discursively constructed as having a personal or 'human' face involving personal experiences, including eyewitnesses (usually by non-elite actors)	References to 'ordinary' people, their emotions, experiences (*Charissa Benjamin and her Serbian husband*; *'It was pretty bloody scary'…*)
Proximity	The event/issue is discursively constructed as geographically or culturally near the target audience	Explicit references to place or nationality near the target audience (*An Australian*, *Canberra woman…*); references to the nation/community (*the **nation**'s capital…*); inclusive first person plural pronouns (***our** nation's leaders…*)

(Continued)

Table 1. (Continued)

News value	Definition	Linguistic resources
Superlativeness	The event/issue is discursively constructed as being of high intensity or large scope/scale (in the view of the target audience)	Quantifiers (*many*, *all*, *thousands*…); intensifiers (*sensational*, *dramatically, extreme*…), including intensified lexis (*epidemic, smashed, stun*, *wreck*…); repetition (*building after building*…); metaphor/personification/simile (*a tsunami of crime; epidemic swallowing Sydney*, *looked like the apocalypse*…); comparison (*most shocking child abuse case*…)
Timeliness	The event/issue is discursively constructed as timely in relation to the publication date: as recent, ongoing, about to happen or otherwise relevant to the immediate situation/time (i.e. current or seasonal)	Explicit time references (*yesterday*, *today*, *now*, *within days*…); verb tense and aspect (*have been trying*, *is preparing*…); references to seasonal/current happenings

> Headline: CATASTROPHIC
>
> Sub-head 1: STORM SURGE SWAMPS 9TH WARD, ST. BERNARD
>
> Sub-head 2: LAKEVIEW LEVEE BREACH THREATENS TO INUNDATE CITY
>
> Lead paragraph:
>
> Hurricane Katrina struck metropolitan New Orleans on Monday with a staggering blow, far surpassing Hurricane Betsy, the landmark disaster of an earlier generation. The storm flooded huge swaths of the city, as well as Slidell on the north shore of Lake Pontchartrain, in a process that appeared to be spreading even as night fell.

The headline, CATASTROPHIC, is an example of maximally intensified negative lexis, and simultaneously constructs Negativity and Superlativeness. The effects of the event (Impact) and its scale (Superlativeness) are emphasised through expressions such as *swamps*, *threatens to inundate*, *with a staggering blow*, *far surpassing*, *flooded huge swaths* and *spreading even as night fell*. The event is also established as highly relevant through references to places that would be well known to the local target audience of *The Times-Picayune*, constructing Proximity (e.g. *9th ward*, *St Bernard*, *Lakeview Levee*, *New Orleans*, *Slidell*, *Lake Pontchartrain*). Timeliness is also established, as the event is identified as recent (e.g. *on Monday*) and ongoing (*appears to be spreading as night fell*), with the use of the present tense in the sub-headlines a conventionalised device of establishing a sense of immediacy (*swamps*, *threatens*).

This brief analysis demonstrates how the *Times-Picayune* emphasises the event's high negative and ongoing impact for its local target audience. Methodologically, we have proceeded via manual discourse analysis. But, as Hunston (2011) notes, 'ways have to be found to translate research questions prompted by discourse analysis into corpus interrogation questions' (p. 167). In this article, we will test whether discursive news values analysis (DNVA) can be applied on a much larger scale through corpus linguistic techniques.

Corpus design and techniques

The Katrina corpus

This study is based on a specialised corpus of newspaper texts from an online news aggregator (Potts, 2013). The sampling unit is a single article (containing headline, byline, dateline, source name, publication date and publication section, where available), and the sampling frame consists of all database articles meeting the following parameters:

1. originally published in a major American print publication;
2. published between 25 August 2005 and 31 August 2006 ;
3. containing search terms *Katrina* AND (*hurricane* OR *storm* OR *flood* OR *disaster*).

This resulted in a corpus of 36,736,679 words in 41,964 texts from 24 publications. As texts were gathered from an online database, metadata pertaining to the original section and page number of texts is erratically included and often unreliable. Therefore, texts from various subgenres of reporting (e.g. letters to the editor, obituaries) are included in the corpus. Although the overwhelming majority of data is news, it is not possible to identify the exact proportion of other newspaper matter without manually checking metadata in 41,964 texts. However, by focusing on high-frequency items we may be able to limit the influence of non-news and the influence of special topic news (e.g. business, sports).

For corpus analysis, we used the web-based corpus analysis system CQPweb (Hardie, 2012), which annotated the data using the seventh version of the Unit for Computer Research on the English Language (UCREL) Constituent Likelihood Automatic Word-tagging System (CLAWS), and the UCREL Semantic Analysis System (USAS), assigning part-of-speech (POS) and semantic tags (semtags) to each text. CLAWS has an error rate of approximately 1.5%, with around 3.3% of ambiguities (such as multiple tags assigned to words that may belong to two classes) left unresolved (Leech et al., 1994: 625). Using the same corpus as in this article, Potts (2013) found that in approximately 90% of cases, USAS provided the most appropriate semtag for a word in its given context (in 85% of cases this was listed first in the string of candidates and deemed 'most likely' by the tagger). This leaves approximately 10% of (mostly low-frequency) types incorrectly matched or unmatched. These were deemed acceptable confidence thresholds for this study.

Using corpus linguistic techniques to analyse news values

In this section, we briefly explain the corpus techniques that have so far been employed in DNVA. Bednarek and Caple (2012b) use frequency lists and concordancing for analysis of news values in *one* environmental news story, complementing this with manual multimodal discourse analysis. Bednarek and Caple (2014) suggest that various corpus linguistic techniques can be used to study newsworthiness. However, for reasons of scope, they focus only on word/bigram frequency and keywords, applying two different methods to a small corpus (approximately 70,000 words).

The first method is to manually identify, from frequency/keywords lists, those forms that seem to have the potential to construct news values. These are called ‘pointers’ (Bednarek and Caple, 2014: 145) to newsworthiness and examples include *yesterday* (Timeliness), *chief executive* (Eliteness), *the most* (Superlativeness), *England* (Proximity for a British target audience), *attack* (Negativity) and *the first* (Novelty). Concordance analysis allows for qualitative examination of these forms, while range analysis can help to identify in how many different texts they occur.

The second method is to investigate topic-associated words using concordancing to gain insights into which news values are associated with particular concepts or entities. For example, the word *car*, which has no apparent potential to construct news value, is frequent in Bednarek and Caple’s corpus, and concordances can show what news values are discursively associated with it.

In this article, we provide a first case study on a *large* newspaper corpus, including both synchronic and diachronic aspects of analysis. The main aim of this case study is to apply and test the integration of corpus techniques in DNVA. We evaluate corpus techniques that were *not* tested previously, in particular tagged lemma frequencies, collocation, key part-of-speech tags (POStags) and key semantic tags.

The main statistic employed here is log likelihood (LL). We use the LL ratio for ranking keyness and identifying statistically significant collocates by measuring the confidence that results are not due to chance (Rayson et al., 2004). This tends to favour higher-frequency words, and is preferred here for this reason to avoid a potential skewing by low frequency and therefore less-dispersed collocates, which might occur in special topic news, obituaries or letters to the editor (cf. section ‘The Katrina corpus’). For this reason, we also set a minimum of five for frequency of node, collocate and collocation. We mainly use a window span of four words to the left to four words to the right (4L-4R), as this is the most commonly used in corpus linguistic studies on English, although a window of five is also common in computational linguistics (McEnery and Hardie, 2012: 129). We recognise that choice of statistic is not unproblematic. Ranking keywords by LL has recently been problematised (e.g. Gabrielatos and Marchi, 2012), but it remains the most widely used keyness statistic, and as a result, no other option has been offered to date in CQPweb and Wmatrix, the concordancers utilised here. Furthermore, since different collocation measures mean different results, this is an area where future experimentation could be undertaken – which collocation measures work best for identifying news values?

Although the main objective of this study is an investigation into the use of corpus-based techniques to explore newsworthiness, we are also working with discourses. Therefore, as a side effect, we will also gain insights into how a specific happening – Hurricane Katrina – was linguistically constructed as newsworthy in major American newspapers.

Findings

Sections ‘The tagged lemma frequency list’ and ‘Collocation’ focus on tagged lemma frequencies and collocation analysis, using a subset of the Katrina corpus, comprising the months August and September 2005. The storm touched down in the last week of August, which makes the data from this month potentially the richest. The subsequent month includes the largest amount of Katrina material. This subcorpus (9.65 million words) hence focuses on the earlier stages of the event and news cycle. Analyses will be synchronic.

In sections 'Key parts-of-speech' and 'Key semantic tags', we perform keyness analysis of POS and semantic tags across a short-term diachrony, to test whether key linguistic devices indicate changed news values in Hurricane Katrina reporting over time. For this section, we compare three subcorpora: August 2005 (608,985 words in 684 texts), September 2005 (9,043,083 words in 10,219 texts) and October 2005 (4,557,231 words in 5360 texts). The August subcorpus contains texts published as the event unfolded, whereas the September subcorpus reflects the immediate aftermath, and the October subcorpus encompasses coverage of the continuing fallout.

The tagged lemma frequency list

Table 2 shows a list of the top 200 most frequent lemmas used as a particular POS, categorised according to perceived potential to construct a specific news value and excluding punctuation (the tags are explained at http://www.natcorp.ox.ac.uk/docs/URG/codes.html#klettpos, accessed 12 June 2014).[2] This categorisation is based on hypothesis – no concordancing or collocation analysis was undertaken at this stage, because of the high raw frequencies of these top 200 (ranging from 4939 to 467,450), unless specifically mentioned below. Where specific examples are provided, they come from the corpus. We have avoided multiple categorisations, but do note occasionally when an item could also be placed in a different category.

As expected, there are many function words that are not clearly identifiable as pointers to a specific news value. This is not to say that they cannot be used to construct newsworthiness, depending on the co-text (Bednarek and Caple, 2014: 147). The table also includes nouns, verbs and adjectives that tell us little in and of themselves about newsworthiness, although in a specific co-text they might construct news values. To give just one example, lemmas like COMPANY, BUSINESS, SCHOOL and CENTER may be used with proper nouns which establish the local (Proximity) or elite (Eliteness) nature of the respective entity.

There is also a considerable amount of high-frequency verbs that do not clearly point to a specific news value, including reporting expressions (e.g. SAY, TELL). As hypothesised by Bednarek and Caple (2014: 147), reporting verbs are nevertheless useful as a starting point, because we can analyse the speakers associated with such verbs in terms of constructed status: ordinary affected people/eye witnesses (Personalisation) versus authorities (Eliteness). Collocation analysis may be helpful here, especially when dealing with a large corpus. Table 3 shows the top 50 collocates, excluding punctuation, for the lemma SAY, which is the most frequent reporting verb at rank 10 with 86,700 instances.

Table 3 suggests that elite sources are often cited, as indicated through the high-status or authority role labels such as *officials*, *analysts* or *executive*, and proper nouns referring to leading politicians (President George *Bush*, New Orleans Mayor Ray *Nagin*, Louisiana Governor Kathleen Babineaux *Blanco*). Concordancing shows that *Maestri* also refers to

Table 2. Pointers to newsworthiness.

Not clearly related to a specific news value: function words	the_ART, to_PREP, and_CONJ, of_PREP, a_ART, in_PREP, for_PREP, that_CONJ, on_PREP, 's_UNC, it_PRON, not_ADV, he_PRON, at_PREP, they_PRON, with_PREP, from_PREP, by_PREP, i_PRON, but_CONJ, his_PRON, this_ADJ, who_PRON, their_PRON, an_ART, or_CONJ, as_CONJ, that_ADJ, she_PRON, up_ADV, you_PRON, as_PREP, its_PRON, which_PRON, what_PRON, about_PREP, my_PRON, out_ADV, if_CONJ, other_ADJ, also_ADV, into_PREP, her_PRON, those_ADJ, no_ART, how_ADV, back_ADV, through_PREP, about_ADV, down_ADV, over_PREP, as_ADV, out_PREP, in_ADV, these_ADJ
Not clearly related to a specific news value: nouns	home_SUBST, school_SUBST, house_SUBST, company_SUBST, center_SUBST, mr._ SUBST, way_SUBST, game_SUBST, business_SUBST, job_SUBST, group_SUBST, thing_SUBST
Not clearly related to a specific news value: verbs	be_VERB, have_VERB, say_VERB, do_VERB, would_VERB, get_VERB, can_VERB, make_VERB, go_VERB, take_VERB, could_VERB, come_VERB, see_VERB, know_VERB, need_VERB, want_VERB, work_VERB, call_VERB, find_VERB, give_VERB, think_VERB, should_VERB, tell_VERB, may_VERB, try_VERB, look_VERB, use_VERB, play_VERB, expect_VERB, ask_VERB
Not clearly related to a specific news value: adjectives	good_ADJ
Potential pointer to Eliteness	official_SUBST, federal_ADJ, Bush_SUBST, president_SUBST, government_SUBST, service_SUBST, team_SUBST, agency_SUBST, red_ADJ
Potential pointer to Personalisation	people_SUBST, family_SUBST, resident_SUBST, child_SUBST
Potential pointer to Proximity	we_PRON, new_SUBST, orleans_SUBST, city_SUBST, state_SUBST, there_PRON, area_SUBST, our_PRON, louisiana_SUBST, gulf_SUBST, st._SUBST, coast_SUBST, national_ADJ, street_SUBST, here_ADV, parish_SUBST, there_ADV, where_CONJ, mississippi_SUBST, local_ADJ, u.s._SUBST, county_SUBST
Potential pointer to Timeliness	will_VERB, year_SUBST, day_SUBST, week_SUBST, after_PREP, time_SUBST, when_CONJ, last_ADJ, now_ADV, still_ADV, before_PREP, sept._SUBST, going_VERB,[a] while_CONJ, begin_VERB, month_SUBST, then_ADV, Monday_SUBST, next_ADJ
Potential pointer to Novelty	new_ADJ, first_ADJ
Potential pointer to Superlativeness	some_ADJ, all_ADJ, than_CONJ, one_ADJ, more_ADJ, many_ADJ, two_ADJ, just_ADV, so_ADV, more_ADV, like_PREP, high_ADJ, million_SUBST, percent_SUBST, three_ADJ, even_ADV, only_ADV, any_ADJ, big_ADJ, much_ADJ, billion_SUBST
Potential pointer to Negativity and Impact	hurricane_SUBST, katrina_SUBST, storm_SUBST, water_SUBST, help_VERB, price_SUBST, disaster_SUBST (also Superlativeness), leave_VERB, emergency_SUBST, relief_SUBST, because_CONJ, oil_SUBST, effort_SUBST, money_SUBST, evacuee_SUBST (also Personalisation), victim_SUBST (also Personalisation), gas_SUBST, move_VERB, damage_SUBST, food_SUBST

[a] Of 5786 occurrences of GOING, *to* is an R1 collocate 5669 times.

Table 3. Collocates of SAY.

Category	Collocates of SAY
Pronouns and general terms of address	*he*, *she*, *Mr*, *they*, *Ms*
High-status role labels	*officials*, *spokesman*, *spokeswoman*, *director*, *chief*, *experts*, *analysts*, *Dr*, *coach*,[a] (*company*), *manager*, (*police*), *executive*, *mayor*, *president*, *Col.*, *Capt*, *Lt*
Proper nouns	*Nagin*, *Blanco*, *Smith*, *Williams*, *Davis*, *Brown*, *John*, *Maestri*, *Mike*, *Bush*, *Robert*, *Jim*, *Michael*
Mode	*statement*, *interview*
Discourse-related	*adding*, *referring*
Other	*it*, *would*, *yesterday*, *that*, *I*

[a]As explained earlier, instances of *coach* (occurring 418 times as collocate of SAY in 286 texts) result from the sports news stories in the corpus and indicate that Eliteness is also a news value in sports news.

two less well-known news actors constructed as authority: Walter *Maestri* (Jefferson Parish Emergency Management Director) and William *Maestri* (superintendent of schools for the Archdiocese of New Orleans). Other collocates are not easily classifiable, either because they are pronouns or general terms of address or because they are common surnames (*Smith*, *Williams*, *Davis*, *Brown*) or first names (*John*, *Mike*, *Robert*, *Jim*, *Michael*), which can be used with both elite and ordinary news actors. Concordancing could be used to establish proportions of 'ordinary' versus 'elite' speakers here. Nevertheless, this very brief collocation analysis confirms the hypothesis that reporting verbs can be a useful starting point for investigating news values.

To return to the broader discussion of Table 2, it may appear surprising that the positive lemma GOOD is so frequent, occurring 7769 times in the corpus. After all, Negativity is more commonly recognised as a news value than Positivity (though see Harcup and O'Neill, 2001: 279 and Schulz, 1982: 152 on positive news); indeed, Negativity has been called 'the basic news value' (Bell, 1991: 156). However, the adjective GOOD (comprising the word forms *good*, *better*, *best*) may not necessarily refer to positive happenings and events around Katrina (collocates of *best* include *way*, *thing*, *interests*, *friends*, *worst*) and may also be negated (e.g. *not/n't* occur 309 times to the left of the word form *good*) or used for comparison (e.g. *than* is the most significant 4L/4R collocate of *better*). Nevertheless, this is an intriguing finding and an example where a frequency list can come up with unexpected results, which can then be explored further.

We now move on to those lemmas that can be classified as potential pointers to a specific news value (based on Table 1). For Eliteness, this includes the proper noun BUSH and his title PRESIDENT, as well as adjectives and nouns pointing to authorities or official services such as OFFICIAL, FEDERAL, RED (as in *Red Cross*), GOVERNMENT, SERVICE, TEAM and AGENCY. For Personalisation, this includes lemmas pointing to ordinary people: PEOPLE, FAMILY, RESIDENT, CHILD – to this we can add EVACUEE and VICTIM (from the Negativity and Impact row). From the perspective of news values analysis, the latter nouns construct ordinary people as negatively affected by the hurricane, simultaneously establishing Personalisation, Negativity and Impact. More detailed analysis of attributes and naming strategies for human referents in the Katrina corpus is provided in Potts (2013).

For Proximity and Timeliness, we have included as ‘pointers’ all potential indicators of place and time, including expressions that may anaphorically, cataphorically or exophorically point to a place (THERE, HERE) or time (THEN), for example,

> We know the devastation is worse **there** [the coast] than it is **here** [Lucedale]. (*Atlanta Journal – Constitution*)

We must emphasise, however, that not all place and time references establish Proximity and Timeliness. The question is whether or not a particular place reference constructs the event as geographically or culturally *near the target audience* and whether or not a particular temporal reference constructs the event as *recent*, *ongoing*, *about to happen* or *seasonal*. Thus, lemmas like AFTER, WHEN, BEFORE, WHILE may more often simply construct temporal relations between happenings, rather than constructing news value, as in this example:

> The water and sewage pumps in the coastal refinery city of Baytown malfunctioned during the hurricane, and residents were asked to conserve water Saturday **while** the city worked on the problem. (*Washington Post*)

A word like THERE can simply be used in existential constructions, too, rather than constructing Proximity:

> The deaths were declared storm-related because **there** was some physical evidence the person made an effort to survive after the storm, Eckert said. (*Times-Picayune*)

For Proximity, pointers further include the first person plural lemmas WE and OUR. When used in a way that *includes the target audience*, these can establish Proximity. However, since there are 38,710 instances of WE and 9941 instances of OUR, extrapolations from small samples might be the only way of establishing the extent of inclusive usage here.

To continue with our discussion of Table 2, pointers to Novelty comprise indications of ‘newness’ (NEW, FIRST), which emphasise the various ways in which aspects of events are new or ‘a first’.[3] This is very common in news discourse in general, but collocation or concordance analysis could tell us more about the specific entities constructed as ‘novel’ in Katrina reporting.

Pointers to Superlativeness include quantification and intensification. Even small numbers (ONE, TWO, THREE) may intensify on occasion (e.g. ***one** billion*), as may the lemmas SOME (e.g. ***some** of the toughest times*) and ANY (e.g. *we don’t have **any** insurance. None. Not a dime*). LIKE can be used in a simile to intensify (e.g. *Home . . . looks **like** a ‘war zone’*; *it looks **like** a bomb went off*), which is why it is included here, but it is multifunctional and not all of its 7800 instances are likely used to intensify. THAN is included in this category because we assume it occurs in intensifying constructions such as *more . . . than*, or ADJ-*er than*. JUST, EVEN, ONLY construct Superlativeness in expressions like *prices tripled* ***in just weeks***; ***even*** *metro Atlanta could feel some effects of the storm*; ***in only four hours*** *Saturday, the Salvation Army raised $10,922*.

Finally, the last row in Table 2 includes pointers to both Negativity and Impact. They are included in the same category here because impact tends to be constructed as negative in disaster reporting. This category includes words which construct cause–effect relations (BECAUSE, LEAVE; e.g. *more than 1 million people could be* ***left*** *stranded*) and hurricane-related labels (HURRICANE, STORM), including its name (KATRINA), since hurricanes are likely to be culturally evaluated as negative. Also included are words that construct the event as a problem (DISASTER, EMERGENCY) and words we hypothesise refer to needing or providing help (HELP, RELIEF, EFFORT) or to the hurricane's negative effects, including on people (DAMAGE, EVACUEE, VICTIM).

Perhaps slightly more problematically, we have also included the lemmas OIL, GAS, FOOD, PRICE, MONEY and MOVE. We have included these here because we hypothesise that they refer to negative effects in terms of prices and availability of oil, gas, food and the movement of people after disasters (non-causative LEAVE may also be used here). This hypothesis is grounded in previous research on the event, as well as an understanding of the corpus and its sociocultural context. Potts (2013: 124) shows that adjectival collocates from the semantic category 'movement, location, travel and transport' co-occur with naming strategies RESIDENT, VICTIM and SURVIVOR in the Katrina corpus. News worthiness is socially negotiated, and so it is critical that any analysis of newsworthiness is buttressed by research into and awareness of the sociocultural context. Although it is impossible to read each text in this corpus in depth, we know that the gulf region is a known industrial hub, so OIL, GAS, PRICE, MONEY are likely to refer to the hurricane's negative impact. Some instances may result from the business news stories in the corpus, as explained earlier.

It is now time to evaluate the technique of using a tagged lemma list to explore news values. Where it appears superior to a pure frequency list is in its power to disambiguate. For example, the use of *new* tagged as substantive shows instances where it is used as part of a location name (*New Orleans* – Proximity), whereas its use as an adjective may construct Novelty. It also allows the identification of high-frequency lemmas such as SAY as well as the frequency of its individual word forms. However, the tagged lemma list does share some of the problems of the frequency list. Word forms and lemmas are multifunctional. Some have a range of different meanings, so that they may construct different news values or no news value depending on their usage and meaning in a given stretch of text. Others construct several news values at the same time (e.g. VICTIM, DISASTER, DAMAGE), making multiple classification a messy necessity. These issues mean that such a list can only be used as a starting point, to formulate a range of hypotheses which need to be confirmed through additional qualitative analysis.

Collocation

As noted earlier, Bednarek and Caple (2014) use concordancing of topic-associated content words to find out what news values are constructed around a given entity. However, when using a large corpus, concordancing may not be viable, and other methods of exposing 'non-obvious' meaning – or 'meaning which might not be readily available to

Table 4. Right-hand collocates of *hurricane*.

Category	R1–R4 collocates of *hurricane*
Proper nouns	*Katrina*, *Rita*, *Andrew*, *Ivan*, *Camille*, *Betsy*, *Gulf*, *Ophelia*, *Charley*, *Georges*, *Hugo*, *Isabel*, *Pam*, *Dennis*, *Florida*, *Frances*, *Miami*, *Mayfield*, *Floyd*
Verbs	*hit*, *struck*, *slammed*, *devastated*, *approached*, *roared*, *ravaged*, *has*, *blew*, *wreaked*, *swept*, *forced*, *passed*, *disrupted*, *bearing*, *caused*, *hits*, *hitting*, *barrelled*, *tore*, *churned*, *ripped*, *destroyed*, *bore*, *blasted*, *killed*, *threatened*, *devastating*, *brought*, *threatens*, *pushed*, *knocked*, *inflicted*
Nouns	*relief*, *victims*, *evacuees*, *efforts*, *fund*, *aftermath*, *center*, *survivors*, *season*, *winds*, *landfall*, *recovery*, *devastation*, *effort*, *damage*, *zone*, *impact*, *path*, *disaster*, *strike*, *flooding*, *havoc*, *strength*, *victim*, *rampage*, *benefit*, *surge*, *protection*, *concert*, *Lemon-Aid*, *fury*, *forecasters*, *activity*
Adjectives	*subsequent*
Other	*ashore*, *1992*, *which*, *'s*, *through*, *1965*, *Aug.*, *May* [*noun/verb*], *1900*, *1969*, *toward*, *29*

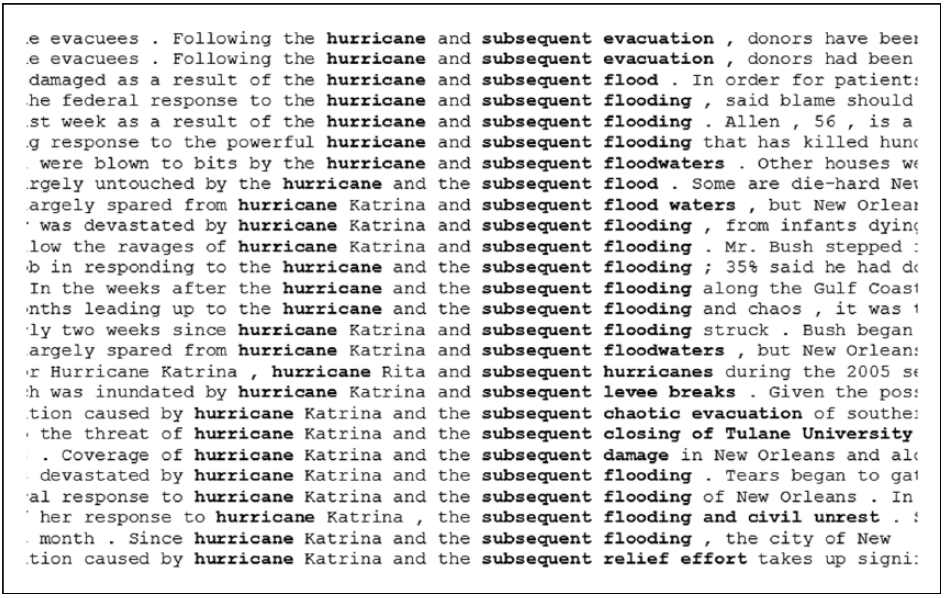
naked-eye perusal' (Partington et al., 2013: 11) – in discourse might be employed to better effect. Here, we explore whether collocation analysis could be used for the same purpose. To do so, we will focus on the top 100 right-hand collocates of *hurricane* (Table 4).

Table 4 shows that the top 100 right-hand collocates of *hurricane* include the names of various hurricanes and other proper nouns, the latter potentially constructing Proximity (*Gulf*, *Florida*, *Miami*). Furthermore, most of the noun collocates establish Negativity, Superlativeness and/or Impact, through reference to providing help (*relief*, *efforts*, *fund*, *recovery*, *effort*, *benefit*, *protection*, *concert*, *Lemon-Aid*), to the negative effects and power of the hurricane (*aftermath*, *winds*, *devastation*, *damage*, *impact*, *disaster*, *flooding*, *havoc*, *strength*, *rampage*, *surge*, *fury*), and to affected people (*victims*, *evacuees*, *survivors*). The latter also constructs Personalisation, as argued previously. Only some nouns appear unrelated to news values, referring to the stages/actions associated with hurricanes: *season*, *landfall*, *zone*, *path*, *strike*, *activity.* Similar to the nouns, most of the verb collocates construct Negativity (e.g. *threatened*), Superlativeness (e.g. *slammed*) or Impact (e.g. *caused*). Often all three news values are established simultaneously: for instance, *devastated* and *ravaged* construct the hurricane as high in negative impact. Only some verbs refer to the stages/actions of hurricanes, for example, *approached*, *passed.*

The adjective collocate *subsequent* primarily co-occurs with word forms referring to the negative impact of the hurricane (*evacuation*, *flood*, *flooding*, *floodwaters*, *levee break*, *closing*, *damage*, *relief efforts*), hence clearly constructing Negativity and Impact (Figure 1).

The year collocates (*1992*, *1965*, *1900*, *1969*) indicate the importance of historical comparisons, which are sometimes used for Superlativeness and/or Novelty, and which may also explain the collocating hurricane names:

> The magnitude of the disaster in St. Bernard has surpassed Hurricane Betsy in **1965.** (*Times-Picayune*)
>
> Katrina would also be the nation's deadliest hurricane since **1900.** (*St. Louis Post-Dispatch*)



```
.e evacuees . Following the hurricane and subsequent evacuation , donors have beer
.e evacuees . Following the hurricane and subsequent evacuation , donors had been
damaged as a result of the hurricane and subsequent flood . In order for patient:
.he federal response to the hurricane and subsequent flooding , said blame should
.st week as a result of the hurricane and subsequent flooding . Allen , 56 , is a
.g response to the powerful hurricane and subsequent flooding that has killed hunc
. were blown to bits by the hurricane and subsequent floodwaters . Other houses we
.rgely untouched by the hurricane and the subsequent flood . Some are die-hard Nev
.argely spared from hurricane Katrina and subsequent flood waters , but New Orlear
' was devastated by hurricane Katrina and subsequent flooding , from infants dyinc
.low the ravages of hurricane Katrina and subsequent flooding . Mr. Bush stepped :
'b in responding to the hurricane and the subsequent flooding ; 35% said he had dc
 In the weeks after the hurricane and the subsequent flooding along the Gulf Coast
'nths leading up to the hurricane and the subsequent flooding and chaos , it was t
'ly two weeks since hurricane Katrina and subsequent flooding struck . Bush began
.argely spared from hurricane Katrina and subsequent floodwaters , but New Orlean:
'r Hurricane Katrina , hurricane Rita and subsequent hurricanes during the 2005 se
:h was inundated by hurricane Katrina and subsequent levee breaks . Given the pos:
.tion caused by hurricane Katrina and the subsequent chaotic evacuation of southe:
' the threat of hurricane Katrina and the subsequent closing of Tulane University
: . Coverage of hurricane Katrina and the subsequent damage in New Orleans and alc
: devastated by hurricane Katrina and the subsequent flooding . Tears began to gat
'al response to hurricane Katrina and the subsequent flooding of New Orleans . In
' her response to hurricane Katrina , the subsequent flooding and civil unrest . :
. month . Since hurricane Katrina and the subsequent flooding , the city of New
.tion caused by hurricane Katrina and the subsequent relief effort takes up signi:
```

Figure 1. Concordances for *hurricane + subsequent.*

Table 4 also indicates that ’s is a collocate, and follow-up analysis can be undertaken to investigate specific frames such as *the hurricane’s* * or *Hurricane Katrina’s* *. For instance, nouns that occur to the right (R1–R4) of *hurricane’s* include *aftermath*, *path*, *impact*, *eye*, *winds*, *victims*, *devastation*, *wake*, *fury*, *wrath*, *effect*, *surge*, *effects*, *damage*, *destruction*. Again, the news values of Negativity, Impact and Superlativeness are clearly visible.

We will conclude this section by briefly evaluating the use of collocation in DNVA. Overall, this method does show some promise, as it can be used with a large corpus and allows identification of those news values that are co-textually associated with a particular word. To explore news values constructed around the entity ‘Katrina’ more fully, we also need to consider the left-hand co-text as well as collocates for additional referential devices (*storm*, *Katrina*). Around the word form *hurricane*, repeatedly constructed news values are Negativity, Superlativeness and Impact. This is an unsurprising result, since we would expect disaster reporting to emphasise an environmental event’s high negative impact on the community. Since collocation analysis uncovers what we would expect it to uncover, rather than any counter-intuitive results, it may in fact be a useful technique for news values analysis: ‘“uncovering the obvious”… gives more credibility to other non-obvious findings’ (Baker et al., 2013: 30). This suggests that collocation analysis can indeed pinpoint repeated constellations of specific news values constructed around entities. However, since it focuses on the *immediate co-text*, rather than the whole news article, it can only provide a partial view of news values constructed around a particular issue. But there is also merit in showing *how* news values are constructed and collocation analysis has the power to identify recurring phraseologies (e.g. the co-occurrence of

Table 5. Key part-of-speech tags in the August 2005 subcorpus.

No.	Part-of-speech	In subcorpus 'Aug 2005'		In reference subcorpus 'Sep/Oct 2005'		Log likelihood
		Frequency (absolute)	*Frequency (per mil.)*	*Frequency (absolute)*	*Frequency (per mil.)*	
1	Past tense of lexical verb	18,923	31,074.04	332,629	24,457.45	+982.9
2	Singular weekday noun	2754	4522.43	36,717	2699.72	+591.79
3	Singular noun of direction	1404	2305.55	16,736	1230.56	+430.21

hurricane with high-intensity verb forms such as *slammed*, *ripped*, *blasted*), figurative devices such as metaphor/personification (e.g. *roar*, *fury*, *wrath*) and rhetorical strategies (e.g. historical comparison). It thus aids the identification of common conventions and clichés of hurricane news reporting.

Key parts-of-speech

Can corpus linguistic techniques help us to understand changes in the construction of newsworthiness over time? To test this, we now discuss a key POStag analysis on three subcorpora; these comprise August 2005, September 2005 and October 2005.

Key POStags from each of these subcorpora were generated using the other two subcorpora as reference. Keyness was calculated using LL with a significance cut-off point of 0.001%.[4] For reasons of scope, only the three items with the highest confidence scores in each category are explored here at greater length. To do so, we identified the specific lexical items making up the greatest proportion of each key POS category, and used collocation analysis to gain a broad understanding of meaning in use, through the lens of common context. When deeper analysis was needed (e.g. to disambiguate instances of polysemous words), we took randomly thinned samples of 50 instances and observed patterns at the concordance level. The first subcorpus to be discussed is August 2005 (Table 5).

The POStag with the highest LL score in August 2005 is 'Past tense of lexical verb', which is somewhat surprising, given that this is the 'breaking news' subcorpus. While the past tense can construct events as recent (the news value of Timeliness), this depends on specific explicit or implied temporal reference (e.g. *yesterday* vs *last year*). A single item – *said* – accounts for over 30% of this result, and occurs with government officials and spokespeople (Eliteness), and, less frequently, 'ordinary' eyewitnesses (Personalisation); top collocates (span ±3, minimum frequency: 5, LL > 15.13) include *he*, *she*, *officials*, *Mr.*, *Blanco*, *Nagin*, *spokesman* (see also section 'The tagged lemma frequency list' earlier). Other, dramatically less frequent, past tense lexical verbs (*lost*, *killed*, *struck*, *died*, *warned*, *caused*, *canceled*, *urged*) indicate Negativity and Impact.

Table 6. Key part-of-speech tags in the September 2005 subcorpus.

No.	Part-of-speech	In subcorpus 'Sept 2005'		In reference subcorpus 'Aug/Oct 2005'		Log likelihood
		Frequency (absolute)	*Frequency (per mil.)*	*Frequency (absolute)*	*Frequency (per mil.)*	
1	Infinitive	208,311	23,035.40	112,748	21,824.18	+219.48
2	Third person plural subjective personal pronoun (they)	35,184	3890.71	17,615	3409.67	+208.11
3	Singular weekday noun	26,310	2909.41	13,161	2547.52	+157.41

'Singular noun[s] of direction' (e.g. *north*, *south*, *east*, *west*) are pointers to Proximity, although top collocates highlight a multitude of contexts: the nouns of direction collocate with proper and common nouns of place (*Florida*, *shore*), and measurement nouns and adverbs (*miles*, *farther*). All of these *may* indicate Proximity, but this relies on the distance between the location of its target audience and the mentioned locations. US locations (such as *Florida*) are culturally/geographically near a US audience (national Proximity), whereas references to local places construct Proximity for the target audience of local newspapers (local Proximity). Nouns of direction are good indicators of local Proximity when capitalised (e.g. '*South* Carrollton Avenue'), indicating proper names of places in a given (known) context.

Finally, instances of 'Singular weekday noun' underscore Timeliness; in 49 out of 50 random concordance lines, these refer to instances in the (very near) past, reporting on the movements, statements, and impacts upon people in the previous week. 'Singular weekday noun' also appears as a key POStag in the September 2005 subcorpus, and is the only repeated key tag across the subcorpora (Table 6). Concordancing shows that these nouns refer to dates that are not recent in relation to the date of publication and hence do not construct Timeliness. Rather, they temporally situate events, such as the hurricane's strike, in relation to the present, and contribute to establishing one of the '"five W's and an H" – who, when, where, what, how, why' – (Bell, 1991: 175) of news reporting.

One key September POStag indicates clear shifts in comparison to August reportage: overuse of the 'Infinitive' in contrast to previous overuse of the past tense lexical verb. The infinitive in English is so flexible – taking its tense from auxiliaries – that it is not a reliable indicator of Timeliness (though it tends to be present-focused and forward-looking in concordance lines). Frequent items in this POS group favour mental/behavioural processes (*see*, *know*, *want*, *think*), including (pro)social activities (e.g. *help*). While the infinitive itself does not point to a news value, these lexical items convey both personal and national Impact. Mental/behavioural processes are individual, and help to convey personal accounts of speakers in their 'own voices', as eyewitnesses, offering Personalisation. Many of these mental/behavioural processes (in the infinitive) are incorporated through inclusion of quotations:

Table 7. Key part-of-speech tags in the October 2005 subcorpus.

No.	Part-of-speech	In subcorpus 'Oct 2005'		In reference subcorpus 'Aug/Sept 2005'		Log likelihood
		Frequency (absolute)	*Frequency (per mil.)*	*Frequency (absolute)*	*Frequency (per mil.)*	
1	Unit of measurement, neutral for number	25,542	5604.72	41,235	4272.15	+1140.54
2	Singular month noun	11,951	2622.43	18,752	1942.80	+641.7
3	Cardinal number, neutral for number	78,559	17,238.32	150,063	15,547.27	+552.62

> 'It's the worst situation you can **think** of', Wooley told state lawmakers last week, when describing his visit. 'It will make you **cry** to see it. It's horrible'. (*Times-Picayune*)

> 'I have two guest rooms in my home that are sitting completely empty', posted one Atlanta resident. 'I would like to **think** that if this were me, someone would **help** and **offer** me a place to at least gather my thoughts for a bit of time'. (*Atlanta Journal – Constitution*)

While quotations from high-status professionals generally construct Eliteness, in the first example above, Louisiana Insurance Commissioner Wooley constructs himself as an eyewitness, with reference to his particular experience and emotions, thus also personalising the event. In the second example, an ordinary citizen is offering help through personally identifying with the affected residents.

What emerges from Table 6 as characteristic of September 2005 compared to the months before and after is overuse of third personal plural personal pronoun (*they*). In only 5 out of 50 sample concordance lines are *they* named individually in the extended context; *they* refers to generalised evacuees, government officials and employees of aid organisations. This indicates a weaker form of Personalisation, as people are incorporated as generalised referents rather than individualised actors who are associated with processes or quotations (see e.g. concordance below). Rather than hearing from individuals themselves about their experiences, we now hear about groups of individuals, some of whom are Elite. This is similar to the distancing and objectification in Hurricane Katrina reporting observed by Potts (2013):

> 'The result is that many of our citizens simply are not getting the help **they** need, especially in New Orleans' he said. (*New York Times*)

We move on now to discussion of the final subcorpus, representing the continuing aftermath of the storm. Compared to August and September, October 2005 features a high amount of measurements and numbers, two of its key POS (unit of measurement, cardinal number) contributing to this preference (Table 7).

Cardinal numbers do not reliably indicate newsworthiness here, as these include dates (of varying distance from the present), counts (from small to great) and radio stations.

However, units of measurement (over 20% of which comprises *percent* itself) construct elements of Impact and Negativity, conveying either financial cost or enumerating more emotional responses:

> The quantity of debris was daunting: Pieces of roofs, trees, signs, awnings, fences, billboards, and pool screens were scattered across several counties. Damage estimates ranged up to **$10 billion**. (*Boston Globe*)

> In a USA TODAY/CNN/Gallup Poll released today, **72%** said they are somewhat or very worried about the effects of the hurricanes on the future of the United States. (*USA Today*)

In the first example, *$10* (unit of measurement) *billion* (numeral noun) in damages are estimated; in the second example, 72% of poll respondents say they are 'worried' about effects (Personalisation, Negativity, Impact), with the high unit of measurement (72%) adding Superlativeness. A total of 40 of the top 50 lexical items contained within 'Unit of measurement' are amounts in US dollars, highlighting the continuing negative impact on financial stability after the storm.

This subcorpus stands apart from the others in one final way: singular month nouns are key in October as opposed to singular weekday nouns, which were key in August and September. Events are no longer conveyed in terms of distance within a week, but in terms of months (i.e. maximally within a year from the present). The most frequent items within this POStag are *Oct.* (18.2%), *Sept.* (12.13%), *September* (10.48%), *Nov.* (8.94%) and *August* (7.38%), which still convey a degree of Timeliness, as they refer backwards and forwards in time only one (or, rarely, two) months into the past or future. It must be noted, though, that Timeliness is usually measured in terms of days rather than months, and one could argue that the use of month nouns in this subcorpus simply situates events in time.

In this section, we tested whether key POStags are useful for investigating the grammatical construction of newsworthiness across a diachrony. We suggested that key POS categories in August, September and October 2005 could be linked to differing emphases upon various aspects of newsworthiness. However, some key POStags are more easily linked to newsworthiness than others. On the whole, key POStags serve as good indicators of differences between various reporting periods, but involve many levels of analysis to uncover the relationships that may (or may not) exist between these occurrences and news values. This includes viewing the constituents of each POStag to gain a better understanding of prominent patterns contributing to keyness as well as collocation and concordance analysis of randomly thinned samples.

Key semantic tags

Using USAS, each corpus item was assigned a tag denoting its correspondence to 1 of 21 major discourse fields and 232 subdivisions. The semantic tags 'show semantic fields which group together word senses that are related by virtue of their being connected at some level of generality with the same mental concept' (Archer et al., 2002: 1). Items sharing semantic tags (e.g. *say* and *claim*) might not be frequent enough in isolation to make it to the top of a frequency list (as in section 'The tagged lemma frequency list'),

Table 8. Key semantic tags in the August 2005 subcorpus.

No.	Semantic tag	In subcorpus 'Aug. 2005'		In reference subcorpus 'Sep/Oct 2005'		Log likelihood
		Frequency (absolute)	*Frequency (per mil.)*	*Frequency (absolute)*	*Frequency (per mil.)*	
1	Weather	5536	9090.83	47,362	3482.42	+3581.31
2	Substances and materials generally: Liquid	3190	5238.40	27,588	2028.48	+2019.79
3	Damaging and destroying	2517	4133.24	29,395	2161.35	+817.99

but in combination, may be a useful pointer for broader meanings. In this section, we consider key semantic tags in August 2005, September 2005 and October 2005 by comparing the corpora, as described above. We consider the categories in the first instance, then describe the most frequent constituents, and finally make use of collocation or concordance line analysis to illuminate patterns.

The most key semantic subcategories in August 2005, as Hurricane Katrina was travelling across the Gulf Coast, directly relate to the event: 'Weather', 'Liquid' and 'Damaging and destroying' all convey the Impact of the storm to varying degrees. Items from the 'Weather' category convey Negativity, and are additionally intensified by their collocates, which construct Superlativeness: Both *hurricane* and *storm* collocate with *devastating*, *major*, *powerful*, *deadly* and *catastrophic*; *winds* collocates with *high*, *strong*, *heavy*, *fierce* and *powerful*; *flooding* is frequently described as *worst*, *severe*, *widespread* and *extensive* (±3 span, LL > 15.13 minimum frequency: 5).

Items tagged 'Damaging and destroying' (e.g. *damage*, *damaged*, *broken*, *broke*, *ripped*, *harm*) describe the immediate aftermath of the storm in terms of negative impact (most prominently), but also in terms of Superlativeness (e.g. with intensified lexical items *devastation*, *destroyed*, *destruction*, *collapsed*, *ravaged*). In an interrelated pattern, words semantically tagged as 'Liquids' describe Impact in terms of direct and indirect consequences: the category contains 42.61% *water* and 5.11% *waters*, contrasting to 29.53% *oil*, 9.09% *gasoline* and 4.29% *crude*. These describe environmental/architectural Impact (flooding, oil spills), but also financial Impact (the rising cost of gasoline). Impact of both of these types is accompanied by Superlativeness; collocates include measurements ('27 feet of *water*') and comparative descriptors ('pushed … *gasoline* prices sharply higher').

In comparison to the key semtags of August 2005 in Table 8, the key semtags of September 2005 (in Table 9) appear at first glance markedly more 'personal'. However, deeper examination uncovers patterns previously explored. The most key semtag in this month ('Evaluation: Good/bad') is made up of 60.35% *disaster*, 13.9% *disasters*, 6.91% *catastrophe* and 1.38% *catastrophes*. Less frequent are adjectives from the same word family (e.g. *catastrophic*, 4.22%; *disastrous*, 0.98%) and the negative evaluative adjective *worst* (1056 occurrences, 10.59% of semtags). This semantic tag thus incorporates both negative lexis and evaluative language (Table 1), but both types of resources construct the

Table 9. Key semantic tags in the September 2005 subcorpus.

No.	Semantic tag	In subcorpus 'Sept. 2005'		In reference subcorpus 'Aug/Oct 2005'		Log likelihood
		Frequency (absolute)	*Frequency (per mil.)*	*Frequency (absolute)*	*Frequency (per mil.)*	
1	Evaluation:- Good/ bad	9976	1103.16	3818	739.04	+468.46
2	Happy/sad: Happy	11,165	1234.65	4508	872.60	+405.21
3	Getting and giving; possession	37,038	4095.73	17,645	3415.47	+403.89

news values of Negativity and Superlativeness: with the word families 'disaster' and 'catastrophe', we are dealing with intensified negative lexis (used in the corpus to refer to environmental happenings or levels of destruction that are socially or scientifically defined, such as '*was declared a federal* ***disaster*** *area*' or '***catastrophic*** *damage/failure*'), whereas *worst* is a negative evaluative adjective in its superlative form.

The next key semtag, 'Happy', seems a counter-intuitive concept in disaster reporting, but closer inspection reveals that 60% of this category comprises a single type: (disaster) *relief*. 'Getting and giving; possession' contributes to the same overall tendency: this category contains tokens such as *supplies* and *donations*, which, together with disaster *relief*, convey Impact, upon both the citizens and organisations providing aid, and the victims in need of it.

Key semtags from October 2005 (Table 10) also show a distinct semantic field preference. Patterns that were uncovered during key POStag analysis also hold strong in the focus on the monetary Impact of Hurricane Katrina: all three of the semantic subcategories are concerned with money and business.

Items from 'Money: Affluence' (e.g. *fund*, *earnings*, *income*, *profit*, *capital*) and 'Business: Generally' (*business*, *businesses*, *economy*, *inc. contracts*, *contract*, *corp*) emphasise this financial Impact, as well as convey Eliteness (of large, powerful corporations and funds). Impact and Eliteness are also conveyed through the items belonging to the 'Business: Selling' semtag (e.g. *sales*, *market*, *buy*, *trade*, *sold*, *sell*, *sale*, *markets*), although Timeliness is also integrated, often through predictions of sales/markets futures (***sales*** *of building products are expected to climb substantially*). Comparison to previous market events and circumstances, such as 11 September 2001, also occur:

> But investors believe that the hurricanes have made third-quarter results somewhat irrelevant in terms of predicting what's to come – not unlike the situation that faced the **market** after the attacks of Sept. 11, 2001. (*St. Louis Post-Dispatch*)

Such a comparison with 9/11 arguably constructs the hurricane's impact both as severe, and as in line with expectations (Impact/Superlativeness/Consonance).

In this section, we devoted some time to exploratory analyses using key semantic tags. In contrast to key POStag analysis, which exposed grammatical features that

Table 10. Key semantic tags in the October 2005 subcorpus.

No.	Semantic tag	In subcorpus 'Oct. 2005'		In reference subcorpus 'Aug/Sept 2005'		Log likelihood
		Frequency (absolute)	*Frequency (per mil.)*	*Frequency (absolute)*	*Frequency (per mil.)*	
1	Money: Affluence	11,224	2462.90	17,396	1802.31	+649.49
2	Business: Generally	8905	1954.04	13,524	1401.15	+579.14
3	Business: Selling	11,484	2519.95	18,244	1890.17	+569.77

'pointed' to areas for deeper corpus-assisted investigation (e.g. through the use of frequency breakdowns, collocation, and concordance analysis), key semtag analysis provided results that were much more easily accessible, either from first glance, or once frequency lists of the constituent lexis were viewed. This is because semantic fields have much more in common cognitively than items sharing a word class. Items from a semantic field may share, for instance, specific negative meanings (as was the case with the negative impact category 'Damaging and destroying'), or attract a set of (grammatically similar) collocates which establish a particular news value (as was the case with intensifying adjectival collocates of *hurricane* and *storm* from the 'Weather' key semtag). For an initial overview of key areas that may contribute to constructing newsworthiness in a given corpus, we recommend key semantic tag analysis over key POStag analysis for transparency and rapidity.

Conclusion

There are a number of reasons why it is challenging to analyse the construction of news values using corpus linguistic techniques.[5] First, there is no closed list of news value devices – news values can be constructed by an open-ended range of lexical or grammatical resources (word forms, lemmas, phrases, whole clauses or sentences). This means that we cannot search the corpus for a defined set of devices. However, as news discourse is conventionalised, we could search for selected devices that are known to be commonly used in news reporting to construct news values (e.g. temporal markers such as *last*, *yesterday*, common emotion nouns such as *fear*, *hope*, *concern* or semantically related adjectives such as *unexpected*, *astonishing*, *shocking*). Additionally, as we argue in this study, collocation analysis can be used to identify news values established in the immediate co-text of topic-associated content words (e.g. *hurricane*).

Second, if we consider newsworthiness as a property of texts rather than individual items, we also need to look at how it is constructed intra-textually. For this, we need to consider the text as a coherent communicative event – reading it horizontally and as a whole – which is not a recognised strength of corpus approaches (Tognini-Bonelli, 2001: 19). But where corpus linguistic methods excel is in the identification of inter-textual patterns in datasets. In so doing, these methods can help to quantify and identify *repeated* sequences, uncovering common practices and conventions in news reporting. As our study has illustrated, corpus techniques can help reveal how phraseologies, figurative devices and rhetorical strategies construct news value. For instance, in the Katrina corpus a cause–effect relation (Impact) is set up by EVENT (e.g. *hurricane*) + (*the*) *subsequent* IMPACT (e.g. *flooding*); Superlativeness is constructed through metaphors and personification (*hurricane* + *roared*; *the hurricane's fury, wrath*), and historical comparison (*comparable in intensity to hurricane Camille of* ***1969*** . . . *only larger*).

Third, the construction of news values is heavily co-text-dependent, because language is multifunctional. In Hunston's (2011) words, 'the meaning of any word cannot be identified reliably if the word is encountered in isolation' (p. 14). Frequency lists and keyness measures, which calculate occurrences without co-text, can be used as input for hypotheses, but concordancing or collocation analysis needs to be undertaken to identify meanings. With a small corpus, this can be more easily achieved; the corpus allows qualitative investigation of all or most instances of a given lemma, POS or constituent of semantic domain. It would be possible, for example, to closely examine all indicators of place and time using automatic tagging with follow-up concordancing. When exploring a large corpus (such as the one used in this study), collocation analysis is helpful, but concordancing is restricted to randomly thinned samples. There is also the possibility of focusing on just one news value, as concordancing of more samples is then more manageable.

What are some of the next steps to be taken? More case studies on different topics and different types of news corpora are needed, and there are at least two additional methods that need to be tested on large datasets for DNVA: n-grams and p-frames, where word forms are not listed in isolation but rather with repeated co-textual patterns.[6] On a more theoretical level, the notion of (de)emphasis needs to be investigated in depth. While one approach would assume that (de)emphasis correlates with frequency of occurrence (repetition), another approach would be positional. From this perspective, news values that are constructed in the *summary* (Van Dijk, 1988), *abstract* (Bell, 1991) or *nucleus* (Feez et al., 2008) – headline and opening paragraph – may be considered as the most emphasised, because this part of the news story arguably comprises the 'most important news element of the story in addition to the choice of angle or "hook", or approach to the subject' (Cotter, 2010: 162). Mahlberg and O'Donnell (2008) and Mahlberg (2009) show that news story structure can be usefully investigated from a corpus perspective.

This study suggests that all tested methods could provide useful insights into the construction of newsworthiness in a large corpus. While POStagged lemmas and POStags require comprehensive follow-up qualitative analysis, semtags appear the most insightful in themselves for providing an overview of newsworthiness, and collocation analysis may be useful for identifying news values established in the co-text of topic-associated words.

Funding

This work was supported in part by the ESRC Centre for Corpus Approaches to Social Science [grant number ES/K002155/1].

Notes

1. In this article we focus on language, although the discursive approach incorporates other semiotic resources (see Caple, 2013).

2. No clear pointers to Consonance were identified in the top 200. This might be because it is rarely individual word forms that establish this news value, with the exception of certain expressions (e.g. *expected*, *familiar*). More research is needed to investigate how Consonance is constructed linguistically.
3. The word *novelty* is ambiguous, indicating 'newness' or 'unusuality'. It would be possible to re-name Novelty as Unexpectedness and to include newness as a sub-category of Timeliness. Expressions such as NOT, JUST, NO, ONLY, BUT, WHILE, EVEN may need to be added as resources for Novelty, as negation and contrast/concession are linked to expectations (Bednarek, 2006: 48–49). On the other hand, these expressions would be frequent in any corpus, are multifunctional, and may not necessarily construct newsworthiness.
4. As the corpus has been tagged using CLAWS7, each instance of punctuation has been tagged as itself (e.g. ;_;). These are not collapsible into meaningful units of investigation, and have been disregarded. Likewise, only unambiguous POStags and semtags have been considered for keyness.
5. Some of these points also apply to other phenomena, such as evaluation (see Hunston, 2011).
6. N-grams are 'recurring sequences of *n* words' (McEnery and Hardie, 2012: 41); with p-frames one slot is variable (e.g. *the * of*).

References

Archer D, Wilson A and Rayson P (2002) Introduction to the USAS category system. Available at: http://ucrel.lancs.ac.uk/usas/usas guide.pdf (accessed 22 July 2014)

Baker P, Gabrielatos C and McEnery T (2013) *Discourse Analysis and Media Attitudes: The Representation of Islam in the British Press*. Cambridge: Cambridge University Press.

Bednarek M (2006) *Evaluation in Media Discourse*. London and New York: Continuum.

Bednarek M and Caple H (2010) Playing with environmental stories in the news: Good or bad practice? *Discourse & Communication* 4(1): 5–31.

Bednarek M and Caple H (2012a) *News Discourse*. London and New York: Continuum.

Bednarek M and Caple H (2012b) 'Value added': Language, image and news value (Special Issue on Journalistic Stance). *Discourse, Context & Media* 1: 103–113.

Bednarek M and Caple H (2014) Why do news values matter? Towards a new methodological framework for analyzing news discourse in critical discourse analysis and beyond. *Discourse & Society* 25(2): 135–158.

Bell A (1991) *The Language of News Media*. Oxford: Blackwell.

Caple H (2013) *Photojournalism: A Social Semiotic Approach*. Basingstoke and New York: Palgrave Macmillan.

Caple H and Bednarek M (2013) *Delving into the discourse: Approaches to news values in journalism studies and beyond*. Working Paper. Oxford: The Reuters Institute for the Study of Journalism, The University of Oxford. Available at: https://reutersinstitute.politics.ox.ac.uk/publications/risj-working-papers.html (accessed 22 July 2014)

Carvalho A (2007) Ideological cultures and media discourses on scientific knowledge: Re-reading news on climate change. *Public Understanding of Science* 16: 223–243.

Cotter C (2010) *News Talk: Investigating the Language of Journalism*. Cambridge: Cambridge University Press.

Cottle S (2009) *Global Crisis Reporting: Journalism in the Global Age*. Maidenhead and New York: McGraw-Hill/Open University Press.

Feez S, Iedema R and White PRR (2008) *Media Literacy*. Surry Hills, NSW, Australia: NSW Adult Migrant Education Service.

Fill A and Mühlhäusler P (eds) (2001) *The Ecolinguistics Reader: Language, Ecology and Environment*. London and New York: Continuum.

Gabrielatos C and Marchi A (2012) Keyness: Appropriate metrics and practical issues. Paper presented at critical approaches to discourse studies 2012, Bologna, 14 September 2012. Available at: http://repository.edgehill.ac.uk/4196/1/Gabrielatos%26Marchi-Keyness-CADS2012.pdf (accessed 31 July 2014)

Galtung J and Ruge MH (1965) The structure of foreign news. *Journal of Peace Research* 2(1): 64–91.

Grundmann R and Krishnamurthy R (2010) The discourse of climate change: A corpus-based approach. *CADAAD* 4(2): 125–146.

Harcup T and O'Neill D (2001) What is news? Galtung and Ruge revisited. *Journalism Studies* 2(2): 261–280.

Hardie A (2012) CQPweb: Combining power, flexibility and usability in a corpus analysis tool. *International Journal of Corpus Linguistics* 17(3): 380–409.

Hunston S (2011) *Corpus Approaches to Evaluation: Phraseology and Evaluative Language*. New York and Oxon: Routledge.

Johnson KA, Sonnett J, Dolan MK, et al. (2010) Interjournalistic discourse about African Americans in television news coverage of Hurricane Katrina. *Discourse & Communication* 4(3): 243–261.

Leech G, Garside R and Bryant M (1994) CLAWS4: The tagging of the British National Corpus. In: *Proceedings of the 15th international conference on computational linguistics (COLING 94)*, Kyoto, Japan, 5–9 August 1994, pp. 622–628. Available at: http://www.sigmod.org/publications/dblp/db/conf/coling/coling1994.html

McEnery T and Hardie A (2012) *Corpus Linguistics: Method, Theory and Practice*. Cambridge: Cambridge University Press.

Mahlberg M (2009) Local textual functions of *move* in newspaper story patterns. In: Römer U and Schulze R (eds) *Exploring the Lexis–Grammar Interface* (Studies in Corpus Linguistics 35). Amsterdam and Philadelphia, PA: John Benjamins, pp. 265–287.

Mahlberg M and O'Donnell M (2008) A fresh view of the structure of hard news stories. In: Neumann S and Steiner E (eds) *Online Proceedings of the 19th European systemic functional linguistics conference and workshop*, Saarbrücken, 23–25 July 2007. Available at: urn:nbn:de:bsz:291-scidok-17005; http://scidok.sulb.uni-saarland.de/volltexte/2008/1700 (accessed 30 July 2014)

Mühlhäusler P (2003) *Language of Environment – Environment of Language*. London: Battlebridge.

Partington A, Duguid A and Taylor C (2013) *Patterns and Meanings in Discourse: Theory and Practice in Corpus-Assisted Discourse Studies (CADS)*. Amsterdam: John Benjamins.

Potts A (2013) *At arm's length: Methods of investigating constructions of the 'other' in American disaster and disease reporting*. PhD Thesis, Lancaster University.

Rayson P, Berridge D and Francis B (2004) Extending the Cochran rule or the comparison of word frequencies between corpora. In: Purnelle G, Fairon C and Dister A (eds) *Le Poids des Mots: Proceedings of the 7th international conference on statistical analysis of textual data (JADT 2004*). Louvain-la-Neuve: Presses Universitaires de Louvain, pp. 926–936.

Schulz WF (1982) News structure and people's awareness of political events. *International Communication Gazette* 30: 139–153.

Tognini-Bonelli E (2001) *Corpus Linguistics at Work*. Amsterdam: John Benjamins.

Tuchman G (1978) *Making News: A Study in the Construction of Reality*. New York: Free Press.

Van Dijk T (1988) *News as Discourse*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Author biographies

Amanda Potts is a Senior Research Associate at the ESRC Centre for Corpus Approaches to Social Science at Lancaster University. Her research interests are in corpus linguistics, (critical) discourse analysis, gender studies, analysis of culture, (new) media discourse, representations of identity, and investigation of discriminatory discourses.

Monika Bednarek is Senior Lecturer in Linguistics at the University of Sydney. Her research on news discourse includes her 2006 monograph on the expression of opinion in British tabloid and broadsheet newspapers (Continuum), and her co-authored book on news discourse with Helen Caple (2012, Bloomsbury). She has also collaborated with Helen on articles on environmental news discourse, verbal-visual play in multimodal news stories and, most recently, on news values.

Helen Caple is Senior Lecturer in Journalism at the University of New South Wales, Australia. Her research interests centre on news photography, text-image relations and the construction of news values in images. She is also exploring more broadly the role of photography in contemporary journalism, including in the online environment. This work has resulted in a monograph with Palgrave Macmillan entitled Photojournalism: A Social Semiotic Approach (2013). She is also the co-author of News Discourse (2012, with Monika Bednarek), which examines the news media from a linguistic and social semiotic perspective.