Analyzing Storytelling in Design Talk using LIWC (Linguistic Inquiry and Word Count)

Lloyd, P.A.; Chandrasegaran, R.S.K.; Akdag Salah, A.A.

Publication date
2022

Document Version
Final published version

Published in
Proceedings of Design Thinking Research Symposium 13 (DTRS 13)

Citation (APA)
Lloyd, P. A., Chandrasegaran, R. S. K., & Akdag Salah, A. A. (2022). Analyzing Storytelling in Design Talk using LIWC: (Linguistic Inquiry and Word Count). In E. Tarazi (Ed.), *Proceedings of Design Thinking Research Symposium 13 (DTRS 13)* (pp. 72-86)

Analysing storytelling in design talk using LIWC (Linguistic Inquiry and Word Count)

Almila Akdag Salah[a], **Senthil Chandrasegaran**[b] and **Peter Lloyd**[b]
[a] Department of Information and Computing Sciences, Utrecht University, a.a.akdag@uu.nl
[b] Faculty of Industrial Design Engineering, TU Delft, {R.S.K.Chandrasegaran, P.A.Lloyd}@tudelft.nl

Abstract. Design thinking concepts such as storytelling, framing, and co-evolution, have been established from close readings of design activity. The increase in easy-to-use computational methodologies provides an opportunity to validate these concepts more widely. Among these concepts, storytelling is already operationalised through various computational approaches. In this paper, we create one corpus of design activity data from the four shared-data DTRS workshops and use Linguistic Inquiry and Word Count (LIWC) in attempting to automatically detect components of stories. However, the conversational nature of the data indicates that further development in methodology is needed. The contribution of the paper lies both in outlining how an automated method for identifying stories could work and showing how the DTRS corpus can be compared with other large datasets outside of the design discipline. This represents a further step on the way to understanding design thinking in conversational contexts.

Keywords: Computational Linguistics, Conversation Analysis, Design Talk, Design Thinking, Storytelling

7 Introduction

Over a period of nearly 30 years the Design Thinking Research Symposium series (Cross, 2018) has conducted four shared data workshops, generating data from design activity in a number of different contexts including: think aloud protocols (Cross, Christiaans & Dorst, 1996), designer-client discussion (McDonnell & Lloyd, 2009), design education (Adams & Siddiqui, 2010), and co-creation (Christensen, Ball & Halskov, 2017). In bringing different research groups and traditions together, the key achievement of these workshops has been to develop a common language with which to describe design activity. A central research thread of these workshops has been to consider the function of different words and the structure of different word forms in discourses of designing and evidencing specific ideas about what constitutes design activity with textual data has formed the basis of this inquiry. The concept of 'framing', for example, as it pertains to both Schon (1983) and Goffman (1974) has been explored by Dorst & Dijkhuis (1995), Glock (2009), Mabogunje et al. (2009), Secules, Gupta & Elby (2015), and Dong & MacDonald (2017). Such extensive exploration and development of a common language to describe designing through analysing common data can equally be applied to other now familiar concepts such as 'co-evolution', 'problem-finding', and 'evaluation'.

One particular concept relating to design activity that has recently found traction is that of 'storytelling'. This is the idea that significant parts of the design process, particularly in naturally occurring datasets, are taken up with narrative structures that place actors, objects and relations in specific contexts. Arguably the early DTRS analyses on episodic memory (Visser, 1995) and precedents (Oxman, 1994) established a foundation for later work on storytelling to emerge (Lloyd,

2000) and develop (Lloyd, 2009). Most recently – in terms of the DTRS – Lloyd & Oak (2018) identified 'value tension' as an important function that stories have in design conversations:
"A story […] provides a dynamic structure through which values can be framed and categorised, and through which design-oriented arguments can be presented." p.109
Concepts like that of storytelling, and the other concepts listed above, have given us a rich language with which to describe and present design activity, both within the DTRS datasets and more widely in the design research field. However, they are (necessarily) derived from the selective use of examples from design protocols. In this respect they represent the first phase of a research process highlighting general features of design activity. A general method of the DTRS workshops when looking at language use has been to 'fit' concepts with textual excerpts, showing how a particular linguistic mechanism might be in play. This is a manual process of close analysis and description, selecting self-defined points of interest in the discourse and describing them in abstract terms.
We might, then legitimately ask: are these concepts robust? Are they features of design discourse that can be readily and reliably identified? Can we show that they are necessary conditions of design activity?
More automated methods of textual analysis have begun to be applied to DTRS data to explore more general features of the design protocols (Menning et al., 2017). This paper extends this recent trend by treating the DTRS data as a single corpus, amenable to computational linguistic analysis. Rather than identifying specific discursive practices within a particular dataset, we seek both to characterise these DTRS data more broadly and begin to validate previously established concepts about designing. We do this particularly in comparison to other types of textual corpuses. With more and more training datasets becoming available and with analytical sophistication increasing rapidly (Archer & Jockers, 2016), it is now possible to use comparative methods to test both specific concepts related to the DTRS data - in the case of this paper that of 'storytelling' - as well as reveal more general regularities.
In this paper we first describe the constitution of the DTRS corpus that we have created. We then go on to describe the Linguistic Inquiry and Word Count (LIWC) methodology, illustrating how it can be applied to the DTRS corpus and compared to other large datasets. The main body of the paper considers the concept of 'storytelling', discussing how it can be constituted and operationalised to reveal narrative and emotional arcs in textual data. We conclude by saying that although we cannot replicate the significant results of others, we do show new methods for automated data analysis of design activity to facilitate cross-corpora comparison.

2 Building a DTRS corpus

The datasets from the four different shared-data workshops of the Design Thinking Research Symposium feature several different disciplines related to design: DTRS2 - industrial design engineering, DTRS7 - architecture and engineering design, DTRS10 - design education and DTRS11 - product design. These datasets encompass a variety of design activities from think aloud protocols (DTRS2), designer-client discussion (DTRS7), design education (DTRS10), and co-creation (DTRS11). Apart from DTRS10, which followed the design process of students and the feedback they received from their lecturers, all other datasets were collected from the meetings of design professionals.
The datasets differ in length and in the number of participants: DTRS2 is the smallest with one 2-hour 'think-aloud' design session with a single designer and another 2-hour session with three designers. The design conversation focuses on a bicycle accessory. DTRS7 has four 2-hour meetings, where two of these feature conversations between an architect and his clients. The other two are between a multidisciplinary design team working on the design of a digital pen. DTRS10 consists of 38 sessions of various lengths, and varying number of participants. As the sessions feature teacher-student

discussions in five disciplines (industrial design, mechanical design, service-learning design, entrepreneurial design, and choreography) a different conversation hierarchy can be observed. DTRS11 contains 20 sessions of varying length with a focus on preparing and discussing the results of co-creation meetings. In total, it has 17 participants. Together, these datasets contain a rich reflection of design activities that occur in diverse stages of design including how designers think, discuss, reflect, evaluate, and talk to their clients.

To build the DTRS corpus from these datasets, we first got the appropriate permissions and anonymized the data by rendering all datasets into a tabulated format with dataset name, session name, speaker, and utterances. We kept filler words such as 'ehm, uhm' etc. as well as descriptions of actions, such as [laughs], to maintain the conversational flow. The resulting corpus has 373,983 words spread across 64 sessions from the 4 datasets. A more detailed description can be found in Lloyd et al. (2021) and in our sister paper for DTRS13, Chandrasegaran et al. (2021). This corpus, though not big enough to be of use as a training dataset for an AI application, is nonetheless a valuable resource to explore hypotheses about the features of design discourse and dialogue.

3 The Linguistic Inquiry and Word Count (LIWC) Dictionary

Linguistic Inquiry and Word Count (LIWC) is a dictionary prepared by Pennebaker et al. (2015) with the purpose of studying the grammatical, linguistic and socio-psychological dimensions of textual data. LIWC provides a 'gold standard' of categories reflecting these dimensions through individual words (Fast, Chen & Bernstein, 2016) and has been used for more than twenty years to analyze a huge range of textual datasets. Over the course of these twenty years, the dictionary has been enhanced to its current format, which now has about 6,400 dictionary words and word stems in 91 categories and sub-categories. These categories cover linguistic (e.g., pronouns, prepositions), other grammar (e.g., quantifiers, common adjectives), and psychological processes (e.g., positive emotion, insight). Table 1 lists some of these categories with word examples.

Table 1. An overview of some LIWC categories with examples. See Pennebaker et al. (2015) for the full list of 91 categories

Main Category	Sub-Category	Examples
Linguistic Dimension	Function Words	
Pronouns	1st per. singular	I, me, mine
	3rd per. plural	they, their, they'd
Impersonal pronouns		it, it's, those
Articles		a, an, the
Prepositions		to, with, above
Negations		no, not, never
Other Grammar		
Common verbs		eat, come, carry
Common adjectives		free, happy, long
Interrogatives		how, when, went
Quantifiers		few, many, much
Psychological Processes		
Affective proc.	Positive emotion	love, nice, sweet

	Negative emotion	hurt, ugly, sad
Social processes.		mate, talk
	Friends	buddy, neighbour
Cognitive processes.	Insight, Tentative, Differentiation	know, think, realise, maybe, perhaps, could, hasn't, but, else
Perceptual processes.	See, Hear, Feel	view, saw, seen, listen, hearing, feels, touch
Drives	Achievement, Power	win, success, better, superior, bully, danger, doubt, chance
Time orientation	Past focus, Present focus, Future focus	ago, did, talked, today, is, now may, will, soon
Relativity	Motion, Space, Time	arrive, bike, go, down, in, thin, end, until, season

In comparison to coding frameworks introduced in the design literature, using LIWC as a coding tool allows us to explore a wider range of linguistic phenomena. Existing design process coding frameworks tend to focus on a particular aspect of design such as different types of reasoning, how the design object is described, or the design stages themselves. In contrast, LIWC identifies more general cognitive, emotional, perceptual and social processes. Since LIWC is a general dictionary used in various disciplines for many different purposes, a LIWC analysis of design discourse can be compared to other corpora, with the consequent potential to create new insights into what constitutes designing talk.

We first use LIWC categories to generate a network structure from text files. The LIWC dictionary assigns words to more than one category. For example, to think is both a verb (belonging to the 'function' category) and part of the 'cognitive process' category at the same time. Using these overlaps, we can build a network where categories are represented by nodes, and words that belong to more than one category, work as links between nodes. We analyse the DTRS corpus we created using this approach, as well as two further datasets for comparison purposes: the Internet Movie Database (IMDB) corpus with 1068 movie scripts (Danescu-Niculescu-Mizil & Lee, 2011) and a Written Corpora with 680,000 blogposts and 500 scientific books (Lloyd et al., 2021). The IMDB corpus gives insight to everyday conversations, and the comparison to DTRS dataset can highlight if there are specifics to designer conversations versus everyday language. Written Corpora in contrast follows rules of writing and constitutes a good example of carefully crafted sentence structure and text flow, which is lacking in conversational datasets.

Figure 1 shows these three datasets' network structures. Nodes represent LIWC subcategories, their sizes indicate frequency of use, links indicate words shared by two categories (i.e., nodes), and proximity of nodes is determined by link strength. Subsequently, if two categories are frequently used together, they will appear closer in the graph, which is useful to visualise conceptual connections. As each dataset is defined by the way they use LIWC categories, the network maps provide us with an overview of both their similarities and differences. For example, the DTRS corpus uses more 'cognitive process' words whereas the IMDB database has more 'biological process' words. 'Cognitive process' contains words related to problem solving, so we might expect to see more of these words in design protocols, whilst the 'biological process' category covers 'body-related' events, which we might expect to find in the action of a film script. The Written corpus has a smaller 'function' category node, which means less use of pronouns, for example, and hence less ambiguity. Beside the node sizes themselves, how the nodes are connected also gives important clues. It is good to stress here that when the nodes are closer together, means that the corresponding categories share words more frequently. For DTRS, this is especially pronounced in the three closely connected categories: 'cognitive process', 'relative', 'verb'. Among these categories, the 'relative' category covers time and space related words. We take the close connection between these nodes as an indication that objects in time and space [relative] are being modelled [verb] through problem solving activity [cognitive process].

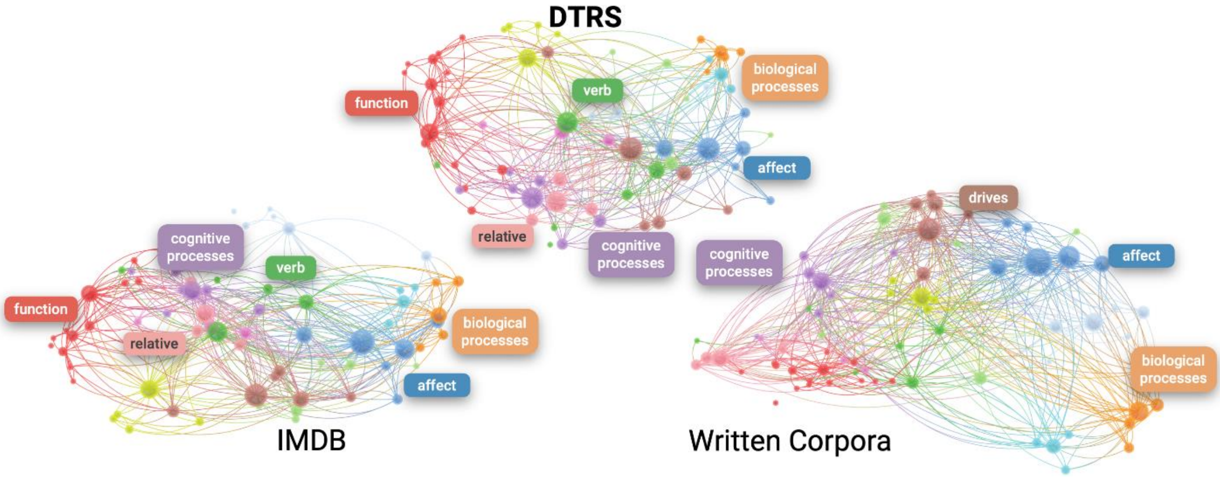

Figure 1. An overview of LIWC Category Network Maps for IMDB data (left), DTRS data (middle) and Written Corpora (right). Each network has 73 LIWC subcategories as nodes and coloured according to the top 12 categories to which they belong. Relative node sizes represent the number of unique words from the dataset that fall into the corresponding LIWC category. Similarly, link thickness between two nodes represents the number of unique words from the dataset that appear in the two corresponding LIWC categories.

In the following sections we explain how LIWC can be used to explore and analyse the concepts of 'story' and 'storytelling' as they appear in textual data. Storytelling is an activity that is, of course, not reserved for designing but is very much a part of everyday life. We tell stories for many reasons during conversations, however, the extent and instances in which we do this has not been quantitatively studied. To look at the idea of storytelling as it applies to the DTRS corpus we need also to look at other datasets with conversations and/or stories for purposes of comparison.

4 Analysing storytelling in conversations

4.1 The structure of stories

What is it that forms the structure of a story? The search to define the basic elements of what constitutes a story stretches back to Aristotle's Poetics and the narrative arc of 'beginning, middle, and end'. In the 19th Century Freytag (1960) argued that narratives consisted of five key components unfolding over time: exposition, rising action, climax, falling action, and resolution. The literature studies as a field offer more theories, but for the purposes of this paper, we focus on the types of narrative distinctions which are applied to quantitative textual data analysis. Boyd et al. (2020) identify three narrative markers – staging, plot progression, and cognitive tension – in several large datasets. van Laer et al. (2019) identify the degree of what they term 'narrativity' in a dataset of TripAdvisor reviews by tagging narrative content (i.e., affective consciousness, cognitive consciousness, spatial and temporal embeddings) and narrative discourse (genres such as drama, tragedy, comedy etc.). Finally, Reagan et al. (2016) classify computationally six types of 'emotional arcs' present in a dataset of 1327 works of fiction to reveal the affective complexity of narratives.

Of the three studies mentioned above, the study by Boyd et al. (2020) draws explicitly on the LIWC dictionary and is the analytical framework we adopt to look at the DTRS dataset. Following Freytag's theory, they divide a story arc into five segments and map these to three story elements: staging, plot progression, and cognitive tension which can be identified in terms of LIWC categories.

Staging: the beginning of a story (i.e., the first segment) is usually reserved for introducing unknowns to the audience. To do so, the writer or speaker needs to use prepositions and articles more than other word categories. Once the staging is done, the number of words from these categories is expected to reduce in the following segments.

Plot progression: after the introduction of characters and the setting, the story moves along by using words from the function category. As an indicator of this, pronouns and auxiliary verbs are expected to rise from the second segment onwards.

Cognitive tension: each story sets up certain challenges and problems that are resolved through the unfolding of events and actions taken by the characters. These points in the story result in the usage of words from the cognitive processes category, with the peak coming in the middle of the story, after which the story moves towards the resolution.

Boyd et al. (2020) established these patterns for three different corpora in their paper resulting in the LIWC dictionary file being updated for this 'story identification' task. As shown in the LIWC network maps of Figure 1, words belong to more than one category. Boyd et al. found that this overlap between categories generates some noise when looking at a narrative, so they published two additional dictionaries to control for the overlaps. These dictionaries are called *Arc of Narrative (AON) with some overlap*, and *AON without overlap*. We used both AON dictionaries as well as the original LIWC dictionary for our analysis. The results between the three were similar, hence, we report the results for the *AON without overlap* as the results were more pronounced for this dictionary. Additionally, Malin et al. (2014) analysed the emotional tone of stories with LIWC, again by segmenting stories first into five equal length and then calculating the positive and negative affect of each segment with LIWC's corresponding categories. They differentiate between redemptive and contaminative stories. Redemptive stories have happy endings, hence a positive emotional rise towards the end, whereas contaminative stories have a sad ending with a negative emotional rise at the end. Since we use the same segmentation in analyzing the narrative arcs, we will also follow Malin et al.'s approach to examine which emotional arcs are prevalent in the DTRS and comparison datasets.

4.2 A LIWC analysis of DTRS stories

For the purposes of this paper, we made the assumption that we would be more likely to find stories in DTRS if we looked at long utterances. We therefore filtered the DTRS dataset first by the length of utterances. As the conversations are between multiple speakers, the majority of the utterances were short. We extracted all utterances with more than 100 words, which returned 530 utterances. Manual inspection of these 530 utterances showed many instances where a description is given about objects, or users, but few full stories with all the story elements of staging, plot progression, and cognitive tension identified by Boyd et al. (2020). With this in mind, the 530 utterances were then filtered by one of the authors using manual annotation to check if the utterance contained story elements. This resulted in 267 utterances. 16 of these were tagged as 'proper' stories, roughly corresponding to the elements outlined above, whereas the majority were deemed as story snippets, i.e., textual pieces that contain one or more story elements, but do not follow the basic principles that are found in every story. The following two excerpts illustrate the difference:

Excerpt 1: Example DTRS story
"But at some point I think you feel like "I'm a human, I'm made to use my hands, I'm made to walk on my feet", and when machines start to do everything I think, at some point we reach a level where you say to yourself, "I want to use my hands, because that's what they're made for, I want to use my feet

because that's what they're made for", and I think they will reach that stage at some point, soon maybe. That's what we feel- at least we feel in Scandinavia, that we want to return to the original human, we want to go out and use our hands, work the old-fashioned way, eat like we did in- in- in one hundred years ago, so."
[Kenny, DTRS11, 127 words]

Excerpt 2: Example DTRS snippet
"This is the funeral that they had for this gentleman. And they may- these are the service sheets that they made up for him obviously showing in a sense his lifetime in pictures really and they had this on the wall itself and it just framed up and down and all the pictures just kept appearing on and off the screen and obviously in our chapel it's so light it's quite difficult to see some of them but this is, I mean this is this is sort of thing now this is started to move forward more now people wanted to do this and I mean in the back it says if you would like to receive a DVD please contact the family. Quite mad in a sense. But that's what you know this is where we are now with this. They're starting to ask for things like this and the concern I've got was that although this man came in to see me he had this erm screen that he had we'd have to bring in a little screen like this, a stand up screen and put it at the front which is again in our chapel quite light, difficult to see and he was then saying about the equipment that you would need. You'd need to have it really in the centre of the chapel sort of looking at the front or doing something and it would be quite difficult to do it in that. We'd have to put it at the side erm and he was then talking about all sorts of expensive equipment that we'd have to be leasing from him and you know all sorts of costs. But I mean what the, the idea is this is starting to become more the norm now. And like I say in Australia they- he saw it in Australia, and he wants to be the main agent in this country. And so obviously it's something that he's seen, and thought was a money spinner for himself I'm not saying that it wouldn't be for us"
[Female A, DTRS07, Crematorium Meeting 02, 347 words]

In the first story excerpt, an idea, or rather a question is introduced as the opening of the story: what makes one a human? The staging is done via the use of the storyteller's emotional reaction to the question, as life has brought them to the point of asking this question "at some point I think you feel like". Then the story unfolds, making the comparison between man and machine, and the resolution comes right after that: in order to keep your humanity, you should use your body, and not leave everything to the machine.

The second story snippet excerpt, although much longer, does not have the five segments and story elements outlined above. It looks like two stories merged into each other. In the first one, a funeral is depicted where technology is used to tell the life story of the deceased person with photographs. There is not a progression of the plot however: we do not learn the details of the deceased person's life. Instead, we hear about the technical problems encountered in the chapel: "in our chapel it's so light it's quite difficult to see some of them". The next part jumps back to the funeral, this time depicting the deceased's friends and family asking for those photographs. Then a new story is introduced: the company that installs the technology, and how much they ask for this service is described. For a human reader, these snippets form the moral of the text: there is demand for such technology, but it is not clear if the storyteller will adopt it. At the end, a resolution does not flow out of these story elements, even for a human observer. The computational challenge is to try and capture these story elements automatically outside of the expected narrative arcs.

4.3 An example analysis

To demonstrate the Arc of Narrative LIWC dictionary in action, the following DTRS story is color-coded, with the words related to staging coloured in orange, the words that are expected to be used for cognitive tension colored in blue, and the words for plot progression colored in green:

"Things will be very different I think, cause like the current generation that we are talking to, are people who have experienced like the- still have a little bit, right? I had an ex-colleague who was twenty-eight maybe and his dad is a lawyer, and he told me that when he grew up he still lived in the alleyways with no bathroom, so that everybody had a common bathroom, and now, you know, completely different, right?, and twenty eight is not that old, it's like our generation, so once you hit ten years from now, the people who already grew up with the bathroom in the house, they will have a completely different perception of this."
[Rose, DTRS 11, Recap with consultants]

Looking at the words highlighted in orange, we see that there is no extra staging in the first sentence, unlike the expectations. The use of article and prepositions is not higher than the remainder of the excerpt. On the contrary, towards the end, we see that articles and prepositions are used more.
Turning to the words highlighted in blue, we see that there are only three words related to 'cognitive process'. Rose's story does not ask the listener to engage in cognitive processes. She does not pose a problem or challenge that needs to be resolved throughout the story. Rose rather asks her audience to remember a time when people did not have toilets in their homes.
Lastly, if we look at words related to plot progression highlighted in green, we see that there are many of them distributed throughout the story. Indeed, there are so many that it would be better to use a quantitative approach to see the changing word use throughout the story. To do that LIWC offers diverse options such as giving the percentage of words appearing from a given category (like function-words) in a sentence, or in a paragraph, or the whole document.

4.4 Comparison datasets

To generate a baseline for datasets that contain short stories we made use of three existing comparison datasets. The reason for this was twofold: First, we would like to test if Boyd et al.'s (2021) hypothesis that certain LIWC categories capture certain parts of a narrative arc would work in conversational datasets, not just in single person monologues (Boyd et al used TED talks and Court-room orders as data). So, our first question is: How does the conversation dimension affect how stories are told?
Second, since LIWC counts words given in a specific category, words used specifically during a conversation, such as filler words, does not constitute a problem. But since we extract only one utterance from a conversational flow, the different narrative markers of a story (staging, plot progression, cognitive tension) may occur over multiple utterances and hence be missed. So, our second question is: What is the effect of the length of a story?
Boyd et al. (2021) cap the minimum length of a story at 250 words, but conversations are dynamic, with turn-taking between participants, and utterances with more than 250 words rather rare. Still, a short story can be easily sketched with less than 250 words. If a story is told in less than 100 words, can LIWC still discern a change throughout the 5 story segments in relation to the 3 narrative elements?
Our first comparison dataset, known as the ROC stories, is a written corpus containing 50,000 stories, each five sentences long, and generated using Amazon Mechanical Turk as part of a study looking at automatic story closure by Mostafazadeh et al. (2016) at ROChester university (hence ROC). The

sentences in these stories are brief to the point that a story has usually word counts around 60-80. The ROC corpus has two important features: it imitates the causal and temporal common-sense relations that are found in daily events, coming close to mini stories that are told in everyday life. Even though these stories contain no conversational phrasing, due to their briefness they are expected to be similar to stories told within the flow of a conversation.

Our second dataset is the IMDB corpus, mentioned earlier in the paper, which contains 1068 movie scripts (Danescu-Niculescu-Mizil and Lee, (2011). The scripts are formatted to show the movie title, speaker, and the utterance. For the purposes of this research no further formatting is needed.

Our third dataset contains 20 years of radio interviews on the US National Public Radio (NPR) network and contains the transcriptions of over 10,000 hours of audio interviews (Zhu et al., 2021). The transcripts are formatted with the program name, session, and the speaker along with the utterance. We assume that, even though both the NPR and IMDB datasets are scripted in some form, they are a close approximation to everyday conversation.

We follow the same pre-processing steps that we applied to the combined DTRS dataset. For both the IMDB and the NPR datasets, we filter for utterances which were above a certain word count (WC). As these datasets are considerably larger than the DTRS datasets, we set the threshold to utterances with more than 140 words. After this filtering, the NPR dataset yields 8048 utterances, and IMDB 145 utterances. The results of filtering reflect the differences in the average utterance length of these datasets, and of course the differences between the different types of conversation flows. Before filtering NPR consists of 105258 utterances, so the dataset contains lengthy utterances. In comparison, IMDB dataset has almost three times more utterances with 304713, but most of the utterances are short, and the 140 WC threshold yields only 145 results. The utterances in the ROC dataset are more-or-less identical in length so we did not do any further filtering and randomly picked 10000 stories for our analysis.

To show the similarities and differences between the stories in our three comparison datasets, we show some an example from each one of them. The five sentence length ROC stories have the expected story-arc elements, i.e., a beginning, a middle and an end:

An example ROC story:

"Jennifer has a big exam tomorrow. She got so stressed, she pulled an all-nighter. She went into class the next day, weary as can be. Her teacher stated that the test is postponed for next week. Jennifer felt bittersweet about it."

In comparison, the IMDB dataset contains stories that are more diffuse: sometimes the beginning or end is not so clear as the conversation takes a different turn, or the story is packed with filler words asking for participation from other speakers:

An example IMDB story:

"There was this kid that I grew up with; he was a couple years younger than me, and sort of looked up to me, you know. We did our first work together, worked our way out of the street. Things were good and we made the most of it. During prohibition, we ran molasses up to Canada and made a fortune; your father too. I guess as much as anyone, I loved him and trusted him. Later on, he had an idea to make a city out of a desert stop-over for G.I.'s on the way to the West Coast. That kid's name was Moe Greene, and the city he invented was Las Vegas. This was a great man; a man with vision and guts; and there isn't even a plaque or a signpost or a statue of him in that town. Someone put a bullet through his eye; no one knows who gave the order. When I heard about it I wasn't angry. I knew Moe; I knew he was headstrong, and talking loud, and saying stupid things. So when he turned up dead, I let it go, and said to myself: this is the business we've chosen. I never asked, who gave the go ahead because it had nothing to do with business."

NPR stories are similar to those of IMDB, especially when the speaker that tells the story is not a radio host, but they are more natural in terms of everyday conversation, with false starts, casual reflections, and repetition:

An example NPR story:
"I walked in and then I looked around. And I realized I was really out of place. These people were at an audition. I mean, the girls had make up. The guys had these sick outfits on. And I look like, pretty much, an athlete that just left the gym, un-showered - oh well. The music comes on and let me tell you guys. I was freaking living in these dance moves. I was like, I don't care what happens. They're going to remember me, even though I don't even live here. So there was like 200 people at this audition. At the end, there ended up being seven guys and five girls. And they put us on video and they were like, you know, if we want you, we'll give you a call in a couple weeks. And I literally left there like, whatever, this doesn't matter. I don't even really care. I went back home to New Jersey. Two weeks later, I was at the laundromat with my pocket full of quarters, when I got the phone call that the agency wanted me to move to LA to pursue a career in dance. So, I get there, I was dancing. I did some TV shows. I did some musicals. I was teaching a dance class and my dance class at the gym was really popular. So my big break came when a friend called me and said, hey, there's this company named Beach Body that wants you to see if you can develop a project with them. So, I go into the Beach Body office. I ended up being there for two hours and that day I left with my contract for Hip Hop Abs."

As with the DTRS dataset, we used the LIWC AON without overlap dictionary to analyse the three comparison datasets. For the ROC dataset, each story is already clearly defined so can be used as a baseline for the other datasets. The ROC story dataset does not have any conversational noise and, even though short, contains the basic elements of staging, plot progression, and cognitive tension as well as having emotional valency. For all datasets AON automatically calculates articles and prepositions separately and reports them as "staging". These words are expected to be used more in the first segment of a story. The rest of the function words (i.e., pronouns, auxiliary verbs, negations, conjunctions, and nonreferential adverbs) are calculated separately as well. These relate to plot progression and are expected to be higher in the middle of a story. Cognitive words are expected to follow a similar curve to plot progression and are the last category that is extracted using AON.
We remarked earlier that the stories told during conversations in the DTRS dataset do not necessarily neatly contain all elements that are used in defining what makes a story; these we define as 'snippets'. One simple difference is in the type of words used: during conversation words that acknowledge the audience, or the interruptions that come from the other participants of the conversation, are common. The stories are often short with sometimes just a short story snippet enough to make the other participants take on the remainder of the story. A simple description of a character, a place or an object can be enough. This raises the question: Will AON categories capture all or some part of the conversational stories in the DTRS dataset?

5 Results

5.1 The narrative arc of conversational stories

For all four datasets the LIWC tool splits each story/utterance into five equal-sized segments. For each segment, the sum of all words relating to the narrative markers of 'staging', 'plot progression' and 'cognitive tension' are calculated and divided by the number of total words for that segment, giving a normalised percentage figure for each narrative marker. A mean percentage is then calculated across all segments for each narrative marker. The results for each segment are then expressed as the standard deviation from this mean. This means that even if stories do have differences in length or have a higher percentage of word use in a certain category, they still remain comparable. The mean of all standard deviations for a particular segment, across all stories in the dataset, forms the final number on the y-axis shown in Figure 2. Table 2 shows an example of how this works for one story only.

Table 2. An example calculation of LIWC results for one story/utterance. Each segment is normalized by calculating the standard deviation from the mean of all segments for each narrative marker.

	Segment 1	Segment 2	Segment 3	Segment 4	Segment 5	Mean %
Staging (%)	12.50	12.50	20.83	4.17	38.10	17.62
Normalized Staging (Standard Deviation from mean percentage figure)	-0.40	-0.40	0.25	-1.04	1.59	-

As noted previously, the DTRS, NPR and IMDB all contains various conversational filler words. The DTRS stories have the highest number, with utterances almost always starting with a turn-taking event acknowledging the previous utterance. As we have not cleaned the data for such conversational effects, we therefore expect to see certain differences from the '5 segment' hypotheses relating to Boyd et al. (2020). Since, even though such filler words are not counted, they do take up space. If the first two sentences in an utterance are only conversational such as: "Well, yeah. I don't know, maybe… But if you think… Let me put it this way…", the first segment of the utterance will, then, have no staging words, which will appear later. Looking at the results for DTRS (top, left in Figure 2) for example, we see that staging starts in the second segment, rather than the first. Cognitive tension is highest in the first and lowest in the last segment, as the start and end of an utterance relates more to a change in speaker than to the story narrated within the utterance. We should also emphasize the confounding factor that the DTRS data shown in Figure 2 is for both story snippets and stories.

IMDB stories do also contain filler words, but not necessarily in the same way as the DTRS data. In contrast to DTRS, IMDB stories are often presented through monologue, and hence the narrative processes can appear similar to classical stories. Lastly, NPR stories, especially if coming from program hosts, follow a written narrative arc the best, though the 8000 utterances do contain a wide variety. ROC stories should in theory all have the same style, all being created from the same rule of being 5 sentences long and should mimic a classical narrative arc. So, we expect to see the most fluctuations from the norm in the DTRS dataset, followed by IMDB, NPR and ROC.

Figure 2 shows that among all datasets NPR (bottom, right) comes closest to the expected narrative arc. After the first part of the story, words related to staging are used less and less, and the words for plot progression and cognitive tension are used more and more. The cognitive tension words are not used as expected, i.e., in the middle, and in general their usage is quite low.

ROC stories do follow the narrative arc for the plot progression and cognitive tension narrative processes, but not so much for staging which stays relatively constant. For both the IMDB and DTRS datasets we can say that although the 'staging' narrative process does take place, the 'plot progression' and 'cognitive tension' narrative processes do not follow any meaningful arc.

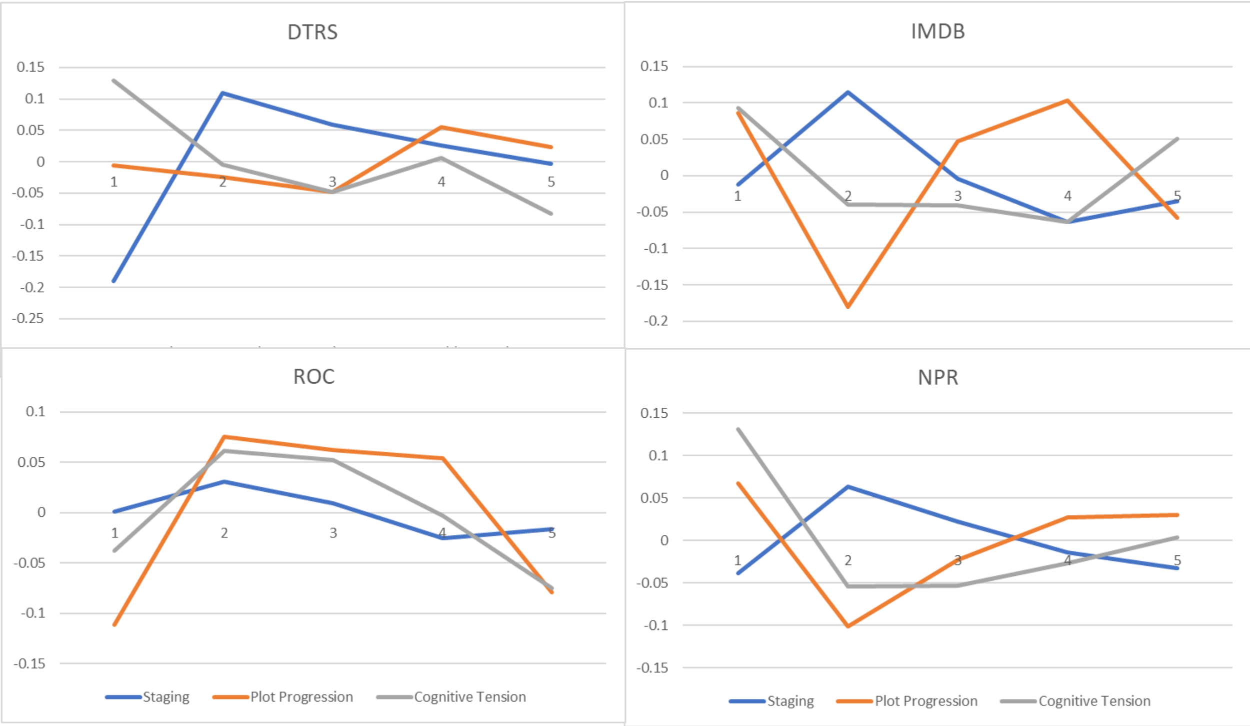

Figure 2. Staging, Plot Progression and Cognitive Tension for stories in the DTRS, IMDB, ROC and NPR Datasets. The x-axis shows the segments, the Y-axis shows the mean standard deviation for each segment across all dataset stories, (narrative markers for each segment for a single story are calculated as the standard deviation from the mean percentage value of one narrative marker across all segments.)

Here we need to note that narrative arc processes for each dataset are mapped within an exceedingly small margin, and do not reflect a real consensus in the data. For example, for DTRS (top, left in Figure 2), the staging starts low, picks up in the first segment and drops steadily till the last segment. However, this latter movement is only due to a change of 0,08 standard deviations, much less than the changes reported in Boyd et al. (2020). If we plot each story in the DTRS dataset separately, we see many different shaped arcs, without a clear overlap. In this respect, Figure 2 shows a clearer picture than the reality.

5.2 The emotional arc of conversational stories

Even though DTRS, IMDB and NPR are conversational datasets, their emotional tone is likely to be quite different. Of the three datasets, DTRS contains conversations recorded at various stages of the design process, and the stories (or story snippets) are told largely to further those design processes. As observed by Lloyd and Oak (2017) and described in the beginning sections of our paper, stories can be used to frame or negotiate "value tension" within the design process. By looking at the emotional arc of conversational stories, we explore the question of whether or not this putative 'value tension' brings with it positive or negative emotional content in the terms discussed earlier of Malin et al. (2014).
In contrast to DTRS stories, IMDB and NPR stories are told with a variety of purposes and to a far wider audience. IMDB, mimicking everyday life, has stories that enhance character, motive, and the overall plot of the film. NPR on the other hand is a radio program reporting newsworthy events along with everyday happenings. The appeal to a wide and diverse audience means that the use of emotional toning may be more pronounced in these datasets. The ROC stories are the shortest stories so their emotional arcs might be difficult to capture using this technique.

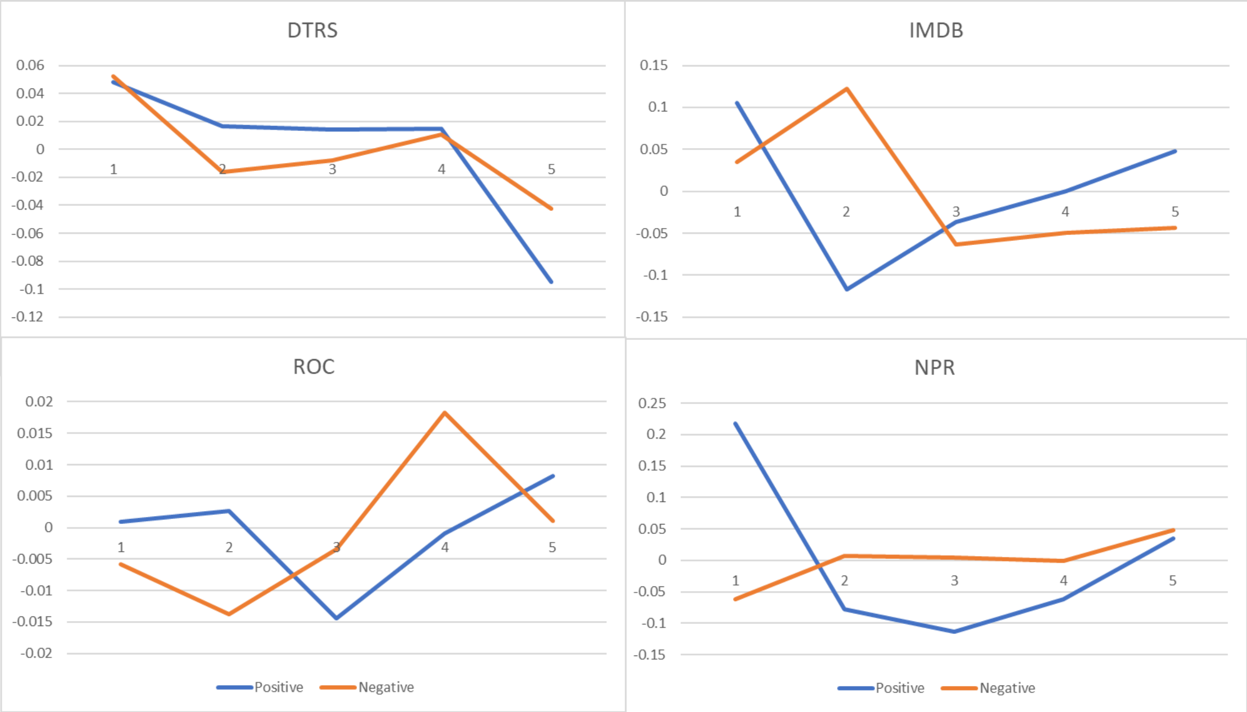

Figure 3. Positive and Negative Emotional Arcs for the DTRS, IMDB, ROC and NPR Datasets. The x-axis shows the segments, the Y-axis shows the mean standard deviation for each segment across all dataset stories, (narrative markers for each segment for a single story are calculated as the standard deviation from the mean percentage value of one narrative marker across all segments.)

Figure 3 shows the results of looking at emotional valence through the five segments of the stories identified in each dataset. For the DTRS stories (top, left) there is a regressive emotional arc, i.e., both for positive and negative emotions there is a fall throughout the stories and snippets. Conforming more to expectations, the IMDB stories (top, right) have a rise-fall shape emotional arc for negative emotions, with the opposite fall-rise shape for the positive emotional arc. This might indicate that in general the IMDB stories have happy endings.

With the NPR stories (bottom, right), the negative emotional arc is steadily rising but more or less stable, whereas the positive emotional arc demonstrates a fall-rise U-shape which can define a more dramatic affective trajectory, but again, largely ending happily. ROC stories do not show a clearly defined trajectory, though it is notable that both positive and negative emotional tone rises significantly during the fourth segment (sentence) of these stories.

For both Figures 2 & 3, although we can discern narrative shapes for the four datasets, they are much less significant than those reported by Boyd et. al. (2020). This may be due to their conversational nature (DTRS, IMDB, NPR) or length (ROC) making the variation of results very high and leading to generally low mean standard deviation rates. Boyd et. al. (2020) report differences of 2 standard deviations, while the biggest difference we found was 0.35 for narrative processes (DTRS) and as low as 0.03 for the emotional arcs (ROC).

6 Discussion and conclusion

Our overall aim in this paper was to create a corpus of the DTRS shared datasets and explore this in relation to well-established concepts relating to design activity. Previous DTRS studies have largely been within a set of data, but building a corpus allows us to compare design activity with other types of activity to reveal more general regularities. We showed how design activity differs from other datasets by applying the LIWC dictionary to find that in general design activity demonstrates a closer relationship between time, space, cognitive process (problem solving) and verb (activity).

We went on to explore (and attempt to validate) one aspect of design discourse (and DTRS data) identified in previous studies, that of ‘storytelling’, to see what results a corpus level analysis might yield. We drew on very recent work using LIWC to look at stories, and again used comparisons with different corpora to explore the shape of narrative and emotional arcs in conversational datasets (as well as in DTRS). Even though previous studies indicate that systematic/automatic ways of tagging stories can yield significant results, our study did not manage to replicate these results in conversational data. We discussed why the need to maintain turn-taking in conversation may interfere with clearly defined stories. Story length also proved to be problematic, with very short stories not having enough linguistic markers to reflect the same scale of the narrative and emotional arcs captured for longer stories.

Before concluding that it is not possible to capture the stories that previous studies of design activity have identified, we must look at the methodology we used to recognise and calculate narrative markers. Though consistent with the literature, fixing utterance length and having equal segmentation may not be suitable for conversational datasets, particularly smaller ones, where stories are likely to be told over multiple utterances. As we noted, this tends to result in story snippets, rather than entire stories. In future studies we will explore the possibility of more dynamically identifying and extracting story shapes of narrative markers independent of fixed utterance and segment size.

LIWC has advantages that are hard to beat. LIWC as a tool is very easy to use, requiring no technical knowledge. The categories of LIWC need to be understood at the conceptual level: for example, what does it mean if someone uses more first-person pronouns? Or what happens linguistically when someone is describing an object? By studying example datasets, one can get acquainted with how these categories are distributed in different type of speech, writing, or even by individual speakers. Combining the use of categories as done by van Laer et al. (2019) enables one to generate complex models to analyse text by simply counting words.

What we hope to have shown in this paper is a new way to think about, and analyse, design protocol data on a larger scale than previously. As more and more datasets become available, the techniques we have used in this paper will become more common in determining how different kinds of design activities are constituted in text, and how they compare with other types of textual data. Our longer-term goal is to create a design dictionary as part of LIWC, to identify designerly ways of talking in large datasets.

In conclusion, although using a LIWC analysis to explore the DTRS data shows considerable promise, we found it difficult to validate the existing concept of storytelling (Lloyd & Oak 2018). Whether this would equally apply to concepts like ‘framing’ and ‘co-evolution’ is an open question, but the conversational nature of the data is likely to be a confounding factor, so methodology needs development. The question remains open as to what type of stories designers typically tell, and how they go about telling them. Despite this, we have shown how an automated method for identifying stories could work. Constructing a corpus of the DTRS shared data and comparing it with other large datasets outside of the design discipline is a further step on the way to understanding design thinking in conversational contexts.

References

Adams, R., Siddiqui, J. (eds.) (2016). *Analyzing design review conversations*, Purdue University Press.

Archer, J., & Jockers, M. (2016). *The bestseller code: Anatomy of the blockbuster novel*, St. Martin's Press.

Boyd, R. L., Blackburn, K. G., & Pennebaker, J. W. (2020). The narrative arc: Revealing core narrative structures through text analysis. *Science Advances*, 6(32), eaba2196.

Chandrasegaran, S., Akdag-Salah, A., & Lloyd, P. (2021). Exploring Distant Reading of Design Discourse:

Examining Expressions of Tentativeness and Causation in DTRS Datasets. In *Proceedings of Design Thinking Research Symposium 13*, in submission.
Christensen, B., Ball, L., & Halskov, K. (eds.) (2017). *Analysing design thinking: Studies of cross-cultural co-creation*, Taylor & Francis.
Cross, N. (2018). A brief history of the Design Thinking Research Symposium series, *Design Studies*, 57 pp 160–164.
Cross, N., Dorst, K., & Christiaans, H. (eds.) (1996). *Analysing Design Activity*, Wiley.
Danescu-Niculescu-Mizil, C., & Lee, L. (2011). Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs, *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, ACL 2011.
Dong, A. & Macdonald, E. (2017). From observations to insights: The hilly road to value creation, in Christensen, B., Ball, L. & Halskov, K. (eds.) *Analysing design thinking: Studies of cross-cultural co-creation*, Taylor & Francis, 465-482.
Dorst, K., & Dijkhuis, J. (1995). Comparing paradigms for describing design activity, *Design Studies*, 16, 261-274, https://doi.org/10.1016/0142-694X(94)00012-3.
Downing, F. (2003). Transcending memory: remembrance and the design of place. *Design Studies*, *24*, 213-235.
Fast, E., Chen, B., & Bernstein, M. S. (2016). Empath: Understanding topic signals in large-scale text, *Proceedings of the 2016 CHI conference on human factors in computing systems*, 4647-4657.
Freytag, G., & MacEwan E. J. (1960). Technique of the drama: An exposition of dramatic composition and art; *An Authorized Translation from the 6th German Edition*, Scholarly Press.
Goffman, E. (1974). *Frame analysis: An essay on the organization of experience*. Harvard University Press.
Goldschmidt, G. (2014). *Linkography: unfolding the design process*. MIT Press.
Glock, F. (2009). Aspects of language use in design conversation, *CoDesign*, 5(1), 5–19.
Lloyd, P. (2000). Storytelling and the Development of Discourse in the Engineering Design Process, *Design Studies*, 21, 357-373.
Lloyd, P. (2009). Ethical Imagination and Design, *Design Studies*, 30, 154-168.
Lloyd, P. & Oak, A. (2018). Cracking open co-creation: Categories, stories, and value tension in a collaborative design process, *Design Studies*, 57, 93-111.
Lloyd, P., Akdag-Salah, A., & Chandrasegaran, S. (2021). How designers talk: Constructing and analysing a design thinking data corpus, *Proceedings of the ASME International Design Engineering Technical Conferences & Computers and Information in Engineering Conference* (accepted, in press).
Mabogunje, A., Eris, O., Sonalkar, N., Jung, M., & Leifer, L. (2009). Spider webbing: A paradigm for engineering design conversations during concept generation. In McDonnell, J. & Lloyd, P. (eds.) *About: Designing – Analysing Design Meetings, Taylor & Francis*, 49-65.
Malin, J. J., Vine, V. J., Stanton, A., Cannava, K., Bodie, G., & Pennebaker, J. W. (2014). The arc of narrative: Using language markers to identify stories. In Poster presented at the annual meeting of the *Society for Personality and Social Psychology*, Austin, TX.
McDonnell, J., & Lloyd, P. (2009). *About: Designing – Analysing design meetings*. Taylor & Francis.
Menning, A., Grasnick, B., Ewald, B., Dobrigkeit, F., & Nicolai, C. (2017). Combining Computational and Human Analysis to Study Low Coherence. In Christensen, B., Ball, L., Halskov, K. (eds.) *Design Conversation* In *Analysing Design Thinking - Studies of Cross-Cultural Co-Creation*, CRC Press, pp.291-309.
Mostafazadeh, N., Chambers, N., He, X., Parikh, D., Batra, D., Vanderwende, L., Kohli, P., & Allen, J. (2016). A corpus and cloze evaluation for deeper understanding of commonsense stories, *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 839–849.
Oxman, R. E. (1994). Precedents in design: a computational model for the organization of precedent knowledge, *Design studies*, 15(2), 141-157.
Pennebaker, J. W., Booth, R. J., Boyd, R. L., and Francis, M. E. (2015). Linguistic inquiry and word count: LIWC 2015. Pennebaker Conglomerates.
Reagan, A. J., Mitchell, L., Kiley, D., Danforth, C. M., & Dodds, P. S. (2016). The emotional arcs of stories are dominated by six basic shapes, *EPJ Data Science*, *5*(1), 1-12.
Schön, D. A. (1984). Problems, frames and perspectives on designing, *Design Studies*, 5(3), 132–136.
Secules, S., Gupta, A., & Elby, A. (2016). Piecemeal versus integrated framing of design activities in Adams, R., Siddiqui, J. (eds.) *Analyzing Design Review Conversations*, Purdue University Press, pp. 155-176.
Van Laer, T., Edson Escalas, J., Ludwig, S., & Van Den Hende, E. (2019). What happens in Vegas stays on TripAdvisor? A theory and technique to understand narrativity in consumer reviews. *Journal of Consumer Research*, 46, 267-285.

Visser, W. (1995). Use of episodic knowledge and information in design problem solving. *Design Studies*,