INJECT: Algorithms to Discover Creative Angles on News

Konstantinos Zachos
Cass Business School
City, University of London
London, U.K.
K.Zachos@city.ac.uk

Neil Maiden
Cass Business School
City, University of London
London, U.K.
N.A.M.Maiden@city.ac.uk

ABSTRACT

INJECT is a new digital tool to support journalists to think more creatively when discovering new angles on stories under development. It delivers interactive and intelligent support embedded in the text editors that journalists work with regularly. This support is generated by combining complex creative searches of millions of related news stories published in multiple languages with entity extraction algorithms and interactive creative guidance tailored to news. This paper reports the tool's architecture, some its algorithms, and the design decisions made to deliver a reliable and usable tool for journalists in different newsrooms and work contexts.

CCS CONCEPTS

• Human-Centred Computing

KEYWORDS

Digital creativity support, creative search, interactive guidance

ACM Reference format:

FirstName Surname, FirstName Surname and FirstName Surname. 2018. Insert Your Title Here: Insert Subtitle Here. In *Proceedings of ACM Woodstock conference (WOODSTOCK'18). ACM, New York, NY, USA, 2 pages.* https://doi.org/10.1145/1234567890

1 Digital Creativity Support for Journalists

Journalism involves the search for and critical analysis of information [1]. How journalists discover and select sources of this information is important, to avoid bias, to be credible and trusted, and to create angles with which to generate new stories of value to readers. Journalist creative thinking, to discover and generate new associations during this searching, can contribute to the generation of stories with new angles. Indeed, journalists are known to seek opportunities to develop new creative skills with which to discover information [4]. However, discovering and examining information sources about complex stories takes time – time that journalists increasingly lack as news organizations reduce staff numbers [6].

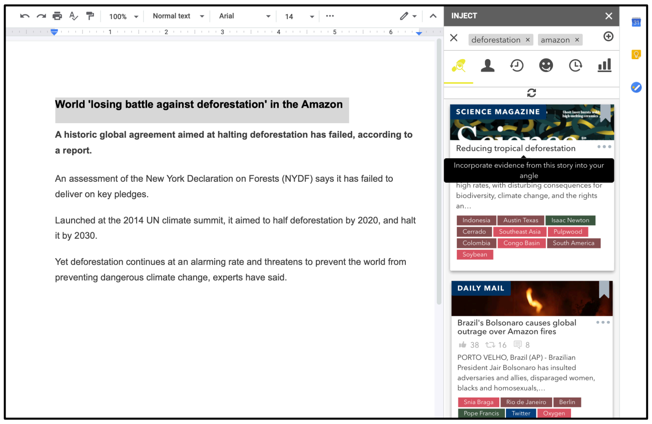
In response, a new digital tool called INJECT was developed to support journalists to discover new associations with which to generate stories with angles more novel and valuable than stories published previously, in less time and with less cognitive effort than with existing digital tools. INJECT invokes creative search algorithms to discover information in already published news stories that it presents to journalists to use to form new associations during their own creative thinking.

Journalists were included throughout the INJECT tool design process. Interviews were held with experienced and inexperienced journalists to discover problems, requirements and constraints. Paper-based then digital wireframes of the INJECT tool were developed and presented to professional journalists. New releases of the working INJECT software were prototyped for their usability and impact with professional and student journalists [2].

2. An Example of INJECT Support for Journalists

One snapshot of journalist usage of INJECT to discover new angles for stories about the deforestation of the Amazon rainforest is depicted in Figure 1. Imagine that the journalist is writing the story in the Google Docs text editor, and at any time, could invoke INJECT's support with one click. When invoked, this support is presented through the INJECT sidebar that appears on the right side of the text editor. The journalist is then able to highlight text in the editor from which to extract topic terms for INJECT to use to discover possible new angles. For example, from the highlighted working title *World losing battle against deforestation in the Amazon*, INJECT extracted the topic terms *world*, *battle*, *deforestation* and *Amazon*, which the journalist then chooses to reduce to *deforestation* and *Amazon*. These topic terms form the inputs to 6 alternative creative strategies, represented by the visual 6 icons at the top of the sidebar.

In our example, the journalist has then clicked the *Quantified information* icon, which triggered INJECT to invoke a creative search of its index of millions of published news stories that were manipulated to discover new angles for stories about *deforestation* and the *Amazon*. INJECT presents the results of these searches in its sidebar. The journalist is then able to scroll through subsets of discovered stories using different sidebar features. The example sidebar in Figure 1 shows two of these stories retrieved from *Science Magazine* and the *Daily Mail*. Each is described using its source, title, first sentence, and a set of 10 entities – places, things, people and organisations – extracted from the story and represented in colored rectangles that the journalist can use to discover multiple new angles.



World 'losing battle against deforestation' in the Amazon

A historic global agreement aimed at halting deforestation has failed, according to a report.

An assessment of the New York Declaration on Forests (NYDF) says it has failed to deliver on key pledges.

Launched at the 2014 UN climate summit, it aimed to half deforestation by 2020, and halt it by 2030.

Yet deforestation continues at an alarming rate and threatens to prevent the world from preventing dangerous climate change, experts have said.

Reducing tropical deforestation

Incorporate evidence from this story into your angle

high rates, with disturbing consequences for biodiversity, climate change, and the rights an…

Indonesia | Austin Texas | Isaac Newton | Cerrado | Southeast Asia | Pulpwood | Colombia | Congo Basin | South America | Soybean

Brazil's Bolsonaro causes global outrage over Amazon fires

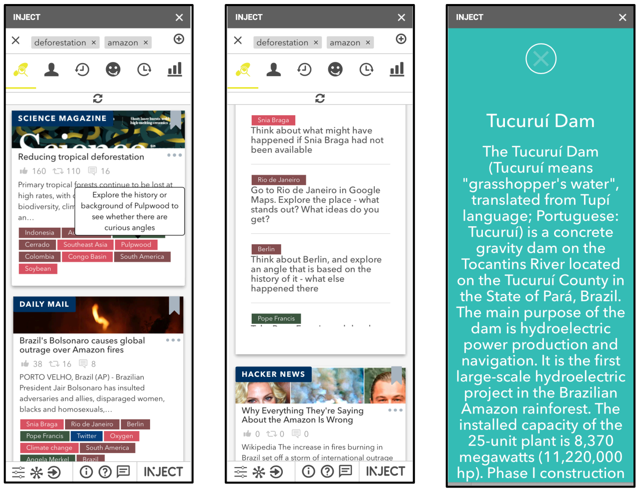
38 16 8

PORTO VELHO, Brazil (AP) - Brazilian President Jair Bolsonaro has insulted adversaries and allies, disparaged women, blacks and homosexuals,…

Snia Braga | Rio de Janeiro | Berlin | Pope Francis | Twitter | Oxygen

Figure 1. Use of the INJECT sidebar on the right side of text editors – in this case Google Docs – used by journalists

Positioning the cursor over each rectangle presents a pop-up creativity spark generated for that place, thing, person or organisation. This form of interactive creativity support was implemented as a mouse hover-over to enable journalists to explore multiple sparks and to discover different associations very quickly. The sparks themselves were designed to direct the deliberate generation of associations and ideas by journalists. Each was generated from a predefined set of spark types to direct journalists to think about, for example, the history and relevance of places, the motives of people and their opponents, the future and emotional impact of objects, and available data about organisations. The sidebar on the left side of Figure 2 shows one such pop-up spark when the mouse hovers over the entity *Pulpwood – Explore the history or background of Pulpwood to see whether there are curious angles.*



Reducing tropical deforestation

160 110 16

Primary tropical forests continue to be lost at high rates, with d… biodiversity, clim… an…

Explore the history or background of Pulpwood to see whether there are curious angles

Indonesia | Au… | Cerrado | Southeast Asia | Pulpwood | Colombia | Congo Basin | South America | Soybean

Brazil's Bolsonaro causes global outrage over Amazon fires

38 16 8

PORTO VELHO, Brazil (AP) - Brazilian President Jair Bolsonaro has insulted adversaries and allies, disparaged women, blacks and homosexuals,…

Snia Braga | Rio de Janeiro | Berlin | Pope Francis | Twitter | Oxygen | Climate change | South America | Angela Merkel | Brazil

Snia Braga
Think about what might have happened if Snia Braga had not been available

Rio de Janeiro
Go to Rio de Janeiro in Google Maps. Explore the place - what stands out? What ideas do you get?

Berlin
Think about Berlin, and explore an angle that is based on the history of it - what else happened there

Pope Francis

Why Everything They're Saying About the Amazon Is Wrong

0 0 0

Wikipedia The increase in fires burning in Brazil set off a storm of international outrage

Tucuruí Dam

The Tucuruí Dam (Tucuruí means "grasshopper's water", translated from Tupí language; Portuguese: Tucuruí) is a concrete gravity dam on the Tocantins River located on the Tucuruí County in the State of Pará, Brazil. The main purpose of the dam is hydroelectric power production and navigation. It is the first large-scale hydroelectric project in the Brazilian Amazon rainforest. The installed capacity of the 25-unit plant is 8,370 megawatts (11,220,000 hp). Phase I construction

Figure 2. Different views of the INJECT sidebar showing published stories retrieved using the creative search with the topic terms *deforestation* and *Amazon*, showing from left to right: a pop-up creativity spark directing the journalist towards new angles, a list of alternative sparks generated from retrieved stories, and more information about entities extracted from the stories

At any time the journalist can also request a list of all current creativity sparks, see the middle of Figure 2, to view and/or download into the text editor window for storage and later use. Furthermore, if the journalist clicks an entity, the INJECT sidebar presents more information about it extracted from Wikipedia, as shown on the right side of Figure 2. Other sidebar features that can support the journalist's creative thinking include the renewal of entities and creativity sparks presented for each story, word clouds of frequently occurring words in stories, and selected Google searches generated directly from retrieved stories. More information about these features in reported in [3].

3. The INJECT Architecture

To deliver these and other forms of support, INJECT was implemented with natural language processing, multi-language creative search and interactive creativity support capabilities. To provide journalists with a sufficiently large external information source to discover associations with, it generated indexes of content from millions of verified stories published by hundreds of news titles in multiple languages.

The tool's three-tier architecture is shown in Figure 1. The architecture separates the tool's functionality, the data and the interactive components so that the tool could be used in different news environments with different context management systems and news archives.

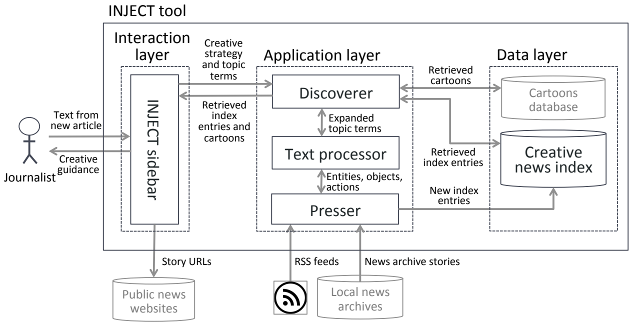


Figure 3. The INJECT tool's three-tier architecture, showing its layers, services and external information sources

The primary interaction layer was the INJECT sidebar, which was implemented with multiple versions to work in different text editors. This embedded sidebar was designed to be simple, fit with existing work practices, and encourage journalists to discover new angles on stories quickly without learning new skills. Furthermore, to support journalists not working with one of the mandated text editors, the interaction layer was extended with a web application version of the sidebar, to use in another desktop window next to the text editor window. The application layer was composed of 3 software services, the presser, text processor and discoverer, which were designed to operate together to generate large numbers of possible associations between information that journalists were writing about using indexed news content from millions of already published news stories. These software services contributed to and retrieved indexed news content from INJECT's data layer, called the creative news index, which was designed so that the discoverer service could undertake divergent creative searches that were more sophisticated than were possible with existing web search and news site APIs. This index was populated by the presser service, which indexed millions of verified news stories as possible starting points for discovering new angles on stories. The text processor service was invoked by the presser to make sense of and generated indexed content from published news, and by the discoverer to expand creative search queries.

The key INJECT services and components in the application and data layers are described in more detail.

3.1 INJECT's Presser

The presser service generated indexes of millions of verified news stories that could be retrieved, on request, as the starting points for journalists to discover associations and to generate new angles on stories. INJECT was designed to be a general-purpose journalist tool, so it manipulated news content on a wide range of topics. It had a crawler component that fetched news stories to index from open RSS feeds, and an importer component that fetched stories from accessible newspapers' archives.

The crawler was directed to fetch verified news stories from over 1000 predefined RSS feeds published by 380 diverse news titles in 6 languages. These feeds, titles and languages were selected by INJECT's editorial team to generate indexes of diverse views and angles on news, and ranged from major daily newspapers in the United States, regional newspapers in the Netherlands, and tabloid titles in the United Kingdom. On a normal news day, it fetched about 15,000 stories. Stories from high-frequency feeds were fetched every 30 minutes, others every 12 hours. During each fetch cycle, the crawler automatically read all news stories accessible via the URLs in the each RSS feed, removed navigation links, adverts and embedded media such as links, images and videos, and sent the remaining text string, along with story's author, URL, image URL and published date, to the text processor service. This text string, author, date and URLs provided a rich external information source with which journalists could discover and generate new associations and angles on news stories.

3.2 INJECT's Text Processor

The text processor service generated new entries to add to the creative news index by analysing the natural language text string of each fetched news story with:

- Named entity extraction mechanisms to index stories using real names such as of people and places. The mechanisms that treated candidate named entities as groups of consecutive words describing a concept such as a person (e.g. *Angela Merkel*), location (e.g. *Indonesia*), organisation (e.g. *United Nations*) or object (e.g. *soybean*). This enabled the processor to extract entities with which journalists might discover associations not described in the text, for example the entity *Angela Merkel* from the text *the Chancellor of Germany*;
- Automatic parser mechanisms that detected nouns and verbs to index stories using object names and action names. The parsers split news text into sentences then applied part-of-speech tagging to mark-up words as belonging to lexical, part-of-speech categories. Shallow parsing was applied to generate a machine understanding of the structure of a sentence without parsing it fully into a parsed tree form. The output was a division of the text's sentences into a series of words that, together, constituted a grammatical unit. To select candidate objects and actions from these units with which journalists might also discover associations, the mechanism applied lexical extraction heuristics on a syntax structure rule-tagged sentence. For example, the processor parsed the text *World losing battle against deforestation in the Amazon* to extract the nouns such as *battle*, *deforestation* and *Amazon*.

The entities generated by the text processor service provided the semantic backbone to each creative news index entry manipulated by other tool services.

3.3 INJECT's Creative News Index

For each fetched story, the creative news index generated a new entry composed of all extracted named entities, objects and actions and frequencies of occurrence, the author, URL, image URL and publication date. A typical entry for a news of about 400 words was between 30 and 50 entities, objects and actions. Early prototyping of the INJECT tool revealed that indexes with this volume and type of content were sufficient to generate new associations that journalists reported could be effective for discovering new angles.

All index entries were uploaded to an external Elasticsearch cluster to be manipulated by the discoverer's creative search algorithms. Elasticsearch is a scalable open source search engine with a REST API that provides scalable, near real-time search. The level of performance offered by Elasticsearch was essential to support journalists to discover new angles on stories more quickly. In November 2019, the Elasticsearch cluster held over 18million entries, with another 450,000 or so new entries being added each month.

3.4 INJECT's Creative Strategies and Algorithms

To discover news stories with which to form associations, journalist could select between 6 predefined creative strategies represented by the 6 icons in the sidebar shown in Figures 1 and 2. These strategies had been designed to mimic the strategies of experienced journalists [3]. Each was implemented in the discoverer service to retrieve creative news index entries with:

A. Quantified information associated with the topics;
B. Information about people associated with the topics;
C. Information about events associated with the background of the topics;
D. Information about future consequences associated with the topics;
E. Data sets and visualizations associated with the topics;
F. Comical information associated with the topics.

For the first 5 of these strategies the discoverer implemented a series of process steps – the disambiguation of the topic terms, invocation of Elasticsearch searches of the creative news index, scoring of retrieved index entries, and filtering by the selected strategy. Each of these process steps is described in some detail.

Disambiguation of Topic Terms

The first step disambiguated each noun topic term by discovering its correct sense in the online lexicon at WordNet using context knowledge from other topic terms in the query (e.g. that *Amazon* is *a South American river; arises in the Andes and flows eastward into the South Atlantic; the world's 2nd longest river* rather than *one of a nation of women warriors of Scythia*).

Assigning the correct sense to a word in context requires syntactic, semantic and pragmatic knowledge about the word itself, its part of speech, and its context [7]. The discoverer service adds syntactic and semantic information to query terms through part-of-speech tagging and WordNet. It disambiguates by iteratively using 3 procedures: (i) defining single term senses; (ii) defining synonyms, and; (iii) defining hypernyms. For each topic term T the algorithm determines its sense S using one of 3 procedures. If Procedure *i* does not provide any positive result, then Procedure *i+1* will be applied. In a generic iteration of the algorithm the input is a list of pre-processing terms $T = [ti,\ldots,tn]$, and a list of associated senses $S = [Sti,\ldots,Stn]$. T represents all terms to be disambiguated and S represents the semantic meaning of T, where Sti is either the chosen sense for ti or the empty set, i.e. the term is not yet disambiguated. A set of ambiguous terms $A = \{ti|\ Sti = \square\}$ is also maintained. T is initialized with the empty set $T = \{\}$ and A with the list formed by all parsed terms. The output is the updated list S of senses associated with the input terms T. In more detail, the 3 disambiguation procedures are:

- **(i) Defining Single-Sense Terms**: the procedure exploits the existence of terms with only one sense in WordNet, called monosenous terms, and tags them automatically with that sense. For example, the noun *deforestation* has one sense defined in WordNet and is tagged with that sense #1;
- **(ii) Defining Synonyms**: the procedure finds query terms that are semantically connected to already-disambiguated terms (i.e. terms with a tagged sense) and for which the connection distance is 0 as computed using WordNet hierarchies. A semantic distance of 0 between two terms defines that both belong to the same synset, and therefore the new term is tagged with the same meaning as the connected term. For example, consider the terms *Amazon* and *river* in *T*. The noun *river* is a monosemous word – *a large natural stream of water larger than a creek* – disambiguated with Procedure 2. One of the senses of the noun *Amazon*, sense #3 (*a South American river that arises in the Andes and flows east-ward into the South Atlantic*) also contains the monosemous term *river*, so the procedure tags *Amazon* with sense #3;
- **(iii) Defining Hypernyms**: similar to (ii), this procedure finds query terms that are semantically connected to already disambiguated terms but for which the connection distance is the maximum 1 as computed using WordNet hierarchies. A semantic distance of 1 between two words indicates that both belong to the same hypernymy/hyponymy relation and therefore the new term is tagged with the same meaning as the connected term.

Two additional procedures – frequency-based senses and context-based senses – were used in earlier versions of the algorithms but not implemented in INJECT due to the smaller number of topic terms input to INJECT's creative search algorithms. The frequency-based senses procedure assigns the most frequent sense to a term irrespective of its context [8]. This heuristic has been used to baseline supervised word sense disambiguation systems. Its high performance is due to the skewed frequency distribution of word senses. The context-based senses procedure exploits the SemCor bigrams method that forms two pairs, one with the previous word, the other with the next word, and searches for these pairs in SemCor corpus [5].

Invoking Elasticsearch Queries

The second step invoked an Elasticsearch search via the news API with the expanded query terms and logic operators set by the journalist to control search depth and breadth. Elasticsearch's distributed and open source search engine was well-suited to the extended textual information extracted from the news RSS feeds. All extracted and extended information was indexed in Elasticsearch. Once indexed, INJECT ran complex queries based on the different creative search strategies codified in the tool.

Scoring of Retrieved Index Entries

The third step scored all returned index entries for their relevance based on frequencies of the original and expanded query terms and names of extracted entities in the title and the body of each retrieved news article. All terms and entities were allocated a computed score, but the scoring mechanism was implemented to reflect the structure of most news stories with the most important information at the start of stories. Therefore:

- Terms and entities in the article title were assumed to be the most important, so if original term(s) and entity name(s) were present in the title, then each occurrence of each term/name was weighted 3x. Moreover, if all original terms/names were present in the title, then each occurrence of each term was weighted 8x. And if an expanded but not original term/name was present in the title, then each term/name was weighted 2x;
- Terms and entities in the more important first two paragraphs of the article body were also weighted. If original term(s) and entity name(s) were present in either of these first two paragraphs, then each term/name occurrence was weighted 2x. Moreover, when all original terms and entity names were present in the first two paragraphs, then the terms were weighted 3x. And if an expanded term/name was present in the first two paragraphs, then the score was weighted 2x.

This weighted scoring process was applied to each retrieved index entry.

Filtering by the Selected Strategy

The fourth step filtered the scored index entries using constraints specific to the selected creative search strategy, so that journalists were presented with information and sparks to form associations consistent with that strategy. To determine the scored index entries for quantified information associated with the topics, the discoverer filtered to retain entries with a minimum number of terms describing quantities, measures and values related to different news sectors including economy, finance, health, education, environment, agriculture and society. For example, to filter index entries for quantities related to economy, the discoverer scored entries that included terms such as *gross domestic product*, *interest rate* and *household disposable income*. Likewise, to filter entries for quantities related to health, the discoverer scored entries that included terms such as *health spending*, *hospital beds* and *body mass index*. The discoverer also scored entries including measurement units such as *degree* and *pint*, and numbers such as *one*, *two* and *hundred*.

To determine the scored index entries for information associated with the background of the topic terms (C), the discoverer filtered to retain entries at least 500 words in length with a minimum number of terms indicative of background articles. The minimum word length was added to reflect that background long-read articles are often longer than other article types. To filter the index entries with this minimum length, the discoverer scored entries with keywords indicative of background and cause – keywords such as *led to*, *result of* and *explanation*. It also scored entries with expressions of previous time periods such as *summer of YYYY*, *prior to YYYY*, and *during YYYY*.

The discoverer filtered scored index entries for the other creative search strategies using similar combinations of keywords and rules.

The outcome of this final step is a set of scored index entries that the application layer presented to the journalist. To manage this presentation, INJECT only presents a maximum of 30 articles corresponding to 30 retrieved index entries, although the sidebar allows the journalist to order the presentation of these articles either by relevance or publication date, or in random order.

3.5 INJECT's Creativity Sparks

The creative sparks service generated the pop-up sparks for each retrieved article and entity extracted from each article. An individual creative spark associated 1 extracted entity or news article to 1 creative instruction. The sets of instructions had been manually generated from websites and blogs that teach journalists to uncover new angles on stories. One set of instructions was generated for each of the 4 types of entity that were extracted – *people*, *events*, *places* and *organizations*. Examples included: *Unpick what the relevance of [Place], as opposed to somewhere else, might have on the story* and; *Explore the history and background of [Organization] to obtain a new perspective on your story*. A total of 34 creative instructions were implemented. One set of instructions was also generated for news articles retrieved with each of the 6 creative strategies – *people*, *causal*, *quirky*, *quantifiable*, *ramifications* and *data visualizations*, and a total of 41 such instructions were implemented. Examples included: *Use data types reported in this story, to generate a new angle*, and: *Make your angle more similar to the causal angle in this story*. When invoked, the service used a randomizing function to attribute one instruction string to one entity string of the same type, then concatenated the strings to generate the spark. So, for the extracted entity *Pulpwood* from Figure 2, INJECT might have presented a spark such as: *Explore the history and background of Pulpwood to obtain a new perspective on your story.*

4 Conclusions

This paper reports the tool's architecture, some its algorithms, and the design decisions made to deliver a reliable and usable tool for journalists in different newsrooms and work contexts. These design decisions have enabled INJECT to deployed quickly and at low cost in different newsroom contexts. The Presser service has also been applied quickly to index 10,000s of digital articles in the news archives of regional news organizations, to provide more localized creativity support for the journalists of these organizations. And the separation of the interaction and application layers has enabled INJECT to be deployed quickly with new text editors, for example with the Adobe InCopy tool used in newsrooms.

Moreover, the separate creative news index has emerged as a valuable standalone information asset. For example, the asset can be used by non-news businesses to understand quickly the news coverage of a topic in different languages or types of newspapers. At the moment, INJECT is being extended with new forms of computational creativity algorithms to landscape news coverage of a topic over a period of time and gain new creative insights for both editors and journalists.

ACKNOWLEDGMENTS

The research reported in this paper is supported by the EU-funded H2020 723328 INJECT innovation action.

REFERENCES

[1] B McNair (1998). The Sociology of Journalism. London: Arnold

[2] N Maiden, K Zachos, J Lockerbie, G Brock and C Traver (2016). Developing and Evaluating Digital Creativity Support in Google Docs for Journalists. In Proceedings 30th International British Human Computer Interaction Conference (HCI'16), Article No. 23. http://doi.acm.org/10.14236/ewic/HCI2016.23

[3] N Maiden, G Brock, K Zachos, A Brown (2018) Making the News: Digital Creativity Support for Journalists. ACM SIGCHI Conference 2018 20-27 April, Montreal, Canada, Paper No 475

[4] N Malmelin and S Virta (2016). Managing creativity in change: Motivations and constraints of creative work in a media organization. Journalism Practice 10,6: https://1041-1054. https://doi.org/10.1080/17512786.2015.10748646

[5] K. Miller (1993) Introduction to WordNet: an On-line Lexical Database, Distributed with WordNet software.

[6] H. Sjøvaag.(2014) Homogenisation or Differentiation? The Effects of Consolidation in the Regional Newspaper Market. Journalism Studies 15,5: 511-521

[7] M Stevenson and Y. Wilks (2001) The Interaction of Knowledge Sources in Word Sense Disambiguation, Computational Linguistics, 27(3): 321-349.

[8] Y Wilks and M Stevenson (1996). The grammar of sense: Is word-sense tagging much more than part-of-speech tagging?' Sheffield Department of Computer Science, Research Memoranda