MAPPING THE FIELD OF ALGORITHMIC JOURNALISM

A revised version of this article is forthcoming in: Digital Journalism

DOI:
10.1080/21670811.2015.1096748

Link, when finally published:
http://dx.doi.org/10.1080/21670811.2015.1096748

AUTHOR

Konstantin Nicholas Dörr

ADDRESS

Konstantin Nicholas Dörr, University of Zurich, Institute of Mass Communication and Media Research, Media Change & Innovation Division, Andreasstrasse 15, 8050 Zurich, Switzerland. Telephone: +41 (0) 44 635 20 35; E-mail: k.doerr@ipmz.uzh.ch, www.mediachange.ch, www.ipmz.uzh.ch

ACKNOWLEDGEMENTS

The author would like to thank Michael Latzer and Natascha Just as well as the two anonymous reviewers for their helpful and constructive comments.

FUNDING

The research was supported by a research grant of the Swiss National Science Foundation (SNF).

MAPPING THE FIELD OF ALGORITHMIC JOURNALISM

With software automatically producing texts in natural language from structured data, the evolution of natural language generation (NLG) is changing traditional news production. The paper answers the question if NLG is able to perform functions of professional journalism on a technical level first. A technological potential analysis therefore uncovers the technological limitations and possibilities of NLG, accompanied by an institutional classification following Weischenberg, Malik, and Scholl (2006). Overall, NLG is explained within the framework of algorithmic selection (Latzer et al. 2014) and along its technological functionality. The second part of this paper focuses on the economic potential of NLG in journalism as well as indicates its institutionalization on an organizational level. 13 semi-structured interviews with representatives of the most relevant service providers detail the current market situation. Following Heuss (1965), the development of the NLG market is classified into phases. In summary, although the market for NLG in journalism is still at an early stage of market expansion with only few providers and journalistic products available, NLG is able to perform tasks of professional journalism on a technical level. The analysis therefore sets the basis to analyze upcoming challenges for journalism research at the intersection of technology and big data.

KEYWORDS

Algorithmic journalism; automation; natural language generation; news production; robot journalism; technology; institution; market potential

Automating News Production

Historically, computerization shows that software has been taking over routine tasks (Frey and Osborne 2013) and "[j]ournalism has always been shaped by technology" (Pavlik 2000, 229). With algorithms editing, aggregating, publishing, and distributing content, processes of media production and consumption are increasingly being automated (Napoli 2014; Mager 2012; Pavlik 2013; Gillespie 2014; Broussard 2014; Diakopoulos 2014).

As newspapers struggle for profitability, market share, journalistic reputation, and readers, recent developments in news production go along with progress in natural language generation (NLG), a subfield of natural language processing (see Jones 2001 for an overview). Terms like "robot journalism," "automated journalism," "algorithmic journalism," or "machine-written journalism" dominate the media and scientific discourse (Anderson 2012 and 2013). Due to the rising availability of digital and digitized data, NLG is defined as software and computer systems, which automatically produce human (natural) language from a computational representation of information (Reiter and Dale 2000). Although companies like Narrative Science or Automated Insights are already able to ad and embed graphics or other media to their generated texts, these additons are not NLG in the understanding of computational linguistics as they rely on other technological premises (Reiter 2010). Therefore they are not part of this analysis. But of course, it is the interplay between NLG and these tools of visualization that enhances the variety and content diversity of journalistic products.

Presently, there are only few studies that discuss the impact of NLG on journalism. For example, Petre (2013) and Coddington (2014) identify and analyze a "quantitative turn"

in journalism in different concepts like data journalism, computational journalism, and computer-assisted reporting. On the other hand, Carlson (2014) discusses this development – using the term automated journalism – mainly in relation to labor and authority. Van Dalen (2012) conducted interviews with journalists to analyze their attitudes concerning the launch of a machine-written sports website. He demonstrates how this technology forces journalists to re-examine their skills. Latar (2015) notes the limits of NLG in a more universal understanding of the potential of algorithms, lacking an institutional discussion. With journalism offering "institutionalized solutions of communication problems" (Neuberger 1996, 12), the rising potential of NLG is also leading to rising user expectations (Bateman 2010, 640). Therefore, initial studies on the perceived quality and credibility of algorithmically generated texts show that differences between human and automated written texts are nearly undiscernible (van der Kaa and Krahmer 2014; Clerwall 2014). Questions about "how decisions of inclusion and exclusion are made, what styles of reasoning are employed, whose values are embedded into the technology, and how they affect public understanding of complex issues" (Young and Hermida 2014, 4) also challenge NLG on an ethical level. Carlson (2014) therefore addresses these questions with the concept of "algorithmic authority" and Gillespie (2014) discusses this development using the term "algorithmic objectivity." This overview shows that algorithms are entering the formerly sacred territory of "human" text production and call for research on various levels (Poynter 2014 and 2015). Besides the institutional discussion of automated news by Napoli (2014) and Anderson (2012 and 2013), communication science has not been able to supply a coherent model of NLG in journalism so far, as well as identifying the possibilities and limitations of this technology. Thus, this leads to the first of two central questions addressed in this article to examine the development at the "intersection of journalism and technology" (Lewis and Usher 2013, 603).

1. Is NLG able to perform functions of professional journalism on a technical level?

As a theoretical basis for an institutional classification of this technology in journalism, this paper answers the question if NLG is able to perform functions of professional journalism on a technical level. First, NLG and its technical functionality are explained within the framework of "Algorithmic Selection on the Internet" by an input-throughput-output-model (I-T-O model; Latzer et al. 2014). Latzer et al. (2014) identify algorithmic selection as the technical-functional core of many successful software applications fulfilling social tasks (e.g. reducing transactions costs). Moreover, they outline a functional typology and subsume text creation applications as a part of it. As outlined above, this paper solely focuses on content creation applications on the basis of NLG and its use in journalism, where they "touch deeply upon human areas of creativity and expression" (Latzer et al. 2014, 8) and point to a new professional role where journalists are "migrating from a direct to an indirect role" (Napoli 2014, 350).

To uncover the technical limitations and possibilities of NLG, a technological potential analysis is used as an analysis tool, following the potential analysis of online journalism by Neuberger (2001) and Wolf (2014) for applications in mobile journalism. Knowing that this analysis can be viewed from a technical and an institutional perspective (Wolf 2014), both positions are identified not as contrary but as two stages during the institutionalization of NLG in journalism. Based on Weischenberg, Malik, and Scholl (2006) it is assumed that journalism is constituted by three spheres: a social, organizational, and professional. The analysis of the technological potential of NLG is discussed within these spheres along with an

analysis of the basic functions, codes, and norms of journalism in the light of various practical and theoretical approaches to defining "journalism" (Neuberger 2002; Neuberger and Kapern 2013; Meier 2011).

Thus, this development is here conceptualized as Algorithmic Journalism and is defined as the (semi)-automated process of natural language generation by the selection of electronic data from private or public databases (input), the assignment of relevance of pre-selected or non-selected data characteristics, the processing and structuring of the relevant data sets to a semantic structure (throughput), and the publishing of the final text on an online or offline platform with a certain reach (output). It is produced inside or outside an editorial office or environment along professional journalistic guidelines and values that meet the criteria of topicality, periodicity, publicity and universality, and thus establishes a public sphere. The technology of NLG is furthermore identified as the central technical innovation that enables Algorithmic Journalism.

To outline the economic potential of NLG in journalism (see Lewis and Westlund 2014 for Big Data Journalism) and to discuss the process of institutionalization on the organizational sphere, the second part of the paper answers the question:

2. Who are the most relevant service providers in NLG with journalistic orientation and how is the present state of the market?

After identifying the most relevant service providers in NLG, 13 semi-structured interviews form the basis of this analysis. The interviews were conducted between 7 December 2014 and 2 September 2015 via face-to-face (3), telephone (1) and Skype (9) and lasted between 27 and 124 minutes. To gain insights in the development of the market and its journalistic orientation, this paper follows Heuss (1965) and classifies NLG into market phases. While publicly available information about the companies and their clients is limited, the interviews allow conclusions on the markets potential and the use of the technology in professional journalism.

The methodological and conceptual approach of this paper is based on triangulation, following Flick (2008, 12) and Denzin (1970; 1989). By triangulation the combination of methods and theoretical perspectives can lead to an increase in knowledge about NLG in journalism and is suitable for the analysis of this new technological development as a "fully grounded interpretive research approach" (Denzin 1989, 246).

The technological potential anlysis as well as the economic overview therefore set the basis to analyze upcoming challenges for journalism research at the intersection of technology and big data.

Algorithmic Selection and Natural Language Generation

First, NLG is explained within the framework of algorithmic selection along an I-T-O model. Although NLG is not a new research field, having first appeared in the 1950s as a minor part of machine translation (Reiter 2010; McDonald 2010), it only became an independent research subfield in the 1980s. The processes of language generation steadily advanced due to growing data availability and the importance of statistical data analysis (Reiter 2010). A popular application of NLG was textual weather forecasting, e.g. FOG (Goldberg et al. 1994) or SumTime (Reiter et al. 2005). Other NLG applications are based on sports (Robin and McKeown 1996), medical (Portet et al. 2007), financial (Kukich 1983) or engineering data (Yu et al. 2007). There are also narrative NLG applications to persuade or

motivate (Reiter et al. 2003) and entertain (Binstead and Ritchie 1997), leading to automated storytelling (Callaway and Lester 2002; Perez y Perez and Sharples 2004)

Media reports may discuss "robots" in journalism but more precisely, algorithms are used to generate text. Latzer et al. (2014), based on Cormen et al. (2009), define algorithms "[…] as a finite series of precisely described rules or processes to solve a problem" and as "a sequence of stages that transforms input through specified computational procedures (throughput) into output" (Latzer et al. 2014, 4). Algorithms are dynamic in nature as they are "constantly adjusted in efforts to improve their performance in accordance with specific criteria" (Napoli 2014, 344).

Depending on the purpose of the application Latzer et al. (2014) use this basic definition for the framework of "Algorithmic Selection on the Internet" where a selection of elements from a basic set is processed according to certain rules, followed by a structuring and ranking of relevant information into an intended output. These two aspects – the selection and the assignment of relevance – are reflected in the functionality of the framework and its applications. This basic functionality can also be applied to applications of NLG as they operate similarly (Reiter and Dale 2000; McDonald 2010; Reiter 2010; Carstensen et al. 2010). By "articulating the specifications of a system through a rigorous examination drawing on domain knowledge, observation, and deduction to unearth a model of how that system works" (Diakopolous 2014, 7), the mode of operation of NLG can be also illustrated by the design of an I-T-O model (figure 1).

Besides this adaption of functionality, NLG is also able to fulfill promises of algorithmic selection. As it will be shown in the technological potential analysis, NLG is able to reduce various kinds of transaction costs due to process automation (e.g., search and information costs) (Latzer et al. 2014, 29). Based on initial programming of the NLG software (e.g., predefined statistical rules), the software is therefore able to draw conclusions and identify interesting facts within the underlying set of data for text generation autonomously and much quicker than a "human" journalist is capable of.

[Here figure 1]

The starting point for NLG and its journalistic use is a data base, for example sports, financial, weather or traffic data (input), which is processed according to predefined linguistic and statistical rules (throughput) to a text (output) in natural language.

A language is a set of strings over some alphabet, which is sometimes specified by grammars within a grammar framework (Pratt-Hartmann 2010, 55). Within this framework, any grammar recognizes a unique language, which has to be coded in order to generate texts in different languages. Thereby the process of generation is often divided into three stages: 1. document planning, 2. micro planning, and 3. realization. The following section refers to the research of Reiter and Dale (2000) and Reiter (2010) for a simplified understanding of the architecture of NLG for journalistic news production.

In general terms, the goal of *document planning* (input level) is to identify the information that is useful to the user or the intended output. The input to the document planner – structured data – is the input to the entire NLG system. This data, the "main ingredient" of NLG in journalism, can be accessed via public APIs or via private data bases (e.g., internal client data). For content-related text generation and therefore every journalistic product based on NLG, individual codes, rules, and dictionaries have to be coded and adapted. Thus, NLG systems only operate according to pre-set specific rules. This includes the linguistic creation process as well as the decision which hidden facts in the data

should be processed and transformed into natural language. Thus, the final result is apart from some linguistic adjustments already almost definite. These pre-structured tasks are identified as a prerequisite at the input level/request level of Algorithmic Journalism (Reiter and Dale 2000, 49). This involves parameters such as text length, content/facts, journalistic form of presentation, theme, tonality as well as the time and place of publication.

The core of this generation process is the throughput level (*micro planning*) that defines the input-output relationship (*text planning* and *realization*). Starting from the input level (request), algorithms apply statistical operations to select elements from a basic data set and assign relevance to them. The NLG system must decide which linguistic structures (words, syntax, sentences) should be used to communicate the desired information and in the *realization* (throughput) stage, it must decide which forms of words to use, and in which order they will appear. Feedback loops thereby indicate the human influence on the generation process which is optimized until the desired result is achieved. As Reiter (2010, 577) notes, this content generation process is complex and requires many decisions, including the *lexical choice* (choosing which content and words should be used to express domain concepts and data reference); *referring expressions* to identify domain entities; *syntactic choice* (choosing syntactic structures in generated sentences and *aggregation* (choosing how many messages should be expressed in each sentence).

The result (output) is a text in natural language. After this generation process, the texts are published mainly automatically on online or offline news outlets.

Technological Potential Analysis of NLG

To answer the question if NLG is able to perform functions of professional journalism, a technological potential analysis is used as a suitable tool to discover the technical possibilities and limitations of a technology (Neuberger 2001, 92).

As pointed out, this analysis can be viewed from a technical and an institutional perspective (Wolf 2014), while both positions are identified not as contrary but as two stages during the institutionalization of NLG in journalism. Also Young and Hermida point out that "the development and application of algorithms in journalism requires both a technological and sociological lens" (2014, 3). Therefore, NLG is regulated and influenced by the characteristics of the Internet as an enabling-technology for data collection and text distribution via cloud based solutions. From a sociological lens (see RQ2), e.g. the integration of NLG products into organizational routines of professional media organizations have to be discussed (Kiefer 2010; Neverla 2001). As for now, this paper focuses on the technological lens first to analyze NLG from a primarily technology-centered view. Knowing that the analysis of its technical potential is only a necessary first step of a broader institutional analysis as expectations of media companies, the industry, and the public have to be taken into account (Wolf 2014, 70).

The technical perspective is illustrated according to the technological potential analysis by Neuberger (2001) for online journalism and Wolf (2014) for applications in mobile journalism. For this analysis ten technical potential dimensions are identified, derived from scientific literature focusing NLG (Reiter and Dale 2000; McDonald 2010; Reiter 2010; Carstensen et al. 2010) and journalism (Wolf 2014; Neuberger 2001). As there is limited research on NLG and its use in journalism, additional technological characteristics of NLG from the interviews were taken into account (here: production routines).

The final potential dimensions therefore include established online characteristics like multimediality, additivity, topicality, selectivity, context sensivity and interactivity

(Neuberger 2003, 57) as well as specific technical features of NLG like data availability, data quality, language diversity, and production routines (see table 1).

To frame this technological development for journalism research, this paper does not propose a new theory or definition of journalism but instead provides a comprehensive classification. Although there are many possible understandings of journalism as well as theoretical approaches (Löffelholz 2008; Quandt 2005, 23), it is agreed upon that journalism is a form of public communication (Quandt 2000, 484). In this paper, "journalism" is seen as "as a social system that enables society to observe itself, as it provides the public independently and periodically with information and issues that are considered newsworthy, relevant, and fact-based" (Weischenberg, Malik, and Scholl 2012, 207). Central to this theoretical understanding are media organizations as institutions, which fulfill specific functions for society (Weischenberg, Malik, and Scholl 2006, 347).

Analyzing NLG in journalism from an institutional perspective, this paper follows Weischenberg, Malik, and Scholl (2006) and argues that journalism is constituted through a social, organizational and professional sphere. On a social level, journalism fulfills certain tasks by observing parts of society and provides the public with relevant information. This also includes fact-based products with a certain reach (Weischenberg, Malik, and Scholl 2006, 346) as well as products of special interest (Rühl 1980, 382). Journalism is traditionally produced on the organizational level within media organizations according to specific rules and routines or by other journalistic actors on a professional level (Wolf 2014, 66; Hohlfeld 2003, 127; Weischenberg, Malik, and Scholl 2006, 346). Furthermore, the principles of topicality, periodicity, publicity and universality (Groth 1960, 360) are *not* linked to a technical artifact such as the newspaper and are also transferable to the general understanding of Algorithmic Journalism outlined above (Wolf 2014, 67). As long as the technology meets the prerequisites above, it is able to perform tasks and fulfill functions of professional journalism on a technical level (see table 1).

[Here table 1]

This preliminary study of the technological potential dimensions serves as a base for the institutional classification afterwards.

Multimediality, Interactivity, and Additivity

While multimedia is the integration of different media for the communication and mediation of information, including text, photo, graphic, video, animations, and audio (Meier 2002, 129), presently, NLG in the sense of computational linguistics is the generation of texts in human language (Reiter 2010). But as outlined above, it is the combination of different tools and other technologies that complement journalistic NLG products. While digital journalism in general is still primarily text-based and is only enriched with photographs, videos and graphics (Quandt 2008, 140), NLG as the core technology constantly progresses and fulfills the prerequisite of multimediality. As there are already NLG systems that include hyperlinks (Reiter 2010, 595), therefore *interactivity* as the technological potential to enable consumer interaction is possible (Neuberger 2007, 43). The more customized the final text product, the more effective and usable it is for the reader or client. This is a major reason for the use of NLG systems in journalism (Bateman 2010, 635). Hypertextuality as a form of journalistic text design therefore enables "a non-linear, modular presentation of texts with multimedia elements" (Neuberger 2003, 63) and enables to link

and distribute already existing content (*additivity*). On the one hand, the text can be the final product, but may also be the starting point for further editing adding additional information like quotes or images. The Internet therefore acts as a carrier medium and distribution channel, while the consumption of this content is not linked to a specific device. Overall, NLG can be seen as a product extension of digital journalism.

Topicality

Publication speed is important in digital journalism and online content can be continuously updated if (live) data is available, thereby ignoring traditional production routines (Neuberger 2003, 60). Presently, structured data sources are hardly available for NLG (e.g., only for specific domains). The term topicality includes the speed of content production within media organizations as well as the relevance of a topic (Meier 2003, 253). The connectivity to data bases (public or private, offline or online) therefore shortens the time for content production, as NLG operates independently after final programming. Depending on the availability, quality and topicality of the data, as well as the technological infrastructure of the provider and the client, NLG is able to generate a large number of texts at any time. Thus, news are far more visible, e.g. to search engine indexing. This is not to say that NLG systems do not produce errors as they are definitely able to do wrong reporting based on wrong coding or wrong data sources (Fox News 2015). This is also leading to ethical challenges. As for example The Associated Press has stopped monitoring every single generated text produced by Automated Insights for their earnings reports as it is too time consuming (Turi2 2015).

Selectivity and Context Sensivity

Selectivity in the sense of social sciences focuses on the user and the potential of the system to personalize and select specific functions for him (e.g., layout or content; Neuberger 2003, 61 and 2007, 44). While NLG needs the technological prerequisites of the Internet (e.g., cloud based distribution of texts and data collection) for a broad integration into journalism, it is possible that users select special topics and journalistic niche products and personalize them based on their individual interests once texts are generated. As Wolf (2014, 94) notes, content refers to the local, temporal, event-related, and interest-specific sphere of the user. This can be news on the favorite football club or the latest developments of the stock market. With NLG being *context-sensitive* and service providers heavily focusing on processing real-time data for text generation (e.g., stock market news), the audience is able to get the right information at the right time and place according to their individual interests (e.g., push messages in mobile use).

Data Availability, Data Quality, Language Diversity, and Production Routines

With data being the main resource, service providers have partnerships with media agencies and data brokers in terms of data acquisition. On the other hand, they process data already delivered and scraped by their clients. While the technological and linguistic capabilities of NLG are steadily increasing, the availability of data is the main issue. By now, companies and clients tend to scrape data via open APIs, however, they often do not disclose these origins. Highly structured and detailed data improve the quality of the generated texts. Today, NLG systems are even able to generate texts in multiple languages

(Reiter and Dale 2000; also see table 3). Moreover, the automation of NLG reduces transactions costs for search and information and journalists are thus able to focus on other relevant tasks as the software is processing the data autonomously after coding.

Considering that NLG is a highly complex process in the field of computational linguistics, algorithms are not able to generate texts without human interference. But still, these new production routines are leading to shifts in journalistic roles. Although evolving journalistic work processes are not new, they often force "new tasks on reporters and editors alike" (Powers 2012, 27). I argue that the direct and active human element *during* the process of content creation is eliminated in Algorithmic Journalism. This is not to say that the human factor is eliminated from content creation altogether, because algorithms are themselves developed by humans. The point is that the individual journalist in NLG is changing to a more indirect role (Napoli 2014) *before*, *during* and *after* text production. As for example source selection (input), fact checking, the actual writing (throughput – both depending on coding) and distribution (output) are automated and pre-selected in Algorithmic Journalism. Journalistic work and knowledge now has to comprise skills like programming even more.

The effectiveness of modern NLG systems is predicated on domain specialization, i.e. they "[…] restrict themselves to a specialized area of discourse with a very focused audience and stipulated content, thereby reducing the options for word choice and syntactic style to a manageable set" (McDonald 2010, 126). Therefore, the development of the semantic web (Berners-Lee, Hendler, and Lassila 2001), which focuses on the large availability of structured knowledge domains, offers vast opportunities for the further development of NLG systems. To date, NLG and its use in journalism is feasible in clearly defined domains such as weather, traffic, finance, or sports with good input data availability. However, once the domain and underlying set of world knowledge expands, NLG struggles with ambiguity of words and phrases (e.g., ball: leather ball or dance?). This is why companies focus on journalistic products of the sports and financial domain where processable and structured data (e.g., .xls, .csv, .xml, .json) are available (see table 3). Data for text generation has to be available in a scalable amount as programming and individualization of journalistic products are cost-intensive.

Another limitation is that NLG is not able to draw conclusion from contradictory data. For example, there is no source checking and the data bases have to be very reliable and well-maintained (Carstensen et al. 2010).

Independent journalistic interpretation and reflection, e.g. of political issues, poses major challenges for NLG. Journalistic narration (e.g., arc of suspense) is also severly limited (Reiter 2010). NLG in journalism is thus more of a starting point and information can be added to the generated text (e.g., quotes from stakeholders, experts etc.).

The following table 2 sums up the extent of the premises, possibilities, and limitations of NLG derived from the potential analysis.

[Here table 2]

Building on the technical perspective, the following part discusses the first research question if NLG is able to fulfill functions of professional journalism on a technical level. This is based on the institutional understanding of Weischenberg, Malik, and Scholl (2006).

Relevant Information, Including Special Interest Topics

With NLG already able to produce articles for different sports (e.g., ice hockey, basketball, American football, football), weather, traffic, finance, and celebrity news, it also produces relevant information for special interest topics (e.g., NCAA college sports such as baseball and lower-division basketball and football; see the results of the interviews in table 3). This adds value to traditional news coverage by offering coverage that was previously not profitable.

Reach

The beginning integration of NLG into the daily operation of media organizations is already able to generate a certain reach (see table 3). For example, the major wire service The Associated Press (AP) uses this technology to generate earning reports for their clients in the AP Style Guide. Such reports are labeled "This story was generated by Automated Insights using data from Zacks Investment Research." AP also invested in Automated Insights, a NLG service provider, via the private equity firm Vista Equity Partners (AP 2015).

Information Produced within Media Organizations / Newsrooms

As news is already produced within media organizations and the newsroom (e.g., the fine dust monitor by Berliner Morgenpost was developed by a team of five journalists and programmers). NLG services and complete articles can also be bought externally from service providers. As long as these services are labelled and communicated within the brand of the media organization, this fulfills the requirements of a production within organization.

Professional Journalistic Actors

To date, already media organizations are testing and using NLG services in journalism (see table 3), evaluating whether products match the standards and quality of their brand and how journalists and users react. These case studies also guarantee the influence of the professional journalistic actor on the individual product, which is tailored to the needs of the client and the audience. Therefore, the software training guarantees the matching of the individual writing style and the expected output. No media organization integrates automated news without serious testing, as media organizations have more to lose than just readers (e.g., credibility of their brand). At this stage of product development, the journalistic actor heavily influences the processing and final version of the text. But questions about "new" actors arise if service providers or other players outside professional journalism (e.g., data brokers) decide to reach an audience without professional journalistic intermediaries.

Information Produced According to Specific Rules and Routines

In general, "the use of algorithms for media production contains an editorial logic based on the socially situated choices of media professionals" (Young and Hermida 2014, 384). The influence of the journalist in NLG is visible as the texts are generated according to different rules and routines specific to the media organization. For example, the adjustment of the AP earning reports matching the AP style guide took nearly one year before launch (ASBPE 2015). This also fits the requirements of an institutionalized journalism. Media

organizations have to be transparent about the data sources and the labeling of the texts. This is closely linked to algorithmic transparency, which Stavelin (2013) argues is borrowed from professional journalistic values. The legal challenges of copyright and accountability are quite unclear and the role of the journalist is in flux. Consequently, the routines of news production are changing.

Principle of Topicality, Periodicity, Publicity and Universality

Argued from a technological-institutional perspective, NLG is able to perform the principle of topicality (e.g., Quakebot reports immediately after the earthquake; sports results after or during the games). The principle of periodicity is fulfilled as texts can be generated regularly – if data is available. When news are integrated within a special journalistic offering, automated texts are accessible publicly via the Internet. With NLG also covering a number of different topics (sports, entertainment, finance, weather), it also fulfills the principle of universality, to date with a limitation to mostly special interest content.

Observation of Society

The above conditions enable the observation of society within different topics, although there are technical limitations (reasoning, reflection, and interpretation). Despite this, NLG is able to add relevant information within special topics to the traditional news coverage in order to fulfill journalistic tasks of orientation and opinion formation. Anderson (2011, 541) suggests that Algorithmic Journalism "might represent the most recent, and thus most unsettling, model for both communication and democracy" and Coddington notes that forms of quantitative journalism like Algorithmic Journalism have "great potential to broaden journalism's ability to make democratic institutions more responsive and legible to the public" (2014, 2).

Thus, the analysis of the technological potential as well as the institutional discussion show that overall, NLG is able to fulfill certain tasks of professional journalism on a technical level. Therefore, NLG is conceptualized on a technical level as Algorithmic Journalism and is defined as shown above.

The Emergent Market for Algorithmic Journalism

After the technological potential analysis, which outlined that NLG is able to perform tasks of professional journalism, a review of available market data gives insights about the present state of the market for Algorithmic Journalism indicating the process of institutionalization of NLG on an organizational level. The classification into market phases therefore allows conclusions on its potential growth as there is a connection between market phases and market structures. Heuss (1965) notes that markets go through five ideal-type stages during their development: the experimental phase where the product is invented, developed, and launched; the expansion phase characterized by exponential growth; the maturity phase where growth diminishes; the stagnation phase with nearly no growth; and finally the regression phase where the market reverts back. These phases are often characterized by high concentration and by temporary monopolies of innovators and early movers (Latzer et al. 2014, 14).

While NLG providers were not able to enter the market on a long-term till the end of 2010 (Reiter 2010), there are companies and start-ups that enter the market with different

products based on NLG. A literature analysis first identified 13 leading service providers with NLG technology as their core business. Given the very limited information available about such companies – particularly on the development and offer of *journalistic* NLG products – interviews were conducted to complement the market overview. These semi-structured expert interviews, which lasted between 27 and 124 minutes, were conducted with the following companies (for a complete list of all interviewees see appendix of this article): Narrative Science (USA) – via Skype; Automated Insights (USA) – via Skype; OnlyBoth (USA) – via Skype; Linguastat (USA) – via Skype; Retresco (GER) – face-to-face; Aexea (GER) – face-to-face; Text-on (GER) – face-to-face; Textomatic (GER) – via telephone; 2txt (GER) – via Skype; Sylabs (FR) – via Skype; Labsense (FR) – via Skype; YSEOP (FRA/USA) – via Skype; Arria (UK) – via Skype.

As a first result, ten out of 13 providers offer products with possible applications in journalism (see table 3). In order to draw a holistic picture of the market, all service providers were asked about the legal form of the company, the founding year, external funding, employees, languages of text generation, topics covered for journalistic use, products available, journalistic clients as well as potential competitors. First, no additional competitors were named indicating that the market of NLG and especially for its application in journalism is served by only a few (see table 3).

With Tencent, a Chinese social and gaming service provider, there is already a new competitor rising. Unfortunately Tencent could not be part of the interviews due to the submission deadline of this paper but was added to the analysis in order to be as up-to-date as possible. Via Dreamwriter, a Tencent-designed NLG tool, they published a business report written in Chinese covering basic financial news (South Morning China Post 2015).

As the companies differ in size, assets, and product portfolio, they decrease their dependency on one market segment as they also offer NLG solutions for e.g. e-commerce, finance, oil industry, healthcare or the farming industry. This likely due to the complexity of NLG, the limited availability of data, the time-consuming individualization of journalistic products due to high quality standards in journalism, and the general view that journalistic products alone are hardly profitable. This complexity also forces media organizations to buy most of these services externally as they lack of newsroom infrastructures and fincancial resources for development.

Considering the company's size, clients, and already launched products, Automated Insights and Narrative Science are the driving forces of Algorithmic Journalism in the US. In Germany these are Retresco and Aexea, in France Syllabs and in Great Britain Arria. Other service providers like YSEOP (FRA/USA), Linguastat (USA), OnlyBoth (USA), who were also interviewed, do not offer products and NLG services for journalistic use and are therefore excluded from further analysis. While the majority of the providers do not receive external funding, Automated Insights, Narrative Science, Arria, and Labsense have external investors. The other companies are privately held. As the market is in constant flow, even companies with long-term experience in the field of NLG like Arria are quickly struggling for survival if clients terminate their contracts (e.g., Arria and Shell Exploration & Production Company; Arria NLG 2015). The investments by external business angels are not an indicator of market success. It is more of a vision of the technology's potential. The interviewees refused to give details regarding the revenue of the companies. Pilot projects can be demarcated at about €25.000. Although service providers report that they are constantly negotiating with well-known media organizations, only few products have officially been launched – those that are available mainly serve financial and sports reporting.

The primary reason why these companies focus on journalistic products for the sports and financial domain is that processable and structured data are available in a scalable amount. Additionally, the language, rules, and settings can be easily defined in terms of programing in a discrete framework and domain.

As a result of the interviews table 3 shows the market and gives an overview of the potential and core uses of Algorithmic Journalism. With few service providers, limited and resembling journalistic products available, Algorithmic Journalism is likely found either in an experimental market phase or in an early stage of market expansion phase. But the integration of these products within professional media organizations already indicates a starting institutionalization on an organizational sphere (Weischenberg, Malik, and Scholl 2006).

[Here table 3]

Conclusion

Based on the framework of algorithmic selection and the technological potential analysis, this paper identified the technological possibilities and limitations of NLG in journalism. Its journalistic application was conceptualized as Algorithmic Journalism with NLG serving as the central technical innovation enabling it. The described technical limitations as well as the dependence of NLG on structured data does not change the fact that NLG is able to perform institutionalized tasks of professional journalism on a technological level.

This mainly technological discussion therefore sets the basis to analyze upcoming challenges for journalism research at the intersection of big data like the shift of journalistic roles before, during, and after content production, ethical and legal challenges in news production, or questions about new intermediaries in journalism as the dependency on data grows.

Becuause news production remains a for-profit business, it is evident that publishers on the one hand seek to reduce costs (Stavelin 2013, 27; Kiefer 2001, 33). On the other hand, they are also desperately looking for new journalistic products and ways to satisfy their audience. The interviews and the classification into market phases show that the market for Algorithmic Journalism is still relatively small. But service providers already pressure traditional media organizations (AI 2015a, b). The integration of AJ into the portfolio of media organizations also indicate a starting institutionalization of Algorithmic Journalism on an organizational level. As the costs of NLG systems are low compared to human journalists, Algorithmic Journalism can be profitable for special interest domains, also due to the possibility to generate news in multiple languages and therefore to reach a broader audience and new markets.

REFERENCES

AI. 2015a. “Robot Reporters: How The AP Embraced Data Automation | Strata + Hadoop World.” https://www.youtube.com/watch?t=31&v=lBzXiMiwQQs.

AI. 2015b. “When Robots Write The News, What Will Humans Do? | SXSW Interactive.“ https://www.youtube.com/watch?v=0LqLHAtLE-c.

Anderson, Christopher W. 2011. “Understanding the role played by algorithms and computational practices in the collection, evaluation, presentation, and dissemination of

journalistic evidence." Paper presented at the 1st Berlin Symposium on the Internet and Society, Berlin, Germany, October 25–28.

Anderson, Christopher W. 2012. "Towards a sociology of computational and algorithmic journalism." *New Media and Society* (15) 7: 1005–1021.

Anderson, Christopher W. 2013. "What aggregators do: Towards a networked concept of journalistic expertise in the digital age." *Journalism* (14) 8: 1008–1032.

AP. 2015. "Private equity firm Vista buying Automated Insights." http://www.ap.org/Content/AP-In-The-News/2015/Private-equity-firm-Vista-buying-Automated-Insights.

Arria NLG. 2015. "Termination of contract and trading update". Email from Stuart Rogers, CEO Arria NLG, 30.4.2015.

ASBP American Society of Business Press Editors. 2015. http://www.asbpe.org/blog/2015/05/03/aps-tom-kent-the-time-has-arrived-for-robotics-journalism-ethical-checklist/.

AX Semantics. 2015. "Sportberichte." https://www.ax-semantics.com/products/sports-content/.

Bateman, John. 2010. "Angewandte natürlichsprachliche Generierungs- und Auskunftsysteme." In *Computerlinguistik und Sprachtechnologie. Eine Einführung,* edited by Kai-Uwe Carstensen, Christian Ebert, Cornelia Ebert, Susanne Jekat, Ralf Klabunde and Hagen Langer, 633-641. Heidelberg: Spektrum.

Berners-Lee, Tim, James Hendler, and Ora Lassila. 2001. "The semantic web". *Scientific American* (284) 5: 24–30.

Binstead, Kim, and Graeme Ritchie. 1997. "Computational rules for punning riddles." *Humor. International Journal of Humor Research*. (10) 1: 25–76.

Broussard, Meredith. 2014. "Artificial Intelligence for Investigative Reporting. Using an expert system to enhance journalists' ability to discover original public affairs stories." *Digital Journalism* (online first), DOI: 10.1080/21670811.2014.985497.

Callaway, Charles, and James Lester. 2002. "Narrative prose generation." *Artificial Intelligence* (139) 2: 213–52.

Carlson, Matt. 2014. "The Robotic Reporter. Automated journalism and the redefinition of labor, compositional forms, and journalistic authority." *Digital Journalism* (online first), DOI: 10.1080/21670811.2014.976412.

Carstensen, Kai-Uwe, Christian Ebert, Cornelia Ebert, Susanne Jekat, Ralf Klabunde, and Hagen Langer. 2010. *Computerlinguistik und Sprachtechnologie. Eine Einführung.* Heidelberg: Spektrum.

Clerwall, Christer. 2014. "Enter the Robot Journalist. Users' perceptions of automated content." *Journalism Practice*, (8) 5: 519–531.

Coddington, Mark. 2014. "Clarifying Journalism's Quantitative Turn. A typology for evaluating data journalism, computational journalism, and computer-assisted reporting" *Digital Journalism* (online first), DOI: 10.1080/21670811.2014.976400.

Cormen, Thomas H., Charles E. Leiserson, Ronald L. Rivest, and Clifford Stein. 2009. *Introduction to Algorithms*. Cambridge, Mass: MIT Press.

Denzin, Norman K. 1970. *The Research Act*. Chicago: Aldine.

Denzin, Norman K. 1989. *The Research Act*. Englewood Cliffs, N. J.: Prentice Hall.

Diakopoulos, Nicholas. 2014. "Algorithmic Accountability. Journalistic investigation of computational power structures." *Digital Journalism* (online first), DOI: 10.1080/21670811.2014.976411.

Flick, Uwe. 2008. *Triangulation. Eine Einführung*. Wiesbaden: VS Verlag.

Fox News. 2015. “Correction: Earns-Graham Holdings story”. http://www.foxnews.com/us/2015/08/07/correction-earns-graham-holdings-story/.

Frey, Benedikt Carl, and Michael A. Osborne. 2013. “The Future of Employment: How Susceptible are Jobs to Computerisation?” Working Paper, University of Oxford, http://www.oxfordmartin.ox.ac.uk/downloads/academic/The_Future_of_Employment.pdf.

Gillespie, Tarleton. 2014. “The Relevance of Algorithms.” In Media Technologies. Paths Forward in Social Research, edited by Tarleton Gillespie, Pablo Boczkowski, and Kirsten Foot, 167-194. London: MIT Press.

Goldberg, Eli, Norbert Driedger, and Richard Kittredge. 1994. “Using natural-language processing to produce weather forecasts.” *IEEE Expert (*9) 2: 45–53.

Gray Jonathan, Lucy Chambers, Liliana Bounegru. 2012. *The Data Journalism Handbook. How Journalists Can Use Data to Improve the News*. Sebastopol: O´Reilly.

Groth, Otto. 1960. *Die unerkannte Kulturmacht. Grundlagen der Zeitungswissenschaft* (Periodik). Berlin: Walter de Gruyter.

Gynnild, Astrid. 2014. “Journalism innovation leads to innovation journalism: The impact of computational exploration on changing mindsets” *Journalism* (15) 6: 713–730.

Hamilton, James T., and Fred Turner. 2009. “Accountability through Algorithm. Developing the Field of Computational Journalism.” A report from Developing the Field of Computational Journalism, a Center For Advanced Study in the Behavioral Sciences Summer Workshop, July 27-31.

Heuss, Ernst. 1965. *Allgemeine Markttheorie*. Tübingen: Mohr.

Hohlfeld, Ralf. 2003. *Journalismus und Medienforschung. Theorie, Empirie, Transfer*. Konstanz: UTB.

Jones, Karen Sparck. 2001. “Natural Language Processing. A Historical Review.” Working Paper, University of Cambridge. http://www.cl.cam.ac.uk/archive/ksj21/histdw4.pdf.

Kiefer, Marie-Luise. 2001. *Medienökonomik*. München: Oldenbourg.

Kiefer, Marie-Luise. 2010. *Journalismus und Medien als Institutionen*. Konstanz: UVK.

Kubicek, Herbert, Ulrich Schmid, and Heiderose Wagner. 1997. *Bürgerinformationdurch „neue“ Medien? Analysen und Fallstudien zur Etablierungelektronischer Informationssysteme im Alltag*. Opladen: Westdeutscher Verlag.

Kukich, Karen. 1983. “Design andimplementation of a knowledge-basedreport generator.” Proceedings of 21st Annual Meeting of the Association for Computational Linguistics, 145–150.

Latar, Noam Lemelshtrich. 2015. “The Robot Journalist in the Age of SocialPhysics: The End of Human Journalism?” In *The Economics of Information, Communication, and Entertainment. The Impacts of Digital Technology in the 21st Century*, edited by Gali Einav, 65–80. Wiesbaden: Springer.

Latzer, Michael / Hollnbuchner, Katharina / Just, Natascha / Saurwein, Florian (2014): The economics of algorithmic selection on the Internet. University of Zurich, Zurich. Online: http://www.mediachange.ch/media//pdf/publications/Economics_of_algorithmic_selection_WP.pdf

Lewis, Seth C., and Nikki Usher. 2013. “Open source and Journalism: Toward New Frameworks for Imagining News Innovation. *Media, Culture and Society*, 35 (5): 602–619.

Lewis, Seth C., and Oscar Westlund. 2014. “Big Data and Journalism. Epistemology, expertise, economics, and ethics.” *Digital Journalism* (online first) DOI: 10.1080/21670811.2014.976418.

Löffelholz, Martin. 2008. “Heterogeneous—Multidimensional—Competing: Theoretical Approaches to Journalism – An Overview.” In *Global Journalism Research*, edited by Martin Löffelholz and David Weaver, 15–27. Hoboken, NJ: John Wiley & Sons.

Mager, Astrid. 2012. “Algorithmic ideology.” *Information, Communication and Society* (15) 5: 269–787.

Mayo, Justin, and Glenn Leshner. 2000. “Assessing the credibility of computer-assisted reporting.” *Newspaper Research Journal* (21) 4: 68–82.

McDonald, David D. 2010. “Natural Language Generation.” In *Handbook of Natural Language Processing*, edited by Nitin Indurkhya and Fred J. Damerau, 121–141. Chapman and Hall: Boca Raton, FL.

Meier, Klaus. 2002. “Neue journalistische Formen.” In *Internet-Journalismus*, edited by Klaus Meier, 21–172. Konstanz: UVK.

Meier, Klaus. 2003. “Qualität im Online-Journalismus.” In *Qualität im Journalismus. Grundlagen,Dimensionen, Praxismodelle*, edited by Hans-Jürgen Bucher and Klaus-Dieter Altmeppen, 247–266. Wiesbaden: Westdeutscher Verlag.

Meier, Klaus. 2011. *Journalistik.* Konstanz: UTB.

Napoli, Philip M. 2014. “Automated Media: An Institutional Theory Perspective on Algorithmic Media Production and Consumption.” *Communication Theory* (24) 3: 340–360.

Neuberger, Christoph, and Peter Kapern. 2013. *Grundlagen des Journalismus*. Wiesbaden: VS Verlag.

Neuberger, Christoph. 1996. *Journalismus als Problembearbeitung. Objektivität und Relevanz in der öffentlichen Kommunikation*. Konstanz: UVK.

Neuberger, Christoph. 2000. “Journalismus als systembezogene Akteurskonstellation. Vorschläge für die Verbindung von Akteur-, Institutionen- und Systemtheorie.“ In *Theorien des Journalismus. Ein diskursives Handbuch, edited by* Martin Löffelholz, 287–303. Wiesbaden: Westdeutscher Verlag.

Neuberger, Christoph. 2001. *Journalismus im Internet. Theoriekontext und empirische Exploration*. unveröffentlichte Habilitationsschrift. Universität Eichstätt-Ingolstadt, Eichstätt.

Neuberger, Christoph. 2002. “Vom Unsichtbarwerden des Journalismus im Internet.“ In *Innovationen im Journalismus*. Forschung für die Praxis, edited by Ralf Hohlfeld, Klaus Meier, and Christoph Neuberger, 25–68. Münster: LIT Verlag.

Neuberger, Christoph. 2003. “Zeitung und Internet: Über das Verhältnis zwischen einem alten und einem neuen Medium.“ In *Online – Die Zukunft der Zeitung?*, edited by Christoph Neuberger, and Jan Tonnemacher, 16–110. Wiesbaden: Westdeutscher Verlag.

Neuberger, Christoph. 2007. “Interaktivität, Interaktion, Internet.“ *Publizistik* 52 (1): 33–50.

Neverla, Irene. 2001. “Das Netz – eine Herausforderung für die Kommunikationswissenschaft.“ In *Kommunikationskulture nzwischen Kontinuität und Wandel. Universelle Netzwerke für die Zivilgesellschaft,* edited by Ursula Maier-Rabler und Ulrich Latzer, 29–46. Konstanz: UVK.

Parasie, Sylvain, and Eric Dagiral. 2012. “Data-driven journalism and the public good: “Computerassisted reports“ and “programmer-journalists“ in Chicago.” *New Media and Society* (15) 6: 1–19.

Pavlik, John. 2000. “The impact of technology on journalism.” *Journalism Studies* (1) 2: 229–237.

Pavlik, John. 2013. “Innovation and the Future of Journalism.” *Digital Journalism* 1 (2): 181–193.

Perez y Perez, Rafael, and Mike Sharples. 2004. “Three computer-based models of storytelling: BRUTUS, MINSTREL, and MEXICA.” *Knowledge-Based Systems* (17) 1: 15–29.

Petre, Caitlin. 2013. “A Quantitative Turn in Journalism?” Tow Center for Digital Journalism, October 30. http://towcenter.org/blog/a-quantitative-turn-in-journalism/.

Portet, François, Ehud Reiter, Jim Hunter, and Somayajulu Sripada. 2007. “Automatic generation of textual summaries from neonatal intensive care data.” Proceedings of the 11th Conference on Artificial Intelligence in Medicine (AIME 2007), 227–36.

Powers, Matthew. 2012. “In Forms That Are Familiar and Yet-to-be Invented: American Journalism and the Discourse of Technologically Specific Work.” *Journal of Communication Inquiry* 36 (1): 24–43.

Poynter. 2014. “L.A. Times reporter talks about his story-writing ‘Quakebot’.” http://www.poynter.org/news/mediawire/243744/l-a-times-reporter-talks-about-his-story-writing-quakebot/.

Poynter. 2015. “AP will use software to write NCAA game stories.” http://www.poynter.org/news/mediawire/324601/ap-will-use-software-to-write-ncaa-game-stories/.

Pratt-Hartmann, Ian. 2010. “Computational Complexity in Natural Language.” In *Handbook of Computational Linguistics and Natural Language Processing*, edited by Alexander Clark, Chris Fox, and Shalom Lappin , 43 -73. Oxford: Wiley-Blackwell.

Quandt, Thorsten. 2000. “Das Ende des Journalismus? Online-Kommunikation als Herausforderung für die Journalismusforschung.“ In *Theorien des Journalismus. Ein diskursives Handbuch*, edited by Martin Löffelholz, 483–559. Wiesbaden: Westdeutscher Verlag.

Quandt, Thorsten. 2005. *Journalisten im Netz. Eine Untersuchung journalistischenHandelns in Online-Redaktionen*. Wiesbaden: VS Verlag.

Quandt, Thorsten. 2008. “Neues Medium, alter Journalismus? Eine vergleichendeInhaltsanalyse tagesaktueller Print- und Online-Nachrichtenangebote.“ In *Journalismus online – Partizipation oder Profession?*, edited by Thorsten Quandt und Wolfgang Schweiger, 131–155. Wiesbaden: VS Verlag.

Reiter, Ehud, and Robert Dale. 2000. *Building Natural Language Generation Systems*. Cambridge: Cambridge University Press.

Reiter, Ehud, Roma Robertson, and Liesl Osman. 2003. “Lessons from a failure: generating tailored smoking cessationletters.” *Artificial Intelligence* (144) 1-2: 41–58.

Reiter, Ehud, Somayajulu Sripada, Jim Hunter, and Jin Yu. 2005. “Choosing words in computer-generated weather forecasts.” *Artificial Intelligence (*167) 1-2: 137–169.

Reiter, Ehud. 2010. “Natural Language Generation.” In *The Handbook of Computational Linguistics and Natural Language Processing*, edited by Alexander Clark, Chris Fox, and Shalom Lappin, 574–598. Oxford: Wiley-Blackwell.

Robin, Jacques, and Kathleen R. McKeown. 1996. “Empirically designing andevaluating a new revision-based modelfor summary generation.” *Artificial Intelligence* (85) 1-2: 135–179.

Rühl, Manfred. 1980. *Journalismus und Gesellschaft. Bestandsaufnahme und Theorieentwurf*. Mainz: v. Hase & Koehler.

South China Morning Post. 2015. “End of the road for journalists? Tencent's Robot reporter 'Dreamwriter' churns out perfect 1,000-word news story - in 60 seconds.” Online:

http://www.scmp.com/tech/china-tech/article/1857196/end-road-journalists-robot-reporter-dreamwriter-chinas-tencent.
Stavelin, Eirik. 2013. *Computational Journalism: When Journalism Meets Programming.* Norway: University of Bergen.
Turi2. 2015. "AP setzt auf Roboterjournalismus." Online: http://www.turi2.de/aktuell/ap-setzt-auf-roboterjournalismus/.
Van Dalen, Arjen. 2012. "The algorithms behind the headlines. How machine-written news redefines the core skills of human journalists." *Journalism Practice* (6) 5-6: 648–658.
Van der Kaa, Hille, and Emiel Krahmer. 2014. "Journalist versus news consumer: The perceived credibility of machine written news." Research Paper presented at the 2014 Computation + Journalism Symposium, Columbia University, New York. Online: http://compute-cuj.org/cj-2014/cj2014_session4_ paper2.pdf.
Weischenberg, Siegfried, Maja Malik, and Armin Scholl. 2006. "Journalismus in Deutschland 2005. Zentrale Befunde der aktuellen Repräsentativbefragung deutscher Journalisten." In *Media Perspektiven* (7), 346–361.
Weischenberg, Siegfried, Maja Malik, and Armin Scholl. 2012. "Journalism in Germany in the 21st Century." In *The Global Journalist in the 21st Century*, edited by David H. Weaver and Lars Willnat, 205–219. New York: Routledge.
Wolf, Cornelia. 2014. *Mobiler Journalismus. Angebote, Produktionsroutinen und redaktionelle Strategien deutscher Print- und Rundfunkredaktionen*. Baden-Baden: Nomos.
Young, Mary L., and Alfred Hermida. 2014. "From Mr. and Mrs. Outlier to Central Tendencies." *Digital Journalism* (online first) DOI: 10.1080/21670811.2014.976409.
Yu, Jin, Ehud Reiter, Jim Hunter, and Chris Mellish. 2007. "Choosing the content oftextual summaries of large time-series data sets." *Natural Language Engineering* (13) 1: 25–49.

INTERVIEWS

- *Frank Feulner,* CTO, aexea communication, Stuttgart, Germany
- *Saim Alkan*, CEO, aexea communication, Suttgart, Germany
- *Wolfgang Zehrt*, CTO, text-on, Berlin, Germany/now: textomatic AG
- *Claude de Loupy*, CEO and Co-Founder Syllabs (Data2Content), Paris, France
- *Ehud Reiter,* Chief Scientist, ARRIA NLG, Aberdeen, United Kingdom
- *Johannes Sommer,* Managing Director, Retresco, Berlin Germany
- *Johannes Bubenzer,* Managing Director, 2txt NLG, Berlin, Germany
- *Raul Valdes-Perez,* CEO, Only Both, Pittsburgh, USA
- *Arden Manning*, SVP Global Marketing, YSEOP, London et al., USA, FR, UK
- *John M. Pierre*, CEO and Co-Founder, Linguastat, San Francisco, USA
- *Linda Hayes*, CRO and COO, Linguastat, San Francisco, USA
- *Edouard de Ménibus*, Co-Founder, Labsense, Paris, France
- *James Kotecki*, Media & Public Relations, Automated Insights, Durham (NC), USA
- *Craig Booth*, Data Engineer, Narrative Science, Chicago, USA

DESCRIPTION OF TABLES

Table 1

Analytical dimensions of the potential analysis of NLG derived from Neuberger (2001), Wolf (2014), Reiter and Dale (2000), McDonald (2010), Reiter (2010), Carstensen et al. (2010) and accompanying interviews with the service providers

Characterstics of the Internet	Characteristics of NLG
– Multimediality – Interactivity – Additivity – Topicality – Selectivity – Context sensivity	– Data availability – Data quality – Language diversity – Production routines

Table 2

Technical premises, possibilities and limitations of NLG in journalism

Premises	Possibilities	Limitations
– Data availability – Data quality	– Production routines, e.g. reduce transactions costs for search & information – Selectivity, e.g. conquer niche markets and special interest content – Language diversity – Topicality; e.g. information and distribution speed; visibility of content (SEO) – Multimediality – Additivity – Context sensivity	– Production routines, e.g. world knowledge (special domains and contexts) – Production routines, e.g. reflection of content & assessment of facts, reasoning – Production routines; e.g. narration (arc of suspense) – Production routines, e.g. source checking – Interactivity

Table 3

Market of service providers of NLG with focus on journalism

Company	Country	Legal Form	Founding Year	Launch NLG Software	External Funding	Employees	Languages (To date)	Topics covered for journalistic use	Journalistic Products	Journalistic Clients
Automated Insights	USA	Inc., subsidiary of Stats LLC	2007	2007, branded as Wordsmith 2014	About $10.8 Million from 11 Investors before being acquired by Vista Equity Partners	About 40	(1) ENG	– Finance – Sports	– Corporate earnings stories (AP) – NCAA College sports (AP) (in dev.) – Yahoo Sports Fantasy Football	– Associated Press (USA) – Yahoo! (USA)
Narrative Science	USA	Inc.	2010	2011 Quill	About $32.4 Million from 7 Investors	About 80	(1) ENG	– Finance – Sports	– Earnings estimates of stock market companies (Forbes) – Sports statistics (e.g. basketball, American football, softball, baseball) (launched).	– Forbes (USA) – Big Ten Network (USA) – Game Changer (USA) – 5-10 signed contracts with US media (not public)
Aexea	GER	GmbH	2001	2009 AX Semantics	no	About 42	(12) ENG, GER, FR, ESP, NL, DNK, SWE, NOR, IT, IDN, PRT, CH	– Sports – Entertainment – Finance – Weather	– Match announcements for all European Soccer Leagues (in German), American Football – Stock exchange reports – Celebrity football news (in German)	– 5 media clients (non-disclosure agreement) + Sports-Information-Service (SID) (GER)
Text-On	GER	GmbH	2013	2014 Text-On	no	About 6, not payed	(1) GER	– Finance	– Pilot-project in the financial sector (in dev.; Berliner Morgenpost) – Share price descriptions	– Berliner Morgenpost (GER) – Finanzen100.de (GER)
2txt NLG	GER	UG	2013	2013 2txt	no	About 5	(1) GER	– Finance – Sports	– Football product (in dev.) – Share price descriptions (in dev.)	– Beginning of negotiations
Retresco	GER	GmbH	2008	2013 Rtr text engine	no	About 27	(1) GER	– Sports	– Preliminary reports of football games in lower German leagues (Kreisklasse)	– FussiFreunde (GER) – Neue Osnabrücker Zeitung – Weserkurier – Radio Hamburg FussiFreunde – Rheinfussball – Goekick.info – Fubanews.org
Textomatic	GER	AG	2015	2015 Textomatic	no	About 5	(6) GER, ENG, ESP, NL, FR, ITA	– Sport – Finance – Travelling – Weather	– Football – Stock exchange reports – Travel advices – Personalized weather reports	– 2 media clients (Handelsblatt and 1 regional newspaper)
Syllabs	FR	LLC	2006	2012 Data2content	no	About 11	(3) ENG, FR, ESP	– Politics	– Project on departmental elections 2015 in France	– Le Monde (FR)
Labsense	FR	SAS	2011	2013 Scribt_	Less than $565.000	About 6	(3) FR, ENG, GER	– Economy	– Project on local news in France (in dev.)	– Beginning of negotiations
Arria	GB	PLC	2011	2012 Arria NLG Engine	About $36 Million from shares	About 50	(1) ENG	– Weather	– Weather report module for two regions in Europe (UK and Germany)	– MeteoGroup (UK)
Tencent	CHN	Ltd.	1998	2015 Dreamwriter	No information	No information	(1) CHN	– Finance	– CPI report on China´s growth	– No information

DESCRIPTION OF FIGURES

Figure 1

I-T-O model Algorithmic Journalism based on Latzer et al. (2014), Reiter and Dale (2000)