**Institution:** University of Aberdeen

**Unit of Assessment:** 11 (Computer Science and Informatics)

**Title of case study:** Mainstream communication of big data using natural language generation (NLG)

**Period when the underpinning research was undertaken:** 2000-2012

**Details of staff conducting the underpinning research from the submitting unit:**

| Name(s): | Role(s) (e.g. job title): | Period(s) employed by submitting HEI: |
|---|---|---|
| Ehud Reiter | Professor (Chair) in Computing Science | 1995-present |
| Somayajulu (Yaji) Sripada | Senior Lecturer in Computing Science | 2000-present |

**Period when the claimed impact occurred:** 2013-2020

**Is this case study continued from a case study submitted in 2014?** Y

**1. Summary of the impact** (indicative maximum 100 words)

Research led by Professor Reiter and Dr Sripada developed natural language generation technology (NLG), which can generate English summaries of complex data. In 2009, this research led to (A) formation of spinout company, now called Arria NLG and (B) creation of an open-source NLG library, Simplenlg. In the 2014-2020 period, NLG technology was adapted by many more types of users (who did not use NLG before 2014). For example, Arria's NLG technology was used by media companies (e.g. BBC) to generate news stories, professional services companies (e.g. Ernst & Young) to generate financial reports, and business intelligence (BI) users to get insights about data; and Simplenlg was used by the travel company Trivago to generate hotel descriptions and Bayerischer Rundfunk, a public-service radio and television broadcaster in Munich.

**2. Underpinning research** (indicative maximum 500 words)

Professor Ehud Reiter and Dr Yaji Sripada are pioneers in the science of natural language generation (NLG) [1-6]. NLG is a subfield of artificial intelligence and represents a process that automatically transforms data into plain-English content, combining analytical output with engaging, contextualised narratives to provide text that is relatable and as 'human' as possible. As stated by Forbes magazine, 'In a paragraph and a few bullet points, we can quickly tell a rich and complex story… the bigger game of NLG is not about the language but about handling the growing number of insights that are being produced by big data through automated forms of analysis.'

Reiter and Sripada's original work on NLG came from a deep interest in finding a way to describe data using natural language that could be used in 'real-world' scenarios, such as medical emergencies, or for patients participating in clinical trials. Their vision was that NLG could help non-analysts benefit from big data instead of being overwhelmed by it. Reiter and Sripada are among the world's foremost authorities in the field of NLG, and have played a key role in the development of NLG architecture.

As reported during REF2014, Reiter and Sripada led an EPSRC-funded research project at the University of Aberdeen from 2000-2003 titled '*SumTime: Generating Summaries of Time-Series Data'* (GR/M76881/01), which applied NLG to meteorological and engineering data [P1]. The versatility of NLG through this first testing phase led to automatically generated narrative weather forecasts from raw weather prediction data [1] and generating summaries for maintenance engineers on gas turbine sensor data [2]. The researchers demonstrated that computer-generated texts could be superior to human-written texts, largely because the computer-generated texts offered more consistency in terms of both content and wording. Since human writers do not necessarily have a good understanding of what a non-specialist audience might find interesting in

terms of content or how they use words, this means that a computer system, which is "tuned" to a reader can often communicate more effectively to a personalised audience than a human writer who lacks knowledge of the reader.

The findings from this project led to further funding and conception of the 2003-2005 ESRC-funded Paccit-Link project '*Automatic Generation of Personalised Basic Skills Summary Reports'* (RES-328-25-0026) [P2] and the *BabyTalk* project (2006-2012), which was supported by several linked EPSRC grants (EP/D049520/1, EP/D05057X/1, EP/H042938/1) [P3] and two EPSRC DTA studentships in collaboration with NHS Lothian [3, 4]. The findings from the *BabyTalk* project demonstrated for the first time that a data-to-text system could be developed to generate useful summaries of complex and diverse physiological time series data such as heart rate, blood pressure information to aid medical personnel in monitoring babies in neonatal intensive care units [3]. These findings enabled Reiter and Sripada to develop novel architecture for implementation of summaries consisting of four stages, organised into a pipeline: Signal Analysis (recognises discrete patters in numeric data), Data Interpretation (higher-level messaging, recognition of relations between messages), Document Planning (allows key messages to be outlined, document and rhetorical structure) and Microplanning and Realisation (creates fluent text capable of expressing concepts and structure) [5].

In 2009, the Aberdeen NLG research group released the *Simplenlg* 'realisation engine', an open source Java library, which was developed to perform realisation and later microplanning [6]. Simplenlg continues to evolve (currently Simplenlg release 4.5.0) and has been updated by the Aberdeen team to accommodate new features [https://github.com/simplenlg/simplenlg].

As previously reported in REF2014, Reiter and Sripada created a spinout company, Data2Text Ltd (2009-2013). The company was acquired by Arria NLG and changed its name to Arria Data2 Text Ltd (company number 355243) in January 2014.

**3. References to the research** (indicative maximum of six references)

**References (citations via Scopus)**


[1] **E Reiter**, **S Sripada**, J Hunter, J Yu, and I Davy (2005). Choosing Words in Computer-Generated Weather Forecasts. *Artificial Intelligence* **167**:137-169, http://dx.doi.org/10.1016/j.artint.2005.06.006, 174 citations

[2] J Yu, **E Reiter**, J Hunter, C Mellish (2007). Choosing the content of textual summaries of large time-series data sets. *Natural Language Engineering* **13**:25-49, https://doi.org/10.1017/S1351324905004031, 69 citations

[3] F Portet, **E Reiter**, A Gatt, J Hunter, **S Sripada**, Y Freer, C Sykes (2009). Automatic Generation of Textual Summaries from Neonatal Intensive Care Data. *Artificial Intelligence* **173**:789-816, http://dx.doi.org/10.1016/j.artint.2008.12.002, 151 citations

[4] J Hunter, Y Freer, A Gatt, **E Reiter**, **S Sripada**, C Sykes (2012). Automatic generation of natural language nursing shift summaries in neonatal intensive care: BT-Nurse. *Artificial Intelligence in Medicine* **56**:157–172 https://doi.org/10.1016/j.artmed.2012.09.002, 46 citations

[5] **E Reiter** (2007). An Architecture for Data-to-Text Systems. In *Proceedings of ENLG-2007*, pages 97-104, https://aclweb.org/anthology/W/W07/W07-2315.pdf, 101 citations

[6] A Gatt and **E Reiter** (2009). SimpleNLG: A realisation engine for practical applications. In *Proceedings of ENLG-2009*, pages 90-93, https://www.aclweb.org/anthology/W09-0613.pdf, 206 citations


**Grants**


[P1] SumTime: Generating Summaries of Time-Series Data, EPSRC; 01/00-09/02; (GBP215,311; [GR/M76881/01])

[P2] Automatic Generation of Personalised Basic Skills Summary Reports, ESRC/EPSRC/DTI; 10/03-09/05; (GBP106,594; [RES-328-25-0026])

[P3] BabyTalk: Generating textual summaries of clinical temporal data, EPSRC; 09/06-08/1012, 09/06-08/10, (GBP330,756; [EP/D049520/1]; GBP148,024 [EP/D05057X/1] GBP134,123; [EP/H042938/1])

**4. Details of the impact** (indicative maximum 750 words)

**ARRIA**

Reiter and Sripada's research led to the formation of a spinout company Data2Text, which was acquired by Arria NLG in 2013. Reiter and Sripada were seconded part-time to Data2Text and then to Arria NLG and became its Chief Scientists. In their Arria role, they were granted over 30 patents on NLG in the 2014-2020 period (https://www.arria.com/core-tech/arria-core-patents/). Most of these patents built on and extended the NLG technology and concepts, which they developed through research at Aberdeen University such as [5].

Since it was founded, Arria has raised over USD80,000,000 in investments to support NLG technology. Arria currently has offices in Aberdeen, London, Auckland (New Zealand), Sydney (Australia), and Morristown New Jersey (USA). It employs over 100 people, including 47 people in the UK (primarily in Aberdeen). The majority of Arria's developers are located in the UK [S1].

In the 2014-2020 period, Arria applied its NLG technology and made an impact on many new markets, including ***business intelligence***, ***financial reporting***, and ***automated journalism***.

*Business intelligence* (BI) tools such as Tableau and PowerBI are used by analysts and data scientists to understand key business data about sales, staffing, customer turnover, etc. BI tools provide excellent visualisations of data which work well for data scientists and skilled analysts but can be confusing for others. Arria's NLG tools supplement BI tools such as Tableau by automatically generating English narratives that summarise BI data and present key insights from it; this allows many more people to identify, extract and understand actionable and useful insights from BI data that might otherwise by hidden in information spread across multiple spreadsheets or data sources. For example, the lead director of the Advanced Analytics Global Business Services at AstraZeneca said, 'Arria NLG Studio's BI analytical and linguistic capabilities provide us with the technology to dramatically improve the speed and efficiency of our reporting. A key benefit for us is that we can use this powerful platform to create our own NLG applications within our department' [S2].

In 2019, Gartner, a global research and advisory firm, reported that Arria is now 'a world leader in NLG, with 23 core NLG patents. Its Arria NLG Studio product as an enterprise-level, self-service NLG design tool with advanced analytics and linguistic functions that enables users to easily create, edit and publish NLG narratives' [S3].

Although most Arria BI use cases are financial, there are exceptions. For example, the analytics company Tibco created a public Covid-19 dashboard, which was supported by Arria technology. It used narratives and visuals to provide up-to-date information to a wide range of users about the status of the pandemic, including trends in infection rates and deaths, and which countries were most severely affected by Covid-19 [S4i]. In 2020, Tibco gave Arria its Global ISV Partner of the Year award 'in recognition of the collaboration between two leaders in technology innovation' [S4ii]. The companies' joint work on the dashboard has enabled public and private sectors draw greater meaning from complex data [S4iii].

A number of large accountancy and financial firms use Arria's NLG Studio tool to produce routine *financial reports*. Ernst & Young, one of the largest multinational professional services firms in the world, developed an NLG portal powered by Arria NLG Studio (https://nlg.ey.com/#/) [S5i]. The portal launched in 2018, now titled as the 'Accelerate Portal', enables users to write reports powered by NLG, turning raw data into insightful narratives at an unprecedented speed and scale, supporting trade activity commentary and investment reporting [S5ii]. BNY Mellon also use Studio (https://www.arria.com/bny-mellon/) [S5iii] and have stated "*We're working with Arria to build a set of capabilities in portfolio management, portfolio analysis and performance measurement that will help our clients significantly reduce the amount of time they spend writing personalised client summaries and reports and allow them to spend more time interacting with their clients*." [S5iiii]

Arria has launched an 'Arria for Accountants' tool, to make its technology accessible to smaller accountancy firms.

Several media companies use Arria's NLG technology for *automatic journalism*; that is to produce initial drafts of articles, which can be checked manually and can be used to fill the gap left by geographically-segmented news articles. In 2017, Urbs Media, news automation specialists, teamed up with leading UK news agency, Press Association to create 30,000 localised news reports every month using Arria [S6i]. In early 2019, the BBC piloted the use of semi-automated journalism Arria Studio [S6ii]; in late 2019, the BBC used Arria Studio to generate election result stories for all 690 individual constituencies after the UK general election in December 2019 [S6iii]. These stories were checked (and if necessary) edited by human journalists, and then released to local BBC new sites and blogs. The editor of BBC News Labs said, "*Using machine assistance, we generated a story for every single constituency that declared last night with the exception of the one that hasn't finished counting yet. That would never have been possible [if carried out by human journalists]*" [S6iii].

**SIMPLENLG**

The NLG technology developed by Reiter and Sripada's research was also released through the open-source Simplenlg package (found on https://github.com/simplenlg). Simplenlg has been used by institutions who want the freedom provided by open-source, and/or want to generate narratives in languages other than English. While Arria has focused on English, Simplenlg is available in 8 European languages and Mandarin Chinese [S7].

In the 2014-2020 period, Simplenlg has also been used by many academic research groups who are exploring applications of NLG. For example, Universidade de Santiago de Compostela (Spain) used Simplenlg to generate weather forecasts, the Università degli Studi di Torino (Italy) used Simplenlg to provide dietary advice, and Université de Genève (Switzerland) used Simplenlg to help authors create interactive narratives [S8].

Examples of commercial usage for Simplenlg in 2014-2020 include:

- Trivago, an internet travel company, which helps customers around Europe make travel arrangements, used Simplenlg to generate narrative descriptions of hotels on its website. This provides up-to-date and consistently written descriptions, which are optimised for search engine optimisation (SEO) [S9i].
- Bayerischer Rundfunk (BR) (Bavarian Broadcasting, a public-service radio and television broadcaster, based in Munich). Although not formally announced until February 2021, BR has been using Simplenlg for sports reporting since December 2020. An article by BR generated with Simplenlg is shown in [S9ii].

**5. Sources to corroborate the impact** (indicative maximum of 10 references)

**ARRIA**

[S1] Details of Arria's expansion and profile since 2014, including new jobs created globally

***Business intelligence*** and ***financial reporting***

[S2] CISION article (2019), corroborates quote from Lead Director, Advanced Analytics Global Business Services, AstraZeneca

[S3] Artificial Intelligence, article (2020), corroborates position as world leader in NLG

[S4 (group)] (i) Tibco Blog (2020): Covid-19 dashboard; (ii) Global ISV Partner of the Year award; Arria's role in project; (iii) improving Covid-19 hospitals operations with analytics

[S5 (group)] (i) Ernst & Young (E&Y) landing page (powered by Arria); (ii) PRNewswire article (2018) corroborating statement from Chief Innovation Officer, (iii) BNY Mellon landing page (powered by Arria), (iiii) PRNewswire article (2018) corroborating statement from Global Head of Data and Analytics Solutions, BNY Mellon

***Media companies***

[S6 (group)] (i) Forbes article (2017), details collaboration between Press Association and Urbs Media and their use of Arria; (ii) Medium article (2019), outlines BBC News Labs experiments with semi-automated journalism using Arria; (iii) BBC News labs article (2019) details use of Arria to cover 2019 general election

**SIMPLENLG**

[S7 (group)] Simplenlg available in 8 European languages and Mandarin Chinese

[S8] Further applications of Simplenlg, as reported by academic research groups in Spain, Italy and Switzerland

[S9 (group)] (i) Research paper detailing Trivago's use of Simplenlg; example of an article utilising Simplenlg by Bayerischer Rundfunk (BR); (ii) email correspondence confirming BR use of Simplenlg to generate articles