IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received June 1, 2018, accepted July 3, 2018, date of publication August 1, 2018, date of current version August 28, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2861987

No Landslide for the Human Journalist - An Empirical Study of Computer-Generated Election News in Finland

MAGNUS MELIN[1], ASTA BÄCK[1], CAJ SÖDERGÅRD[1], MYRIAM D. MUNEZERO[2], LEO J. LEPPÄNEN[2], AND HANNU TOIVONEN[2]

[1]VTT Technical Research Centre of Finland Ltd., FI-02044 Espoo, Finland
[2]Department of Computer Science, University of Helsinki, 00100 Helsinki, Finland

Corresponding author: Magnus Melin (magnus.melin@iki.fi)

This work was done as part of the Immersive Automation Project, which is funded by Business Finland, The Media Industry Research Foundation of Finland, The Swedish Cultural Foundation in Finland, the media companies involved as well as the research organizations participating in the project (VTT Technical Research Centre of Finland Ltd., University of Helsinki, Sanoma Media Finland Oy, Alma Media Oyj, Conmio Oy, Keski-Pohjanmaan Kirjapaino Oyj, Streamr, and KSF Media Ab).

ABSTRACT In an age of struggling news media, automated generation of news via natural language generation (NLG) methods could be of great help, especially in areas where the amount of raw input data is big, and the structure of the data is known in advance. One such news automation system is the Valtteri NLG system, which generates news articles about the Finnish municipal elections of 2017. To evaluate the quality of Valtteri-produced articles and to identify aspects to improve, $n = 152$ users were asked to evaluate the output of Valtteri. Each evaluator rated six preselected computer-generated articles, four control articles written by journalists, and four computer-generated articles of their own choice. All the articles were evaluated along four dimensions: credibility, liking, quality, and representativeness. As expected, the texts written by Valtteri received lower ratings than those written by journalists, but overall the ratings were satisfactory (average 2.9 versus 4.0 for journalists on a five-point scale). Valtteri's best rating (3.6) was for credibility. The computer-written articles that the evaluators could freely select got slightly better ratings than the preselected computer-written articles. When looking at the results by demographic groups, males aged 55 or more liked the automatic articles best and females aged 34 or less liked them the least. Evaluators mistook 21% of the computer-written articles as written by humans and 10% of the human-written articles as computer-written. The share of users making these mistakes grew with the age. Overall, the male evaluators made less writer-identification mistakes than female evaluators did.

INDEX TERMS Artificial intelligence, automated content generation, automated storytelling, natural language processing, robot journalism.

I. INTRODUCTION

The media sector has been a rapidly changing landscape in the last few decades. Internet has altered the way we as a society consume news in terms of the physical medium (paper to electronic), immediacy (daily papers to instant updates), format (static text and pictures to interactive stories) and many more. Further change is ongoing with the introduction of computer systems that automatically produce textual content for the customers to consume [9]. However, the current efforts in automated news generation have been focused on well-understood domains with fairly limited space of possible article archetypes, such as sports news and news flashes regarding earthquakes. The reasons for this are complex, but can be boiled down to the fact that producing complex news articles requires complex systems that either employ opaque statistical methods (i.e. machine learning, neural networks) that are problematic for the news industry, or alternatively are transparent but costly to set up.

Van der Haak *et al.* [14] summarize the three key functions of journalists as 1) data collection, 2) interpretation, and 3) storytelling. Computers can certainly provide assistance in all of these functions, but in the end, current systems can not generate non-repetitive stories independently. However, technical advances will inevitably lead to computers writing stories independently based on data, without anybody having to tell them explicitly what to write about. The interesting question is what quality level they will achieve within which timeframe. We can easily perceive a future where journalists

input story constraints, provide original background materials like interviews, and automatically get an almost finished article that they just have to polish. For journalists this means they will have to define their profession by the tasks that are fulfilled rather than the persons who possess the skills and knowledge to fulfil them [13], and perhaps shift their focus to tasks where machines do not perform well. For news consumers, the automation opens up the possibility of reading news even on niche topics, but there are scary prospects as well. For instance, individually tailored news with a heavy political agenda can be automatically produced in big volumes.

In this work, we evaluate the Valtteri Natural Language Generation (NLG) system [7], [8], a case study in producing an NLG architecture that reduces the costs of transparent automated news generation in complex domains by employing re-usable and generic components. The Valtteri case study produces news articles about the results of the 2017 municipal elections in Finland, a significant political event in Finland. While the system itself is tri-lingual, producing news in English, Finnish and Swedish, the evaluations were conducted solely on the Finnish output. The Finnish language has significant morphological complexity, which makes it in a sense the ''worst case'' in terms of performance among the languages produced by the case study system.

The main research questions this paper investigates are:

1. What are the user perceptions of the news stories generated by the Valtteri NLG system? (RQ1)
2. Which areas of the automated storytelling need most attention to achieve adequate quality of user perception? (RQ2)

The paper is structured as follows: Section II discusses previous work in the field and describes the perception factors we will evaluate. Section III describes the NLG system design in short, and in Section IV we describe the user evaluation procedures. The results of the evaluations are presented in Section V, with further reflections in Section VI. Limitations are listed in Section VII.

II. BACKGROUND

Users' thoughts on automated journalism have been described in only a few papers up to now. Using Swedish test subjects, Clerwall [3] measured the perception of automatic articles in English, in a setting where the news source (human-written or computer-written) was not declared to the test users. No significant differences in users' perceptions of the texts were found except that the human-written news got more positive ratings for the ''pleasant to read'' descriptor. However, this study did not use the same data for the human versus the machine written news, making the comparisons problematic.

Van der Kaa and Krahmer [15] examined the user perception of computer-written news articles with by-lines potentially manipulated to falsely state a human author, on the dimensions of expertise and trustworthiness. In this evaluation with Dutch content and Dutch speaking respondents, they found no strong differences in perceived expertise nor in trustworthiness, amongst regular news consumers.

Graefe *et al.* [4] built on this and performed evaluations on the impact of the actual and declared source (human-written, computer-written) of the news, on three dimensions: credibility, readability, and journalistic expertise. For this study, they developed a measure for content perception using 12 items and performed tests in an all-German context by varying the actual news source. Their study found that computer-written articles were rated as more credible and higher in terms of expertise than the human-written articles. Regarding readability, human-written articles were rated significantly higher. However, the finding that the declared source has an impact on perception taints these ratings, as computer-written articles were rated substantially higher on readability if declared as written by a journalist.

The impact of declared source appears to vary between countries. Jung *et al.* [5] found that, in South Korea, articles attributed to a computer received higher ratings than those attributed to a human. While South Korea ranked 25th of 26 developed countries in that news trustworthiness survey, Finland stood out as the clear number one, with 65% agreeing with the statement ''you can trust most news most of the time'' [11]. Findings from a mixed European nationality study - although with 3/4 being Germans - by Wölker and Powell [16], showed that credibility perceptions of computer-generated news can be considered equal to human-created content. That study found that for special topics like sports, automatically generated news can even be perceived as more credible than a human-written story.

To our knowledge, there is no literature referencing user perception evaluations for automatically generated textual news content in a Finnish context or the Finnish language. Contrary to the earlier studies about user perceptions, our test also employed a large pool of tested articles.

Our test had the practical goal of using the results to guide ongoing technology development. We wanted to evaluate the general quality of the news stories produced by Valtteri (described more in Section III) and pinpoint areas that needed more development.

In addition, we wanted to take a closer look at the impact of giving users control over the content, since our assumption was that there is great potential in using automated news to serve fringe market segments. The long tail market theory [1] yields that it could be a lucrative business opportunity to automatically produce significant amounts of news items, since even if each individual news items has only a tiny readership, the overall readership over all the news items will be large and the cost of an individual news item is miniscule.

To achieve comparability in the quality evaluation analysis, we opted to use the rating factors from Sundar [12] as the basis of our study. These factors have also served as inspiration to Clerwall [3] as well as for Graefe *et al.* [4]. As the evaluation measures extracted by Sundar [12] were targeted at online and printed news, they are a good fit for evaluating computer-generated content in the written domain.

However, respondent fatigue [2] makes it hard to evaluate all of the 21 Sundar measures successfully when scaling up the number of articles to evaluate. As we set out to evaluate the perceived language generation quality, we did not want to limit ourselves to testing only a few articles, since a particular story topic could result in various biases. We decided to use a quantitative approach of analyzing ten auto-generated articles. Due to the large number of articles, even using 12 items like in [3] and [4], to evaluate each of the articles was deemed excessive.

Thus, the current paper uses a reduced set of measures to evaluate the content quality as efficiently as possible, while staying compatible with earlier research. Sundar [12] includes descriptions of the four main factors found through factor analysis. The factors, each comprising of several measures, were originally described as follows:

1. *Credibility:* The concept of credibility, as applied to a news story, may be defined as a global evaluation of the objectivity of a story.
2. *Liking:* Liking is overall affective reaction. Applied to a news story, liking is an indicator of a news receiver's feelings toward - or - evoked by - the overall content of the news story.
3. *Quality:* Quality means the degree or level of overall excellence of a news story.
4. *Representativeness:* Representativeness of a news story is a summary judgment of the extent to which the story is representative of the category of news. In other words, it is the answer to the following question: What is the probability that the story, taken as a whole, belongs to the class of entities that we call "news"?

These descriptions are rather verbose, which makes them prone to evaluation errors. Spelling out just the associated measures of each factor makes the descriptions unambiguous and more comprehensible for respondents. We therefore listed the evaluation criteria with a set of lead words as follows:

1. Credibility: fair, objective
2. Liking: enjoyable, interesting, lively, pleasing
3. Quality: clear, coherent, comprehensive, concise, well-written
4. Representativeness: important, relevant

Contradictory terms ("biased" from credibility since we can not rate fairness and biasedness with one single number, "boring" from liking, since a combined rating is not appropriate for boring and interesting), and irrelevant ("timely" from representativeness, our sample news are old) terms were excluded. Like all the other material in our online survey, the evaluation criteria were translated to Finnish. Translating the criteria did require slightly longer descriptions, as some English words do not have direct equivalencies in Finnish, and other words could have been confusing in the given context. E.g., the Finnish word for objective ("puolueeton") literally means without any party, a wording that could have created confusion as we were evaluating automated election news - where naturally parties play one of the main roles. Luckily, the main evaluation factors did not face similar problems.

III. VALTTERI NLG ENGINE

Valtteri [7], [8] is a case study of a modular NLG architecture that is largely domain and language independent. In April 2017, the system was set up to produce news articles based on the Finnish municipal elections of 2017. The system took structured results data, released by the Finnish Ministry of Justice as input, and output news articles in three languages: Finnish, Swedish and English. The data included party and candidate results pertaining to the whole country, each of the 13 electoral districts, 311 municipalities, and 2,012 polling stations. The data also contained some candidate backgrounds such as sex, party affiliation, and whether they are currently a member of the European Parliament.

For the generation, Valtteri broadly follows the "classic" pipeline generation process consisting of three stages: content selection, document planning, and surface realization [10]. However, Valtteri expands on this architecture to make most of the components either domain independent, language independent or both, see Fig. 1. The input data is transformed into facts, six-tuples corresponding to the "who, what, where" schema employed often in news.

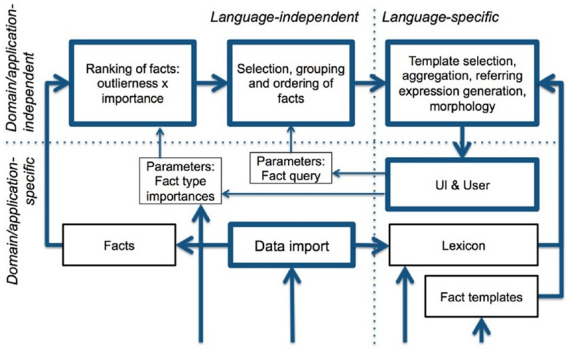

FIGURE 1. Valtteri architecture overview [7]. Thick boxes represent software components and thin boxes data structures.

In the Finnish election domain, entities ("who") can be of two types: party or candidate. Location ("where") has four possible types: the whole country, electoral district, municipality, or polling station. The "what" variables are more varied. In total there are 14 different variables per candidate and 21 per party. These are described in more detail in [7].

Next, each fact is associated with a newsworthiness score. Determination of what is newsworthy is based on the topicality, outlierness, interestingness, and relevancy of the facts to the reader. The document planning stage then builds a document plan, which contains newsworthy facts in a sequence of themed paragraphs. No strict story structure is defined in advance. Each paragraph is created by selecting the most newsworthy unused fact as a nucleus and by adding sufficient supporting facts, with certain additional rules ensuring that each paragraph has an additional theme.

Once the document plan has a sufficient number of paragraphs for an article, it proceeds to the surface realization stage, where it undergoes several transformations. First, each fact is mapped to a phrase-level template. Next, the aggregation component determines whether two or more templates should be aggregated into a longer sentence (e.g., with the use of the conjunction ''and'') or realized as is. The Referring Expression Generation component then determines whether to use full names, short names or pronouns for entities in consecutive sentences in a paragraph. Finally, a realization component turns the transformed document plan into plain text that is displayed to a user. This stage includes the morphological realization, a non-trivial effort in the case of Finnish which has significant morphological complexity. More detailed descriptions of these processes can be found in [7] and [8].

The election Valtteri system was showcased as a website, publicly available at www.vaalibotti.fi. On the website, a user could select the focus of the article, by selecting the location, and optionally the party and or candidate of interest. The system then automatically generates an article based on that user selection. With this tailoring, the election system was able to produce over 700,000 news articles per language, at varying levels of locality – a volume that widely exceeds the human capacity. Users could elect to read about anything from the overall national trends to highly specialized articles detailing a single candidate's success on an individual polling station.

IV. PARTICIPANTS AND PROCEDURES

Doing the tests in Finnish was the most natural choice as the election related news is typically consumed in the native language and the native speakers of Swedish an English speakers in Finland are both minorities. We used the results from the municipal elections 2017 in Finland as data source for the user evaluations. News about election results are commonly understood without respondents needing domain specific knowledge about the topic at hand, and it is a topic of somewhat general interest. However, a downside with election results is that their newsworthiness drops drastically after the first day after publication, so in evaluations it is not feasible to show respondents fresh exciting news. At the time of the evaluation, the news were already a few months old and thus most participants had likely at least some perception of the general nation-wide trends observed in the elections.

We selected ten municipalities to be featured in the evaluation articles. To obtain a ground truth for comparisons we contracted with two professional journalists to write election news for the selected municipalities. The places were chosen so that they were not the home municipality of any participant. The two journalists were provided with the same raw data as Valtteri, along with instructions on how long the news stories should be. The names of the municipalities, parties and candidates were encoded so that the journalists could not use their general knowledge in writing. The project researchers decoded the stories back to real names before using them in the evaluation. Both journalists were assigned with writing five stories each. Having two journalists writing stories ensured we got some diversity in the writing style. It turned out, not only the writing style, but also some viewpoints were slightly different; only the female journalist reported about the gender distribution of the elected councils.

Evaluations were conducted online, using a custom site set up for the evaluations. Prior to the evaluations, users were asked to give demographic background information, as well as tell how much news they read per day and how familiar they are with automated news. Furthermore, they were asked to evaluate the perceived importance for each of the evaluation criteria.

After filling the background survey, the users evaluated ten preselected stories where the source, computer or human, was not declared. Of the preselected stories, six were computer-written and four were human-written. In the 1st, 3rd, 4th, 7th, 8th and 9th slot there was always a computer-written story. From the fixed set of selected places, we then randomized the slot in which a given municipality would appear. Finally, the users were asked to look up and evaluate four articles of their own choice - places, parties, or candidates of special interest to them.

The users were asked to rate each article on credibility, liking, quality and representativeness on a 5-step Likert scale (1 being worst, 5 being best). In addition, they were requested to give written feedback on what was bad and what was good about the story. For the preselected stories, users had to indicate whether they thought the story had been written by computer or by a human.

The evaluation panel was recruited through a commercial partner specializing in online test panel provision, with the aim of having an age and gender diverse group with enough geographical dispersion to eliminate physical location as a cause of bias. This aim was achieved, and in the end, we collected article evaluations from 152 respondents from all around Finland. On average, people spend roughly two and a half minutes reading and answering each story.

V. RESULTS

A. MAIN FACTORS

Out of the four evaluation criteria, the credibility ratings for computer-written preselected articles ranked the highest, followed by representativeness. As can be seen in Table 1, liking got the lowest ratings, which were slightly lower than those for quality.

Compared to the human-written stories, the computer-written stories received statistically significantly lower scores on all criteria. Only the credibility values for preselected computer-written stories are comparable to those for the journalist written one: ratings for the three best-rated computer-generated stories were on the same level as those given to journalist-written stories.

The average ratings between stories differs somewhat, with the best story rated 3.2 and the worst story rated 2.6 (Fig. 2). We can see that the ratings order sequence is the same for all

TABLE 1. Ratings for the evaluations of the main factors, comparing human-written to computer-written articles.

Criteria	Computer	Human	Statistical significance (T-test)
Credibility	$\mu = 3.59$ $\sigma = 1.07$	$\mu = 4.10$ $\sigma = 0.90$	$p < 0.01$
Liking	$\mu = 2.33$ $\sigma = 1.04$	$\mu = 3.98$ $\sigma = 0.96$	$p < 0.01$
Quality	$\mu = 2.58$ $\sigma = 1.13$	$\mu = 3.96$ $\sigma = 0.98$	$p < 0.01$
Representa-tiveness	$\mu = 3.15$ $\sigma = 1.00$	$\mu = 3.96$ $\sigma = 0.91$	$p < 0.01$
Average	$\mu = 2.91$ $\sigma = 0.86$	$\mu = 4.00$ $\sigma = 0.81$	$p < 0.01$

FIGURE 2. Average ratings for all preselected automatically generated news stories ordered according to the overall average rating.

articles, and the difference in average rating between the best and worst article is 0.5 points.

As stated, the 1st, 3rd, 4th, 7th, 8th and 9th articles were always computer-generated stories, selected randomly from a pool of ten. There was no correlation with the order of the article evaluation and the average rating. This supports the notion that the test indeed successfully provided ratings about the writing style of the story, and not the general interestingness of the content. The article evaluation order did, however, have an impact on some of the criteria. For liking, the first ratings were the lowest (Fig. 3), whereas there was a slight downward trend for credibility (not shown). The ratings for quality and representativeness were stable throughout the different positions (not shown).

The users rated the self-selected stories higher than the preselected ones (Fig. 4, Table 2). The difference is statistically significant for liking and quality. The higher ratings support the hypothesis that targeted, more interesting topics have higher end-user value.

To examine the differences between demographic groups we divided the respondents into groups, based on gender and age. For ages, the following ranges were used: young (18-34 years), middle (35-54 years) and old (55-74 years). The differences in ratings are fairly small for the human-written and the preselected computer-written stories, but

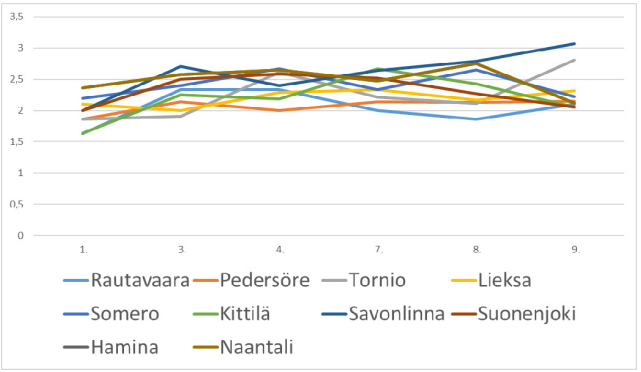

FIGURE 3. The average ratings for liking for the automatically generated articles groups, based on the evaluation order (1st, 3rd, 4th, 7th, 8th and 9th).

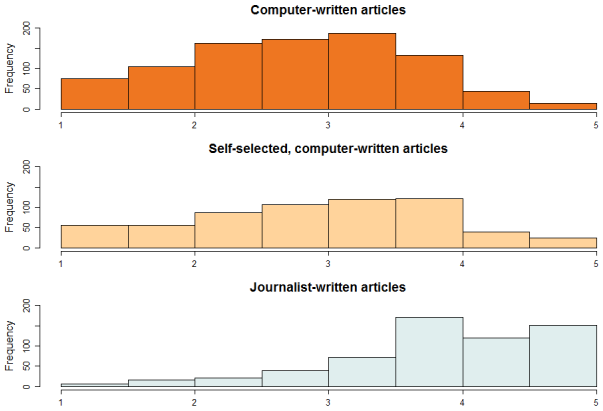

FIGURE 4. Histograms of the average ratings for the articles in the different groups: computer-generated preselected articles (top) self-selected automatically generated articles (middle), and journalist written articles (bottom). The average is the average rating of the four different evaluation criteria.

TABLE 2. Ratings for the evaluations of the main factors, comparing preselected computer-written articles to self-selected automated articles.

Criteria	Preselected	Self-selected	Statistical significance (T-test)
Credibility	$\mu = 3.59$ $\sigma = 1.07$	$\mu = 3.61$ $\sigma = 1.07$	no, $p=0.66$
Liking	$\mu = 2.33$ $\sigma = 1.04$	$\mu = 2.68$ $\sigma = 1.14$	yes, $p < 0.01$
Quality	$\mu = 2.58$ $\sigma = 1.13$	$\mu = 2.75$ $\sigma = 1.20$	yes, $p < 0.01$
Representa-tiveness	$\mu = 3.15$ $\sigma = 1.00$	$\mu = 3.18$ $\sigma = 1.09$	no, $p=0.55$
Average	$\mu = 2.91$ $\sigma = 0.86$	$\mu = 3.06$ $\sigma = 0.96$	yes, $p < 0.05$

clearer for the self-selected automated articles (Fig 5). Young female evaluators gave the lowest average ratings ($\mu = 2.7$, $\sigma = 1.0$), and old male evaluators gave the highest ratings ($\mu = 3.0$, $\sigma = 0.9$). For the preselected computer-written stories, the liking rating of young females ($\mu = 2.0$, $\sigma = 0.9$) differs at statistically significant level from the three

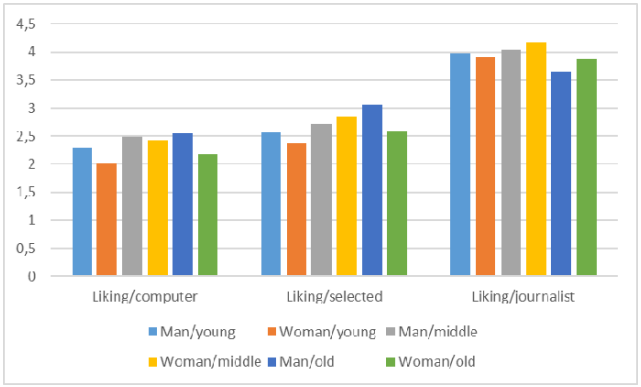

FIGURE 5. The average ratings for liking across demographic groups and different types of articles.

groups giving the highest ratings, namely old males ($\mu = 2.4$, $\sigma = 0.9$, T-test $p < 0.01$), middle-aged males ($\mu = 2.5$, $\sigma = 1.0$, T-test $p < 0.01$), and middle-aged females ($\mu = 2.4$, $\sigma = 1.1$, T-test $p < 0.01$). Old males gave the lowest and middle-aged females gave the highest ratings for the human-written articles.

In the user background details, we had asked about the evaluators' news reading habits and their familiarity with automated news. Based on this background data, we divided the users into four groups: reading a lot and familiar, reading little and familiar, reading little and unfamiliar, and finally, reading a lot and unfamiliar. Users who indicated that they read a lot of news and that they were familiar with automated news gave the highest rating for both preselected and self-selected computer-generated stories ($\mu = 3.0$, $\sigma = 1.1$). The users who read a lot of news but were unfamiliar with automatic news generation gave the lowest ($\mu = 2.8$, $\sigma = 1.0$) ratings for computer-generated stories. The difference between average ratings of the groups were however small (0.13-0.25) for computer-generated stories.

B. FREE TEXT FEEDBACK

The free text feedback regarding the computer-generated content contains some interesting comments, such as accusations of political bias since not all parties were mentioned in the article. Notably the system had no concept of any party being implicitly more "interesting" or newsworthy than the others. Mentioning all facts would have conflicted with the goal of the NLG process defined to select only the most important facts.

On the positive side, there is a trend of praising the fact-basedness and that the story is clear and to-the-point. Further praise was given for the neutral and objective language.

The most frequent complaints were about language errors and deficiencies such as the overuse of the Finnish word for "it". Another main complaint was about repetition and dry language. Some of the repetition is due to the lack of enough alternative phrases to tell the same fact. Some repetition is technically not repetition of data but perceived as repetitiveness by a human: for instance, the computer wrote "the party got second most seats" and then later "the party got second most votes" - these are two different facts, but closely related and experienced by evaluators as repetition. For comparison, the journalist wrote "the party remained second largest and gained two additional seats", avoiding the numerical confusion over seats and votes while at the same time adding contextual details.

We extracted the most common words of the free-text comments, also combining the different forms of the same word and very similar words into one, to get an overview of the things to improve. The top meaningful words in the negative feedback of the automatic stories were:

- boring ("tylsä")
- repetition ("toistoa")
- numbers ("numeroita, lukuja")
- confusing ("sekava")
- listing-like ("luettelomainen")
- stiff ("kankeaa, tönkköä")
- writing mistakes ("kirjoitusvirheitä")
- order ("järjestys")
- monotone ("monotoninen")
- abrupt ("töksähtelevää")
- robot like ("robottimaista")
- annoying ("ärsyttää")
- writing style ("kirjoitusasu")
- grammar mistakes ("kielioppivirheitä")
- dry ("kuiva")
- incoherent ("epäjohdonmukainen")

The same procedure was done for the positive comments resulting in the following top words:

- facts ("faktat")
- clear ("selkeä")
- matters ("asiat")
- numbers ("luvut, numerot")
- informative ("informatiivinen")
- equal ("tasapuolisesti")
- readable ("luettavaa")
- most important ("tärkeimmät")
- structure ("rakenne")
- credible ("uskottava")
- essential ("oleellinen")
- coherent ("johdonmukainen")
- neutral ("neutraali")
- to the point ("ytimekkäästi")
- objective ("objektiivinen")
- comprehensive ("kattava")
- interesting ("mielenkiintoinen")

C. COMPUTER OR JOURNALIST

For each of the preselected articles the users were required to state who they thought had written the story - a human or a computer.

Human-written articles were mistaken for computer-written content in 10.3% of the evaluations. The share of mistakes was fairly evenly distributed across the articles.

The one exception was the Tornio article where only two mistakes were made in total (3.2%).

Automatically generated content was mistaken for human-written articles in 21.1% of the evaluations. Here the mistake distribution is a bit more diverse, but as it happens, here too the Tornio article is an exception with 33.3% mistaking it for human-written content.

We can assume that the higher the share of the mistaken-as-human rate is, the better the general quality of the article is. This assumption is supported by comparing the mistaken-as-human-rate to the mean rating values, for which we get a Pearson product moment correlation coefficient $r = 0.77$. The four articles with least mistakes in identifying them as computer-written are the four articles with lowest average ratings (Fig. 6).

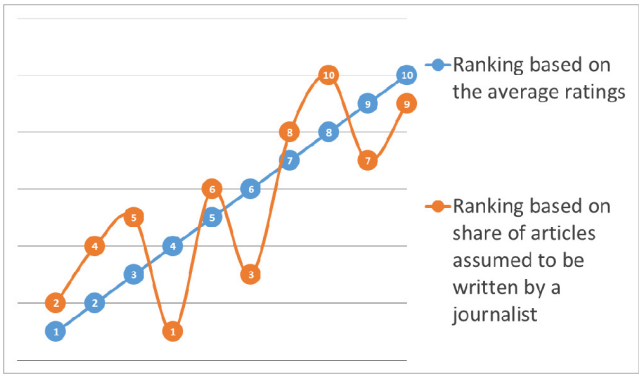

FIGURE 6. The orders of the computer-written news stories based on the average of the four evaluation criteria (blue values) and the share of evaluations where the articles was assumed to be written by a journalist (orange values).

What do these numbers say about the quality of the texts? We were surprised that 10.3% of the human-written stories were mistaken as computer-generated, since they appear to be of good quality - an assessment supported also by the perception ratings. The fact that only 21.1% of the computer-written articles passed as human-written shows that the general quality is not on par with a human journalist.

Looking more closely at the text of the best rated automatic story, about Tornio, we can see that the facts come together nicely and that the flow of the story is good; there were no major language mistakes or confusingly phrased facts. The Lieksa story, correspondingly, which was rated as the worst one, contained some illogical and confusing sentence aggregations. At least to some extent, giving more context with absolute numbers for comparison (saying 4 out of 12 seats, instead of 4 seats) might have made the story easier to understand.

Overall, we can see that a little more than half of the test users made at least one mistake in recognizing the automatically generated stories. Pivoting into age analysis, it can be noted that the share of mistaken test users increased with the age of the test person (Table 3).

We can see in Fig. 7 that the first computer-written article was correctly identified as such by more users than the following articles. We can speculate that the amount of numbers in the articles gave it away, but after they see article number two which is human-written, they realize this type of text will always have much numbers. Then for computer-generated articles number three and four expectations of what a computer-written article should look like have become more blurred - doubling, and then tripling the average error rates compared to the first article. An explanation for the converging curves could be starting point bias, with learning reducing the bias [6]. Starting point bias, or anchoring, refers to the phenomena that when faced with an unfamiliar situation, users make estimates by starting from an initial value and then adjusting the value to yield the final answer. In our case, the final (last) value is around 22.7%, which is close to the total average of 21.1%. For the human-created content, the same number converges at 9.4%, which is close to the 10.3% average.

TABLE 3. The share of test users who made at least one mistake with automatically generated articles taking it for a journalist-written one.

Age \ Gender	Male	Female
18-34 years (young)	16/33 = 48%	9/18 = 50%
35-54 years (middle)	16/29 = 55%	21/35 = 60%
55-74 years (old)	10/13 = 77%	16/24 = 67%
Total	42/75 = 56%	46/77 = 60%

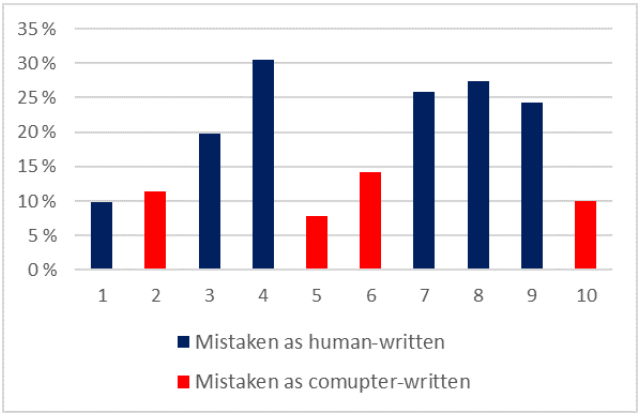

FIGURE 7. Share of misrecognitions by evaluation order. The percentage of times a computer-written story was thought to be a human-written story, and vice versa.

TABLE 4. Share of mistakes by gender. Margin of error at 95% confidence is 1.3% for computer-written stories and 1.5% for human-written stories.

Mistaken as… \ Gender	Male	Female	Avg.
… human-written	18.9%	23.3%	21.1%
… computer-written	11.6%	9.1%	10.3%

Male evaluators were better at recognizing computer-written content (Table 4), but on the other hand, they made

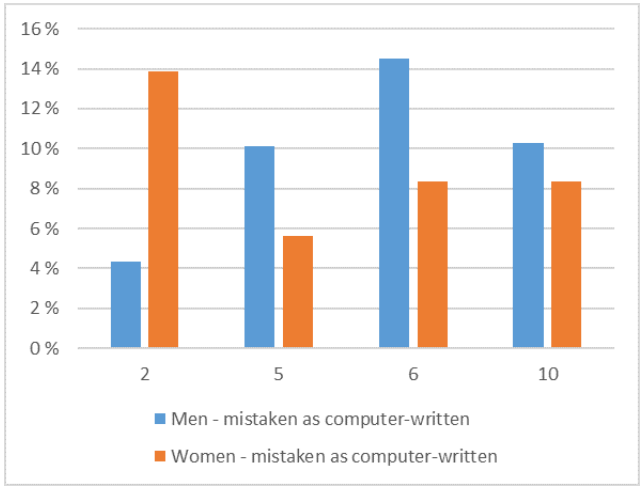

FIGURE 8. Share of misrecognitions (by gender) recognizing a human-written story as computer-written story. The X-axis is story number.

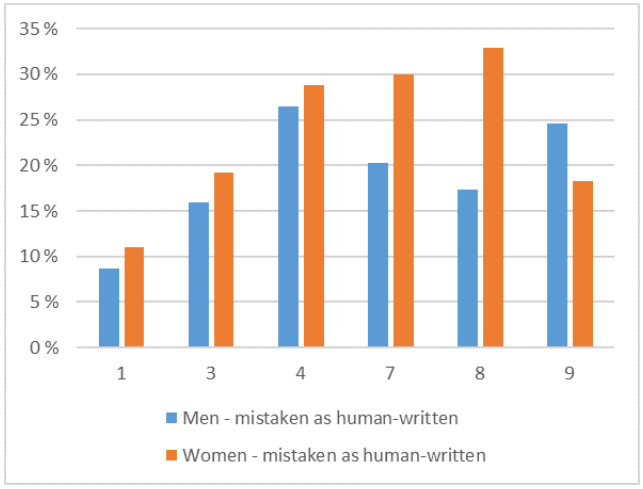

FIGURE 9. Share of misrecognitions (by gender) where users thought a computer-written story was human-written story. The X-axis is story number.

more mistakes recognizing human-written content. Female evaluators were not as quick to recognize the computer's writing patterns, but once they caught on, for the two last stories evaluated they made fewer mistakes than the males (Fig. 8 and Fig. 9). There is no clear reason for this difference, but it is possible that the genders simply had different pre-expectations of what computer-written content would look like. The expectations can have been set by movies and popular literature that target a certain gender.

The reported familiarity with automated news did not seem to play a major role in how well the users recognized, or learned to recognize, automatically produced news (Fig. 10 and Fig. 11).

D. PERSONAL IMPORTANCE

Prior to evaluating the articles, we asked the users to state their perceived importance of each measured factor. This also served as warm-up so that the users would have it easy to

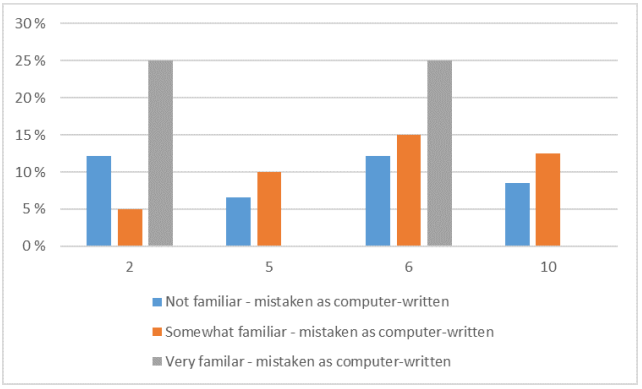

FIGURE 10. Share of misrecognitions where users thought a human-written story was computer-written, per familiarity with automated news. The X-axis is story number.

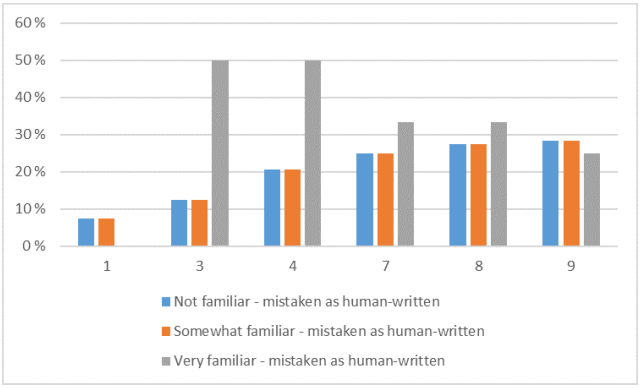

FIGURE 11. Share of misrecognitions where users thought a computer-written story was human-written, per familiarity with automated news. The X-axis is story number.

recognize the evaluation factors from the beginning, when evaluating news articles in the next step.

Users rated the credibility factor highest, with an average of 4.7 ($\sigma = 0.5$) on the 5-point Likert scale. Quality was the second most important factor, averaging 4.6 ($\sigma = 0.6$). Perhaps surprisingly, the liking of the article was stated as the least important factor with a rating of only 3.8 ($\sigma = 0.8$), a bit below representativeness, which obtained an average rating of 4.1 ($\sigma = 0.8$). The order of importance was the same for both genders, but females gave some 0.2 higher importance ratings on average ($\mu = 4.4$, $\sigma = 0.7$ for females vs. $\mu = 4.2$, $\sigma = 0.7$ for males). The pattern does not change when looking at the average and grouping the users by their reading habits. One observation can however be made: all those who indicated not reading a lot of news rated the importance of credibility with the top value of 5.

VI. SUMMARY AND CONCLUSIONS

Conducting user evaluations with 152 participants and analyzing the results gave us clear answers for the Research Question 1 (the user perceptions of Valtteri). While texts written by Valtteri, as expected, got lower ratings than the ones written by journalists, the ratings were still satisfactory

overall: the average was 2.9 vs. 4.0 for journalists on a 5-point scale. Valtteri succeeded best at credibility (3.6 vs. 4.1), where one computer-generated article was rated even higher than the corresponding journalist-written article. Results for representativeness were medium (3.2 vs. 4.0), whereas quality (2.6 vs. 4.0) and liking (2.3 vs. 4.0) were the weakest.

Our findings align with those of Clerwall [3], who found the strengths of the software-generated content to be accuracy, trustworthiness and objectiveness. Also like in our study, Graefe *et al.* [4] found that readability of computer-written articles was poorer than for the articles created by humans, while computer-written articles rated very well for credibility. This echoes the credibility perception findings of Wölker and Powell [16] as well. When evaluating the importance of the factors, the users in our study rated credibility as the most important factor and liking as the least important factor. From that premise, the outlook for computer-written news appears promising.

The news articles that the test users chose by themselves got slightly better ratings than the preselected articles, but the difference was statistically significant only for liking (2.68 vs. 2.33) and quality (2.75 vs. 2.58). This indicates that people are willing to endure some imperfections if the subject is interesting enough and they are part of the dialogue.

We found some differences between age groups regarding liking. Male evaluators aged 55-74 liked the automatic articles best, and female evaluators aged 18-34 liked them the least. One interpretation of this is that young female evaluators preferred more embellished language, whereas older males did not mind the colorless language if they got the facts.

Evaluators thought that 21% of the automatically generated articles were written by humans and mistook 10% of the human written news as automatically generated. The share of users making these mistakes grew with age. Our results show that males were better than females at distinguishing computer-written from human-written content. Considering the facts that the system is in early development and it features an unusually high level of automation, this is an encouraging result, especially as the stories have a writing style that is easily distinguishable once you figure out the pattern. During the tests, the users saw six of the automatic articles close after one another, making it comparatively easy to notice the pattern. This could clearly have skewed the numbers a bit to Valtteri's disadvantage.

Evaluators with the habit of reading a lot of news and familiarity with automatic news generation gave the highest overall ratings. However, users with the habit of reading a lot of news but no familiarity with automatic news generation, gave the lowest ratings. One possible explanation is that the test users with previous familiarity with automatic news generation have a more positive attitude towards automation and understand the challenges of the technology.

In the negative comments of the automatic articles the terms ''boring'', ''incoherent'', ''repetitive'' and ''listing like'' show up - this is mirrored in the lower ratings for liking and quality. Correspondingly, the good ratings in credibility mirror the frequent words of the positive free text feedback: ''clear'', ''facts'', and ''matters''.

Especially the free text feedback provided input to answer the Research Question 2 (what part of automated storytelling needs most attention in future development) - the automatically generated text should be more diverse, and the language more colorful. In addition, the reported numbers must be given with more context to provide tangible information. At least for this context, it is questionable whether automatically generating articles for the end-user to consume is the best way to employ the system. An alternative would be to use the system as a tool for the journalist, spelling out the most interesting facts as text, but allowing the human to provide final modifications. However, when the target audience of a single article is small, the choice is likely between publishing an automatically generated story or no story at all, since it would not be possible to hire enough journalists to write such stories.

Our results indicate that it is feasible to use NLG for generating credible fact-based news stories. In the present state, the resulting text can be successful for highly tailored and targeted use cases where economics makes it impossible to attach human resources to writing. Using it to produce complete ready to publish stories intended for larger audiences requires significant steps to be taken in improving the textual fluency.

VII. LIMITATIONS

The evaluations were not without limitations: the first limitation is that the topic was only news stories about municipality election results. This was an easy to understand topic, motivating its use in the test, but it was not the most exciting topic, particularly because the election had been carried out several months before and the data was not fresh. As the trials were conducted some time after the elections, the stories could have had been seen as more boring than if they were presented immediately after the elections. At the same time, using a large pool of news stories is likely to have removed the possibility of random artefacts in the stories having affected the results.

A second validity limitation is that the setup evaluates only the quality factors for texts produced by the Valtteri system, and it is unclear how representative that output is. Results would be at least slightly different when redoing the evaluations with another system. It is especially noteworthy that the single Valtteri system produces news in three languages compared to most news generation systems employed by news companies that produce news in only a single language. Care should thus be taken when contrasting the results regarding fluency to single-language systems, where much more effort can go to ensuring the fluency in the target language.

Furthermore, the system's architecture emphasizes reusable, generic components. Thus, drawing inferences regarding the performance of bespoke single-context systems should be avoided.

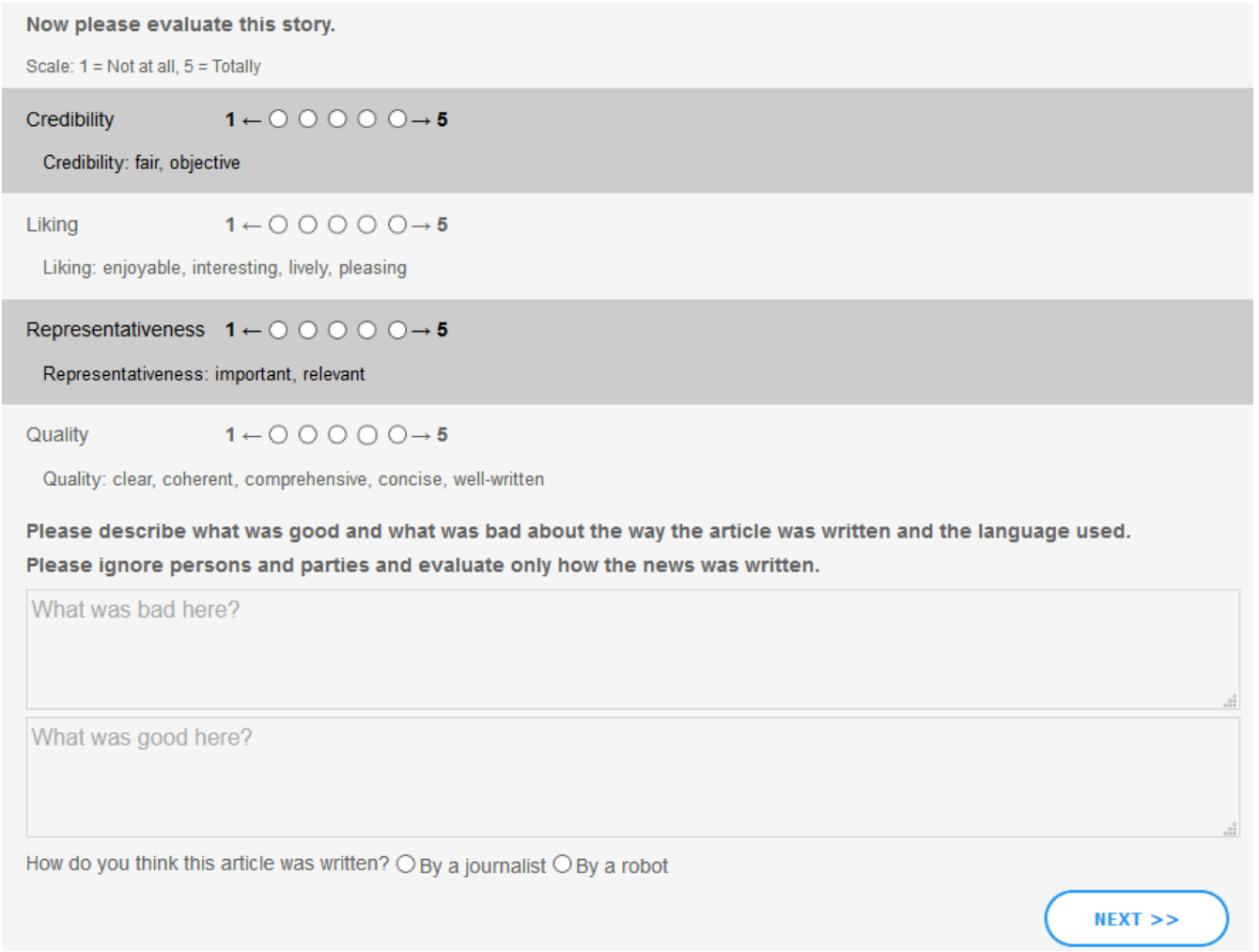

FIGURE 12. Screenshot of the evaluation criteria in the user tests (English version).

In a sense, this trial should be seen as providing a lower bound for the abilities of systems like Valtteri, rather than as an upper bound.

APPENDIX

A. EVALUATION CRITERIA (ENGLISH VERSION)

See Fig. 12.

B. SAMPLE STORY (TORNIO, ENGLISH VERSION)

Most Seats go to the Centre Party of Finland in Tornio:

The Centre Party of Finland is the largest party in the council in Tornio and has 22 seats in the new council. The party received most votes and secured 47.4% of the vote. The party increased their number of seats the most and got 4714 votes.

Katri Kulmuni (cent.) secured 7.9% of the vote and received most votes. She got 784 votes and was elected as a councillor. 648 voted for her in the previous municipal election. She represents The Centre.

The Left Alliance secured 2nd most seats in the new council in Tornio and has 7 seats in the new council. 16.7% of the vote went to the party. The party received the second most votes and had secured 2nd most seats in the previous election. The party took 1663 votes.

3.4% of the vote went to Janne Olsen (sd.). He received the 2nd most votes and got 337 votes. He was elected as a councillor and represents The Social Democratic Party of Finland. Janne Olsen is currently a councillor.

SDP has 6 seats in the new council in Tornio and secured 3rd most seats in the new council. The party secured 15.0% of the vote and got 4.3 percentage points more votes than in the last municipal election. The party secured the second largest increase in council seats and increased their voter support by the greatest margin.

C. SAMPLE STORY (LIEKSA, ENGLISH VERSION)

The Finns Party Drop Most Seats in Lieksa:

The Finns Party dropped the most council seats in Lieksa and lost 3 seats. The party got 8.4 percentage points fewer

votes than in the last municipal election and decreased their voter support by the greatest margin. The party dropped 641 votes since the last municipal election and has 5 seats in the new council.

The Social Democratic Party of Finland is the largest party in the council and has 13 seats in the new council. The party secured 34.8% of the vote and received most votes. The party secured one more seat and got 1946 votes.

The Centre Party of Finland secured 2 more seats in Lieksa and 2nd most seats in the new council. The party has 12 seats in the new council and secured 33.6% of the vote. The party got 7.0 percentage points more votes than in the last municipal election and received the 2nd most votes.

The Left Alliance got the same number of seats and has one seat in the new council. The party secured 5.1% of the vote and 5th most seats in the new council. The party received a council seat in the previous election and got 1.1 percentage points more votes than in the last municipal election.

The National Coalition Party got the same number of seats in Lieksa and 2.1 percentage points fewer votes than in the last municipal election. The party has 3 seats in the new council and dropped 198 votes since the last municipal election. The party lost 2nd most voter support and secured 4th most seats in the new council.

D. HUMAN-WRITTEN STORY (TORNIO, FINNISH VERSION, ENGLISH VERSION NOT AVAILABLE)

Keskustalle Enemmistö Tornion Valtuustoon:

Vaalivoittaja Suomen Keskusta nousi täpärästi yksinvaltiuteen Tornion kaupungissa. Tornio meidän kunta-listan Jari Sainmaa ylsi valtuustoon.

Lähes puolet annetuista äänistä kerännyt Suomen Keskusta saa 22 paikkaa uuden valtuuston kaikkiaan 43 paikasta. Näin Keskusta hallitsee tulevalla kaudella Tornion valtuustoa täpärästi yksinkertaisella enemmistöllä.

Muuten vaalitulos ei tuonut suuria yllätyksiä Tornion kaupungissa. Kärkikolmikon järjestys pysyy samana kuin kuluvallakin kaudella. Toiseksi suurin puolue Vasemmistoliitto menetti yhden paikan ja kolmanneksi suurin Suomen Sosialidemokraattinen Puolue sai yhden paikan lisää. Tästä puolueen käy kiittäminen ennen muuta Janne Olsenia, joka yli nelinkertaisti henkilökohtaisen äänisaaliinsa viime vaalien 65:stä äänestä 337:ään ääneen ja keräsi Tornion toiseksi suurimman äänisaaliin.

Tornion ylivoimainen äänikuningatar oli Keskustan Katri Kulmuni 784 äänellä. Kahden hengen ehdokaslistalta valtuustoon pyrkinyt Jari Sainmaa keräsi 197 ääntä ja nosti Tornio meidän kunta -ryhmän uutena valtuustoon yhden hengen ryhmäksi. Ryhmän toinen ehdokas Marko Koivisto keräsi 35 ääntä ja jäi varalle.

Vaikka kuntakentässä nähtiin paikoin suurtakin myllerrystä, Tornion kaupungissa vanha valta lujittuu entisestään. Suomen Keskusta on periaatteessa valtuuston yksinvaltias, mutta pienimmällä mahdollisella yliotteella. Kolmen suurimman puolueen valta on sen sijaan kyseenalaistamaton. Yhdessä Keskusta, Vasemmistoliitto ja Suomen Sosialidemokraattinen Puolue hallitsevat neljää viidesosaa valtuustosta.

REFERENCES

[1] C. Anderson. (2006). *The Long Tail: Why the Future of Business is Selling Less of More*. [Online]. Available: http://www.wired.com/wired/archive/12.10/tail.html
[2] M. Bradley and A. Daly, "Use of the logit scaling approach to test for rank-order and fatigue effects in stated preference data," *Transportation*, vol. 21, no. 2, pp. 167–184, 1994.
[3] C. Clerwall, "Enter the robot journalist," *Journalism Pract.*, vol. 8, no. 5, pp. 519–531, 2014, doi: 10.1080/17512786.2014.883116.
[4] A. Graefe, M. Haim, B. Haarmann, and H.-B. Brosius, "Readers' perception of computer-generated news: Credibility, expertise, and readability," *Journalism*, vol. 19, no. 5, pp. 595–610, 2016, doi: 10.1177/1464884916641269.
[5] J. Jung, H. Song, Y. Kim, H. Im, and S. Oh, "Intrusion of software robots into journalism," *Comput. Hum. Behav.*, vol. 71, pp. 291–298, Jun. 2017, doi: 10.1016/j.chb.2017.02.022.
[6] J. Ladenburg and S. B. Olsen, "Gender-specific starting point bias in choice experiments: Evidence from an empirical study," *J. Environ. Econ. Manage.*, vol. 56, no. 3, pp. 275–285, 2008.
[7] L. Leppänen, M. Munezero, M. Granroth-Wilding, and H. Toivonen, "Data-driven news generation for automated journalism," in *Proc. 10th Int. Natural Lang. Gener. Conf., Stroudsburg, Assoc. Comput. Linguistics*, 2017, pp. 188–197.
[8] L. Leppänen, M. Munezero, S. Sirén-Heikel, M. Granroth-Wilding, and H. Toivonen, "Finding and expressing news from structured data," in *Proc. 21st Int. Acad. Mindtrek Conf.*, New York, NY, USA, 2017, pp. 174–183.
[9] C.-G. Linden, "Decades of automation in the newsroom: Why are there still so many jobs in journalism?" *Digit. Journalism*, vol. 5, no. 2, pp. 123–140, 2017.
[10] E. Reiter and R. Dale, *Building Natural Language Generation Systems* (Studies in Natural Language Processing). Cambridge, U.K.: Cambridge Univ. Press, 2000.
[11] Reuters Institute. (2016). *Digital News Repot 2016*. [Online]. Available: http://reutersinstitute.politics.ox.ac.uk/sites/default/files/Digital-News-Report-2016.pdf
[12] S. S. Sundar, "Exploring receivers' criteria for perception of print and online news," *Journalism Mass Commun. Quart.*, vol. 76, no. 2, pp. 373–386, 1999, doi: 10.1177/107769909907600213.
[13] A. van Dalen, "The algorithms behind the headlines," *Journalism Pract.*, vol. 6, nos. 5–6, pp. 648–658, 2012, doi: 10.1080/17512786.2012.667268.
[14] B. Van Der Haak, M. Parks, and M. Castells, "The future of journalism: Networked journalism," *Int. J. Commun.*, vol. 6, pp. 2923–2938, 2012. [Online]. Available: http://ijoc.org/index.php/ijoc/article/viewFile/1750/832
[15] H. A. J. van der Kaa and E. J. Krahmer, "Journalist versus news consumer: The perceived credibility of machine written news," in *Proc. Comput. Journalism Conf.*, New York, NY, USA, 2014. [Online]. Available: https://pure.uvt.nl/portal/en/publications/journalist-versus-news-consumer(b36bc9d3-3a56-4ce9-aa2c-3fe726c775a2).html
[16] A. Wölker and T. E. Powell, "Algorithms in the newsroom? News readers' perceived credibility and selection of automated journalism," *Journalism*, to be published. [Online]. Available: http://journals.sagepub.com/doi/full/10.1177/1464884918757072, doi: 10.1177/1464884918757072.

MAGNUS MELIN received the M.Sc. degree in electrical and communications engineering from the Helsinki University of Technology in 2005. He is currently pursuing the D.Sc. degree in telecommunications software with Aalto University, Espoo, Finland.

He is currently a Research Scientist with the VTT Technical Research Centre of Finland Ltd., Espoo. His current research interests include entity extraction and linked data utilization.

ASTA BÄCK received the M.Sc. degree in media technology from the Helsinki University of Technology in 1983.

Since 1983, she has been with the VTT Technical Research Centre of Finland, Ltd., in different positions, such as a research scientist, the team leader, and a project manager, where she is currently a Principal Scientist with the VTT Big Data Industrial Applications Team. Her research has been published in the *Journal of Innovation Management*, the *Journal of Future Studies, Strategic Thinking and Policy*, and the *International Journal of Social and Humanistic Computing*, among others. Her expertise and research interests include media innovation development, utilization of social media to support innovation and marketing, and tools and methods for analyzing social media data.

CAJ SÖDERGÅRD is currently a Research Professor in digital services. He has developed big data methods and applications for over 30 years, starting from image processing mainly for the media industry and more recently for applications within nutrition, environment, learning, and bioeconomy. He has over 250 publications and five patents. After working in industry, he has served in a variety of positions at VTT, including the team leader and the center leader. He is currently on the Board for the European Big Data Value Association and a member with the EU High Level Expert Group on European Open Science Cloud. He leads the Focus Area Big Data in industrial applications at VTT.

MYRIAM D. MUNEZERO received the B.S. degree in computer science and mathematics from the University of Namibia, the M.S. degree in computer science from the University of Joensuu, and the Ph.D. degree in computer science from the University of Eastern Finland. She is currently a Post-Doctoral Researcher with the University of Helsinki.

She has several years of working experience at the Academy of Finland and Business Finland funded projects, in areas ranging from natural language processing and generation to continuous experimentation. Her research in these areas has resulted in several conference papers and journal papers in respectable venues.

LEO J. LEPPÄNEN received the B.A. degree in language technology and the M.Sc. degree in computer science from the University of Helsinki, Helsinki, Finland, in 2015 and 2017, respectively, where he is currently pursuing the Ph.D. degree in computer science.

From 2014 to 2016, he was a Research Assistant with the RAGE Agile Education Research Group, University of Helsinki, where he was a Research Assistant with the Discovery Research Group from 2016 to 2017. Since 2017, he has been a salaried Doctoral Candidate with the Discovery Research Group. He has authored 10 peer-reviewed articles. His research interests are in educational data mining, learning analytics, and data-to-text natural language generation.

HANNU TOIVONEN received the M.Sc. and Ph.D. degrees in computer science from the University of Helsinki, Finland, in 1991 and 1996, respectively.

From 1990 to 1993 and from 1999 to 2002, he was a Researcher and a Principal Scientist at Nokia Research, respectively. From 1993 to 1999, he was a Researcher with the University of Helsinki, where he has been a Full Professor with the Department of Computer Science since 2002. He is involved in the areas of artificial intelligence and data science, especially in computational creativity, self-aware systems, and analysis and generation of natural language. He holds ten patents. According to Google Scholar, he has been cited over 20 000 times and has an h-index of 51.

Prof. Toivonen is a member of the Finnish Academy of Science and Letters and the Finnish Academy of Technology. He served as the Program Chair of the IEEE International Conference on Data Mining in 2014.