Simon Meier-Vieracker*

Automated football match reports as models of textuality

https://doi.org/10.1515/text-2022-0173
Received November 18, 2022; accepted January 11, 2024; published online January 26, 2024

Abstract: This paper deals with automated football match reports as a common genre of automated journalism. Based on a corpus of automated and human-written reports (n = 1,302) on the same set of matches and with reference to linguistic concepts of text and textuality, the textual properties of these texts are analyzed both quantitatively and qualitatively. The analysis is based on the idea that the task of text generation can be described as the task of automatically selecting cues of textuality such as connectives or signals of thematic relatedness. The results show that automated and human-written texts differ significantly in the use of these cues, particularly in the use of linguistic means for creating evaluation and contrast, and thus allow to trace in detail, how these cues contribute to cohesion, coherence and narrative qualities. Different from computational linguistic approaches focused on optimizing text generation algorithms, this paper proposes to use automated texts, which are to some extent imperfect, as models of textuality that through their imperfection can say something about the nature of texts in general. The paper thus contributes to the field of (mostly communication studies) research on automated journalism in which the texts themselves are rarely investigated.

Keywords: automated journalism; text generation; textuality; cohesion; coherence; narrativity

1 Introduction

Among the technological innovations of digital writing in recent years (Lobin 2014: 123–153), the fully automated generation of texts seems to trigger particular fascination as well as suspicion. Apart from small and highly schematic texts as system messages or automated emails as reminder letters, the production of texts seemed to be a human activity for a long time. Recently, however, especially in the journalistic field, technologies and providers emerge that are challenging precisely this certainty.

***Corresponding author: Simon Meier-Vieracker**, Institute of German Studies and Media Cultures, TU Dresden, D-01062 Dresden, Germany, E-mail: simon.meier-vieracker@tu-dresden.de. https://orcid.org/0000-0002-0141-9327

Under terms such as “robot journalism” or “automated journalism”, new technologies are subsumed, advertised, and discussed both in mass media and research, which fully automatically generate and publish journalistic content on the basis of structured data in impressive quality. Even more recently, AI-based text generation technologies that use large language models have made great strides. Especially the so-called transformer architectures like GPT-3 or ChatGPT can generate texts of various kinds and genres (Floridi and Chiriatti 2020; Meier-Vieracker 2024). However, since these models only ‘remix’ the data they were trained on, they cannot yet be used for reporting real world events. Thus, automated journalism is currently making use of template-based software and rule-based algorithms (see Section 2 for details), the products of which are the object of the present paper.

Natural text generation itself is a classic subject of computational linguistics (Gatt and Krahmer 2018), which is often inspired by discourse analytic approaches like *Rhetorical Structure Theory* (Mann and Thompson 1988). Conversely, however, discourse analysis has not yet paid further attention to the phenomenon of automatically generated text. Even in the study of language in the media, which has increasingly dealt with digital genres (Brock et al. 2019; Heyd 2016), interest in automated texts has so far been rather low. Where they are mentioned at all, this is done cursorily and rather in a thesis-like manner (Antos 2017). Empirical analyses of the automated texts themselves are nearly non-existent (de Cesare 2021; Jukneviˇcienė and Viluckas 2019). Even in communication science and journalism studies, where intensive empirical work on automated journalism is done, the texts themselves are hardly ever subjected to a detailed, let alone linguistically founded analysis.

The goal of the present paper is an attempt at filling this gap. Based on a corpus of both automatically generated and human-written football match reports on the same set of events, I aim to analyze the textual properties of these texts. Unlike computational linguistic approaches, however, my analysis is not primarily focused on optimizing technical procedures. I rather start from the observation that automated texts are modelled on their human-written equivalents, so that their analysis allows us to investigate the nature of textuality. The way texts are (re)produced by a machine, and in which aspects this reproduction may remain inadequate, can inform linguists about what makes texts texts. The present paper thus seeks to read automated texts as results of modelling procedures (Scharloth 2016: 318–320) and to use the model character of automated texts for heuristic purposes.

In the following, I will first give an overview of the field of automated journalism and how it is being researched. I will then discuss theoretical concepts of text, texture and textuality that will help to reframe the task of text generation from a linguistic perspective, before presenting the data. Using quantitative, i.e., corpus linguistic methods of key item analysis, as well as qualitative methods, I will then show in

detail that while automated texts exhibit a high degree of connectivity and thematicity, they lag behind human-written texts in their narrative elaboration, particularly through their limitations in creating contrast. Automated texts thus do not exploit fully the typical arrangement of events that creates a plot from a series of events (Gülich and Hausendorf 2000). Finally, I will discuss how these findings can contribute to a better understanding of the nature of texts in general.

2 Literature review

2.1 What is automated journalism?

By definition, automated journalism is a form of automated text generation (also called natural language generation) aimed at the production and publication of journalistic content (Haim and Graefe 2018). The human part in this process is limited to the development of the algorithms (Carlson 2015: 416), after which the process of text production and sometimes even publication is fully automatic without human intervention.

In journalism studies, automated journalism is usually seen as part of a broader trend which is termed computational or algorithmic journalism (Thurman 2019). These terms cover the broad range of the use of algorithms for "the gathering, evaluation, composition, presentation, and distribution of news" (Thurman 2019: 180) and include things as diverse as data-driven graphics (Meier-Vieracker 2020) or personalization of news services based on usage metrics (Tandoc 2014). Although automation plays an important role in all these forms of computational journalism (Diakopoulos 2019), I will focus here on automated journalism in the narrow sense, i.e., the automated generation of journalistic texts.

The standard procedure of automated text generation in journalism is the so called template-based approach building on rule-based algorithms, which runs as follows (Haim and Graefe 2017: 2f., 2018: 150–152). In a first step, the software selects and evaluates structured data against the background of statistical trends based on a set of pre-defined rules. In the second step, texts are generated on the basis of pre-written templates. These templates range from phrasal patterns below the sentence level, which are selected, filled and syntactically adjusted up to textual macro structures. As a result, complete texts are generated, which include prefabricated parts, some randomly selected variants, and varying information taken or derived from the raw data (de Cesare 2021: 88; Diakopoulos 2019: 99).

Applications of such procedures, which in principle have been the same for about 30 years now (Schmitz 1994), are bound to areas where, firstly, enough

structured data are available and, secondly, the composition of texts is so repetitive, formulaic and expectable that even rather strict rule-based approaches will lead to acceptable texts (Haim and Graefe 2017: 3). These include weather reports, (stock) market news and sport reports (see the overview in Dörr 2016). Especially in football, extensive databases are available that make it relatively easy to create match reports that not only reflect the pure facts of the reported match, but also contain statistically derived interpretations and evaluations, such as noting a team's "strong initial phase". These evaluations might be the reason why in journalism studies automated journalism is described as the generation of "narrative news texts" (Carlson 2015: 416; see also Diakopoulos 2019: 137), even though, from a linguistic point of view and as will be shown below, the narratives they present are limited to sequencing reported events rather than actually building a plot.

Numerous experimental reception studies in the field of communication and journalism studies have investigated if and to what extent automated texts are recognized as such by human readers (Graefe et al. 2018; also cf. the works cited in Thurman 2019: 185). Also, readers' judgements of these texts according to approved categories of journalistic text quality like credibility have been investigated. Most studies show that in experimental settings automated texts are not or hardly distinguishable from human-written texts. The evaluations of readability, credibility and journalistic expertise do not differ significantly, either (Graefe et al. 2018; Lermann Henestrosa et al. 2023).

More fine-grained studies have suggested that automated texts are perceived as more objective and informative, but also as more boring, while human-written texts were judged as more pleasant to read (Clerwall 2014). However, the observed differences are hardly ever related to the concrete texts, and which textual elements might have led to the respective judgements is not further investigated. By the way, and this is rather irritating from a linguistic point of view, the texts used as stimuli in the experiments are usually not reported in the studies.

Moreover, there are studies on the reception of and opinions about automated journalism as a whole by professional journalists (Kunert 2020; Thurman et al. 2017). Many journalists emphasize the benefits of opinion-based and stylistically more individual texts by human writers compared to the most objective descriptions in automated texts (van Dalen 2012: 652). In these studies, however, the reference to texts (both automated and human-written) is all the more lacking, which could be used to show what might be a linguistic manifestation of, for example, individual style. It therefore seems instructive to subject the texts themselves to a detailed linguistic analysis. But, before doing so, we need a better theoretical understanding of text, texture and textuality.

2.2 Texts, texture and (cues of) textuality

Any task of developing an algorithm for automated text generation draws on a theoretical concept of what constitutes a text, although this may not be explicitly reflected by the developers. An everyday notion of text that could serve as a reference point here aims at the production of grammatically well-formed sentences that are then combined into a larger but clearly delineated unit. Beyond mere concatenation, however, a text represents a coherent set of sentences that as a whole fulfills a recognizable communicative function (Brinker et al. 2018) – in our case the report of a football match.

In linguistics, this notion of text has been elaborated into a detailed text theory. In their groundbreaking work "Cohesion in English", Halliday and Hasan (1976: 1) defined the term *text* as referring to "any passage, spoken or written, of whatever length, that does form a unified whole". The basic unifying relation is referred to as cohesion, which can be further divided into grammatical cohesion and lexical cohesion (Halliday and Hasan 1976: 6). Linguistic means used as cohesive ties include pro-forms, syntactic constructions such as parallelisms, and connectives on the side of grammatical cohesion. Lexical cohesion can be established through semantic relations such as hyponymy as well as through expanding or condensing paraphrases (Schubert 2017: 319). While sentences are internally marked by structure, the patterns of cohesive ties, both grammatical and lexical ones, constitute the "texture" (Halliday and Hasan 1976: 2), which is defined as "the property of 'being a text'" and of being distinguished from non-texts.

In a similar approach, De Beaugrande and Dressler (1981) in their foundation of text linguistics introduced the term *textuality* in order to capture constitutive properties of texts. Expanding the lexico-grammatical approach of Halliday and Hasan (1976), textuality is further differentiated into seven "standards of textuality" (De Beaugrande and Dressler 1981: 3). *Cohesion* as the set of mostly grammatical relations at the text surface is complemented by *coherence* as a continuity of sense building on (possibly implicit) conceptual relations like causality or temporal sequence used for explanations, narrations, etc. Moreover, De Beaugrande and Dressler (1981) determine *intentionality*, *acceptability*, *situationality*, *informativity* and *intertextuality* as standards of textuality. While intertextuality as the connectedness with other texts is rather intuitive, the other standards are more difficult to comprehend, as they seem to mix text structural properties on the one hand and producer- and recipient-sided properties as well as situational-contextual aspects on the other (Hausendorf et al. 2017: 2).

For this reason, Hausendorf et al. (2017: 8f.) have proposed to reframe the concept of textuality as a heuristic concept aimed at investigating linguistic means in

texts which *signal* or *indicate* their textuality to their recipients. These means can be regarded as *sources* and *cues* of textuality, i.e., as features of texts which will lead the recipients to perceive them as texts. Specifically, Hausendorf et al. (2017) identify the following classes of sources of textuality associated with textuality cues in the texts: *Delimitability* concerns the various means for drawing the boundaries of a text, e.g., by headlines. *Connectivity* compromises the whole range of cohesive ties in the sense of Halliday and Hasan (1976), while *thematic relatedness* covers thematic coherence as well as patterns of thematic unfolding like narrative, explanatory or argumentative sequences. *Pragmatic utility* covers various features that signal the possible functions and uses of the text. Finally, *intertextuality* concerns the links to other texts through references, quotes and so on, while *patternedness* concerns patterns on different levels that a text shares with other instances of the same genre.

This approach provides a suitable theoretical basis for the research question on the textual properties of automated texts in comparison to human-written ones for three reasons. First, it allows to capture the textual features of (automated as well as human-written) football match reports in a consistent theoretical framework: match reports are delimited, grammatically cohesive and thematically coherent sequences of sentences, which are used to inform about the course and outcome of a football match. They may contain intertextual references to other texts and they are highly patterned texts both through a highly standardized text structure and through genre-specific formulaic sequences (Meier 2019). Second, the approach is particularly well applicable to the task of automated text generation. In the terms of this approach, this task can now be described as the task of automatic selection and use of suitable cues of textuality. This process works without intentions and understanding on the part of the text generating machine and therefore has to be boiled down to the processing of a sophisticated, but static set of rules which nevertheless succeeds in emulating textuality. Third, the focus on cues of textuality fits with the analytic interest identified above as a research gap, i.e., the investigation of textual elements which guide recipients' interpretations and judgements on the texts.

3 Corpus data

In order to enable a precise analysis of automated texts, a large corpus was built. It contains both automated and human-written football match reports on the two highest German football leagues *Erste* and *Zweite Bundesliga.* The automated reports are produced by the Berlin based provider *Retresco.* The company has developed a template-based text generation software which is used for a variety of domains

including sports reports. In the case of football reports, the software generates texts based on match and season data by selecting, filling, and combining predefined templates. According to the company, these templates were mostly developed manually together with professional sports journalists, e.g., by identifying typical formulations for certain constellations and contexts.

To give an impression, a complete automated match report on the match 1 FC Köln against Hertha BSC (15.08.2021) shall be quoted here:

> ***Doppelpack: Kainz sichert 1. FC Köln den Sieg***
>
> *Zum Auftakt in die neue Spielzeit kam der 1. FC Köln gegen die Hertha BSC zu einem 3:1.*
>
> *Stevan Jovetic traf nach einer Vorlage von Matheus Cunha per Rechtsschuss zur 1:0-Führung für die Hertha (6.). Als manch einer bereits mit den Gedanken in der Halbzeitpause war, besorgte Anthony Modeste auf Seiten des FC das 1:1 (41.). Zum Seitenwechsel hatte keine Mannschaft die Oberhand gewonnen. Unentschieden lautete der Zwischenstand. Mit einem schnellen Doppelpack (52./55.) zum 3:1 schockte Florian Kainz den BSC. Mit dem Abpfiff durch den Schiedsrichter war dem 1. FC Köln der Start ins neue Fußballjahr geglückt. Gegen die Hertha BSC fuhr der FC einen 3:1-Sieg auf eigenem Platz ein.*
>
> *Als Nächstes steht für die Mannschaft von Steffen Baumgart eine Auswärtsaufgabe an. Am Sonntag (17:30 Uhr) geht es gegen den FC Bayern München. Die Hertha tritt bereits einen Tag vorher gegen den VfL Wolfsburg an.*

> **'Double: Kainz secures victory for 1. FC Köln**
>
> At the start of the new season, 1. FC Köln won 3:1 against Hertha BSC.
>
> Stevan Jovetic scored with a right-footed shot after an assist from Matheus Cunha to give Hertha a 1-0 lead (6th). When some were already thinking about the half-time break, Anthony Modeste scored for FC to make it 1-1 (41st). At the change of ends, neither team had gained the upper hand. The intermediate score was a draw. Florian Kainz shocked BSC with a quick double (52./55.) to make it 3:1. When the referee blew the final whistle, 1. FC Köln had a successful start to the new football year. Against Hertha BSC, the FC scored a 3:1 victory on their own pitch.
>
> Next up for Steffen Baumgart's team is an away task. On Sunday (5:30 p.m.), they will face FC Bayern München. Hertha will take on VfL Wolfsburg the day before.'

Retresco's automated football match reports have been published for example in the online edition of the newspaper *Die Welt*.[1] Moreover, a demo version of the software is freely available on the company's website, with which users can generate any number of reports on the current *Bundesliga* match day.[2] According to the company, the algorithm underlying the demo software is the same as the one used for the

1 https://www.welt.de/sport/fussball/bundesliga/article234111500/RB-Leipzig-VfL-Bochum-Bochum-kommt-nicht-aus-dem-Keller-Bundesliga.html.

2 https://www.retresco.de/branchen/medien.

Table 1: Corpus composition and size[a].

	kicker.de		**sky.de**		**retresco.de**		***Sum***	
	Texts	Tokens	Texts	Tokens	Texts	Tokens	Texts	Tokens
2018_19	135	91,792	133	66,523	135	61,417	*403*	*219,732*
2021_22	306	182,134	287	107,498	306	86,311	*899*	*375,943*
Sum	*441*	*273,926*	*420*	*174,021*	*441*	*147,728*	*1,302*	*595,675*

[a]The number of texts in the sky corpus is slightly smaller because match reports were not published for all matches.

delivery of texts for publication on *welt.de*. Thus, the demo software can be used for corpus construction purposes.

Texts from two periods were included in the corpus of the present study. First, I included one match report each on 135 matches (15 match days) of the second round of the German Bundesliga 2018/19 (January – May 2019). Second, I included 103 match reports each from the first leg of the first and second Bundesliga 2021/22.

For comparative purposes, match reports written by humans on the same set of matches were collected as published on the websites *kicker.de* and *sky.de*. In that manner, a corpus was built that contains automated match reports that can be exactly matched with two human-written counterparts each. The corpus was part-of-speech tagged and lemmatized with the software TreeTagger (Schmid 2003). Moreover, a sentence splitter (Whitener 2017) was used. The resulting corpus contains 1,302 texts with 595,675 tokens. Table 1 shows the composition and size of the corpus:

All data and the Python code for replicating the quantitative analyses presented in the paper can be found in the repository.[3]

4 Results

4.1 Quantitative results

Before going into detail, some global observations on the texts will be helpful. Figure 1 shows the average text lengths and sentence lengths grouped by the three sources.

Figure 1 indicates that automated texts are shorter and consist of shorter sentences. This may be a first hint to a greater complexity of the human-written texts. In addition, the size of the boxes shows that there is less scatter in the dataset of automated texts, which points to the rule-based generation of the texts. The same holds true for the average number of paragraphs as shown in Figure 2.

3 https://osf.io/8bn6p/.

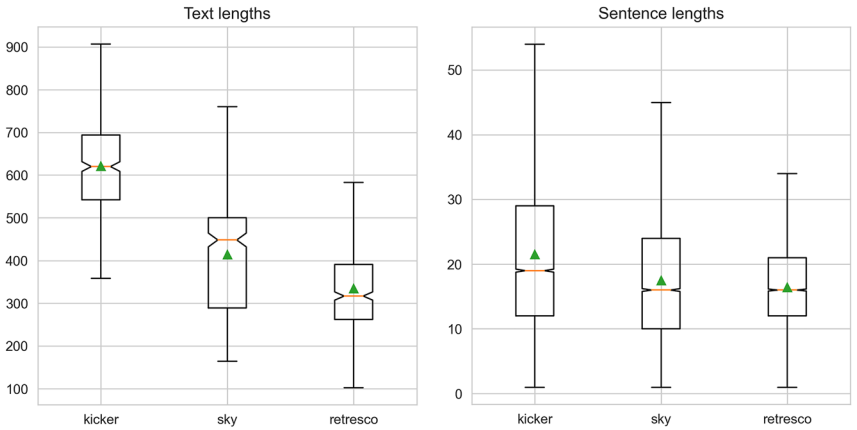

Figure 1: Text lengths and sentence lengths.

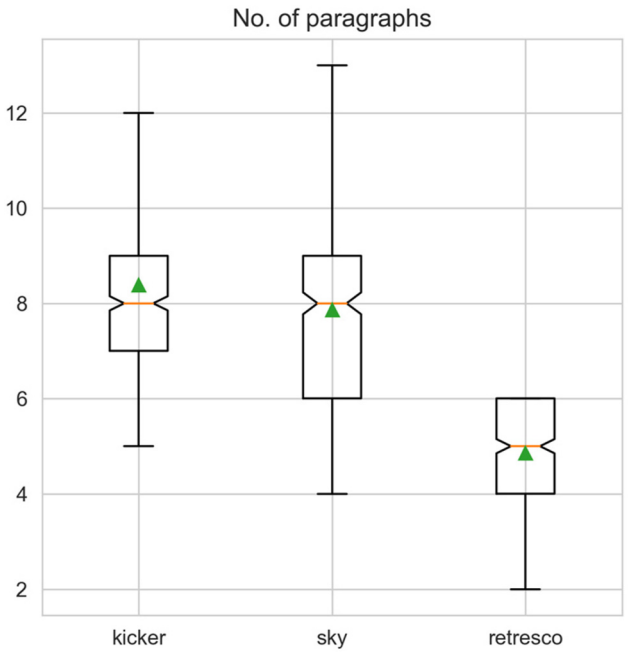

Figure 2: No. of paragraphs.

The fact that the human texts are subdivided more finely may seem like a purely layout-related detail. But assuming that paragraphs are delimitable units of meaning, often with typographically marked subheadings, this finding suggests that

human writers are able to subdivide the thematic macrostructure of the text into smaller parts. In contrast, the algorithm only distinguishes between the main parts of the report like initial situation, course of the game and outlook for the coming matches, and it does not set any subheadings. Thus, the textuality of delimitability is better exploited in human-written texts through "structuring signals" (Hausendorf et al. 2017: 150) such as subheadings. Moreover, the automated texts show a very static kind of patternedness through a most formulaic and recurrent structure of the texts.

4.2 Corpus linguistic results: Lexico-grammatical cues of connectivity and thematic relatedness

The results reported so far remain rather abstract and do not take into account the lexico-grammatical features of the texts. Therefore, I will now focus on more linguistically substantial categories, i.e., parts of speech and lemmas. I will continue to adopt a quantitative perspective and conduct a contrastive frequency analysis called keyword analysis. Keywords (or key items) "are those whose frequency (or infrequency) in a text or corpus is statistically significant, when compared to the standards set by a reference corpus" (Bondi 2010: 3). I will use the automated text corpus as the target corpus, which will be compared to the human-written texts using the Wilcoxon rank-sum test, which has been shown to be more adequate than the widely used keyword metric Log Likelihood Ratio as it also takes the dispersion of keywords into account (Sönning 2023). As I will show, the automated and human-written texts differ significantly in how cues of connectivity (Hausendorf et al. 2017: 161) and cues of thematic relatedness are being used.

4.2.1 Key parts of speech

The results of a key part of speech[4] analysis is shown in Figure 3. All items displayed were found to be statistically significant with a significance threshold of $p < 0.05$. However, the bars show the difference in relative frequencies to make the actual difference easier to estimate. The left side shows the parts of speech used more frequently in the automated texts and the right side the parts of speech used less frequently.

Some results are expectable given the operating principle of the text generation algorithm. On the one hand, NE (proper names) and CARD (cardinal numbers) are

4 The part of speech tags follow the Stuttgart Tübingen Tagset (STTS), https://www.ims.uni-stuttgart.de/forschung/ressourcen/lexika/germantagsets/.

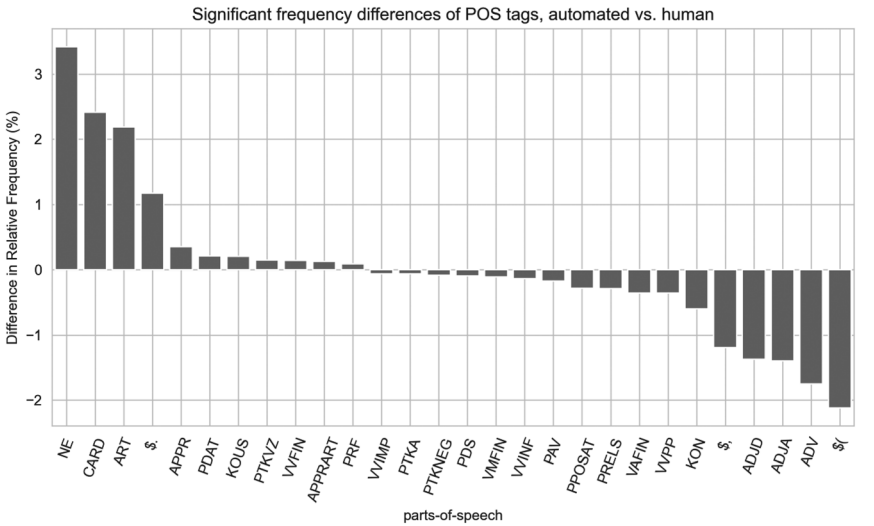

Figure 3: Significant frequency differences of POS tags.

massively overrepresented in the automated texts, since these items represent the information gathered in the databases by the algorithm. Even if there are little or no embellishing passages, the hard facts of the game still remain to be reported. On the other hand, ADJA (attributive adjectives as in *schöne Flanke* 'beautiful cross') and ADJD (adverbial or predicative adjectives as in *perfekt zurückgelegt* 'perfectly laid off') are strongly underrepresented in the automated texts. Such evaluative embellishments, which cannot be derived from the raw data alone, remain the preserve of human writers.

PDAT (attributive demonstrative pronouns) are overrepresented in the automated texts. From a text-analytical perspective, this is an interesting finding. Demonstrative pronouns are often used as coreferential, anaphoric means which refer back to a referent already introduced in or derivable from prior discourse (Diessel 1999: 6; Schwarz-Friesel and Consten 2011: 357), even across sentence boundaries. Thus, they can be regarded as a cohesive tie par excellence. Example (1) contains a complex anaphora *diese Niederlage* ('this defeat'), the referent of which is the whole event expressed in the preceding sentence *musste … Punkte abgeben* ('lost the points'):

(1) *Gegen Bayer 04 Leverkusen musste man zum zweiten Mal in Folge die Punkte abgeben. Durch diese Niederlage fiel Wolfsburg in der Tabelle auf Platz acht zurück.* (retresco)
'Against Bayer 04 Leverkusen, Wolfsburg lost the points for the second time in a row. This defeat dropped Wolfsburg down to eighth place in the table.'

A look at the corpus data, however, shows that the actual complex-anaphoric use of demonstratives is restricted to a small range of expressions like *diese Niederlage* ('this defeat'), *dieser Sieg* ('this win') and *dieses Remis* ('this draw'). Different from prototypical complex anaphora, they reformulate and condense the preceding proposition, but do not add much information or evaluation (Schwarz-Friesel and Consten 2011: 358). Also, the entire cross-sentence construction seems to be part of a pre-established pattern which is selected as a whole.

While attributive demonstrative pronouns are overrepresented in the automated texts, substituting demonstratives pronouns (PDS) are underrepresented. In particular, the sentence-initial demonstrative pronoun *das* ('that') distinguishes human-written from automated texts. This pronoun is used as a complex anaphora, too. Its antecedent (Schwarz-Friesel and Consten 2011: 355) is a whole proposition rather than a single expression, and their referents are complex bundles of events, as in (2).

(2) *Danach machte der Club ein wenig auf, ohne zunächst gefährlich zu werden. Das aber änderte sich in der 73. Minute als Behrens nach einem Pereira-Abschluss im Nachschuss aus kurzer Distanz am Tor vorbeizielte.* (kicker)
'After that, the club opened up a bit without becoming dangerous at first. But that changed in the 73rd minute when Behrens missed the goal from close range after a Pereira shot.'

As can be concluded from this example, the complex-anaphoric use of *das* presupposes a conceptual representation of a complex state of affairs, which then can be referred to as a whole. This seems to go beyond the capabilities of the text generation algorithm. In most cases of substituting demonstrative pronouns in the automated texts, they refer back to a concrete noun phrase as in (3):

(3) *Am Zwischenstand änderte sich weiter nichts, sodass dieser auch gleichzeitig der Pausenstand war.* (retresco)
'The intermediate score remained unchanged, so that this was also the score at the break.'

Again, this construction turns out to be a pre-fixed pattern, which the algorithm seems to select as a whole instead of linking smaller components into a larger unit as human writers would do.

Among the parts of speech overrepresented in the automated texts, we also find subordinating conjunctions (KOUS), which connect two propositions and are thus prototypical cohesive ties, too (Stede 2018: 27). Table 2 shows the five most frequent conjunctions tagged as KOUS in automated and human-written texts.

The most frequent conjunction in the automated texts is the conjunction *als* ('when'), which usually establishes a temporal relation of coincidence between two

Table 2: Subordinating conjunctions.

automated (*n* = 973; 6,586 pMW)		**human-written (*n* = 2034; 4,541 pMW)**	
als ('when')	36 %	*dass* ('that')	16 %
sodass ('so that')	16 %	*weil* ('because')	13 %
während ('while')	14 %	*ehe* ('before')	12 %
dass ('that')	12 %	*als* ('as' or 'when')	11 %
da ('because')	5 %	*während* ('while')	8 %

events (Breindl et al. 2014: 300). However, the corpus data show that it is used in a quite particular way in the automated texts as in (4):

(4) *Der FC Bayern München verpasste den Ausgleich, als ein Kopfball von James Rodriguez das Tor verfehlte (73.). Eine gute Chance für den FC Bayern vergab Lewandowski, als sein Kopfball das Tor verfehlte (77.).* (retresco)
'FC Bayern München missed the equaliser when a header by James Rodriguez missed the goal (73.). A good chance for FC Bayern was missed by Lewandowski when his header missed the goal (77.).'

In this example, the conjunction *als* connects two propositions that describe the same event in two different ways, which therefore are necessarily coincident. In a way, such a construction allows to build a syntactically complex sentence, which, however, does not correspond to a semantically complex proposition. In contrast, the same scenes are described in the human-written text as follows (5):

(5) *Weiser blockte einen Kopfball von James aus vier Metern (73.), Lewandowski nickte vorbei (77.).* (kicker)
'Weiser blocked a header by James from four metres (73.), Lewandowski nodded past (77.).'

The event is reported in a much more condensed manner without unfolding it into syntactically complex sentences. In the human-written texts, we find uses of *als* that actually establish a more substantial temporal relation between two events as in (6).

(6) *Die reguläre Spielzeit war längst abgelaufen, als die Frankfurter noch einen Schreckmoment überstehen mussten.* (kicker)
'Regular playing time had long since expired when the Frankfurt team had to survive a moment of shock.'

Here, the regular time being expired sets the ground on which the actual event, the moment of shock, appears. This pattern, prototypical for narratives (Gülich and Hausendorf 2000: 374), seems to go beyond what the algorithm's set of rules can achieve.

The second most frequent conjunction in the human written text is *weil* ('because'), which points to the fundamental conceptual relation of causality as a basis for text coherence. Match reports must not only report the mere facts, but also provide explanations for what happened on the pitch (Schütte 2006). The construction of cause-effect relations is therefore crucial. Typical examples from a human-written texts are (7) and (8):

(7) *In der Schlussphase wurde es noch einmal turbulent, weil beide Teams den Sieg wollten.* (kicker)
'In the final phase, things got turbulent again because both teams wanted the victory.'

(8) *Augsburg hielt nun besser mit, auch weil Werder zeitweise den Fuß vom Gaspedal nahm.* (sky)
'Augsburg now kept up better, also because Werder slowed down at times.'

In both examples, the facts reported in the subordinate clauses provide causal explanations of what happened on the pitch. In the automated texts, however, *weil* is used only 11 times and makes up only 1 % of the subordinating conjunctions. Comparable with *als* discussed above, *weil* is used in a peculiar way as in (9):

(9) *Die verbleibende Zeit der ersten Halbzeit blieb ohne weitere Treffer, weil die Chancen durch Fink, Gießelmann und Usami ohne Erfolg blieben.* (retresco)
'The remaining time of the first half remained without further goals because chances by Fink, Gießelmann and Usami were unsuccessful.'

In this example, the facts reported in the subordinate clauses do not provide a causal explanation in the proper sense. They rather give a further specification or elaboration of the fact reported in the main clause by increasing its level of detail (Scheffler and Stede 2016). The two propositions connected by the conjunction seem to describe the same fact, just in different granularity. The syntactic complexity and the conceptual relation evoked by *weil* does not adequately correspond to a causal relation on the propositional level.

Heavily underrepresented in automated texts are adverbs (ADV). Again, not only the frequency of adverbs in general differs between automated and human-written texts, but also which adverbs are used. Table 3 shows the five most frequent adverbs in both sub-corpora.

A most noticeable difference can be observed with *aber* ('but'), which is – in its adverbial sense[5] – used more than twenty times as often in human-written texts

5 *Aber* can also be used as a conjunction as in *7 Niederlagen, aber 15 Siege* ('7 defeats but 15 victories'). The relative frequencies of this type of *aber* (according to the part-of-speech tagging) is nearly balanced between automated and human written texts (426 pMW vs. 609 pMW).

Table 3: Adverbs.

automated (*n* = 5,538; 37,488 pMW)		human-written (*n* = 24,647; 55,022 pMW)	
nur ('only')	12 %	*auch* ('also')	8 %
noch ('still)	6 %	*aber* ('but')	7 %
nun ('now)	5 %	*noch* ('still')	6 %
bisher ('so far')	5 %	*nur* ('only')	5 %
auch ('also)	5 %	*wieder* ('again')	4 %

(4056 pMW vs. 176 pMW) (cf. Juknevičienė and Viluckas 2019: 68f. for a similar observation on generated football match reports in a video game). As an adversative connective, *aber* signals or even builds up a contrast between two propositions (Breindl et al. 2014: 516). As the examples (10) and (11), taken from human-written texts, show, *aber* can indicate a contrast between certain circumstances on the one hand and an actual event on the other:

(10) *Der FSV dominierte, lief aber in einen Konter.* (kicker)
'The FSV dominated but ran into a counterattack.'

(11) *Die Eintracht schien in dieser Phase deutlich unterlegen, schwamm sich aber in der Schlussphase des ersten Durchgangs noch einmal frei.* (kicker)
'Eintracht seemed to be clearly outclassed at this stage, but broke free once again in the final phase of the first half.'

In both examples, expectations derived from apparent dominance or inferiority are thwarted by the events introduced with *aber*. This contrast is further emphasized by the connective *aber*, which thus takes a concessive interpretation (Blakemore 1989; Stede 2004: 276).[6] It goes with the conditional presupposition that a dominant team will not concede a goal under normal conditions as in (10) (Breindl et al. 2014: 520) so that the counterattack comes as a surprise. In (11), the verb *scheinen* ('seem') indicates the mere appearance of being outclassed, but still *aber* marks the team's breaking free as a non-expected event. As both examples show, the connective *aber* plays an important role in dramatic embellishments, since it allows to highlight or even to construct unexpected turning points as indispensable features of tellable stories (Baroni 2014). All this seems to presuppose abstract and at the same time evaluative representations of game situations which cannot be derived directly from the raw data alone and thus remain a preserve of human writers. Accordingly, the

6 As a test, *aber* with a concessive interpretation can be paraphrased with *obwohl* ('although') or the like (Stede 2004: 277): "Although FSV dominated, they ran into counterattack."

overall few uses of adverbial *aber* in the automated texts fall into a small set of pre-fixed patterns, which the template-based algorithm selects as a whole according to quantifiable rules as in (12).

(12) *Mit bisher nur vier Treffern zeigte der Sturm des 1. FC Nürnberg erhebliche Defizite. Dafür stand die Defensive aber ziemlich sicher.* (retresco)
'With only four goals so far, the offense of 1. FC Nürnberg showed considerable deficits. But the defense was pretty solid.'

It is revealing that the passages with *aber* in the automated texts do not describe dynamic game scenes, but relations as they can be derived from the mere numbers of goal ratios.

4.2.2 Key lemmas

The observation that human writers seem to have other options for creating contrasts and staging moments of surprise in their narrative representations of the games can also be further supported by keyword analysis at the lemma level. Results are shown in Figure 4. Again, all lemmas displayed are highly significant with $p < 0.001$, but the bars show the difference of relative frequencies. Only lemmas overrepresented in human-written texts are shown here.

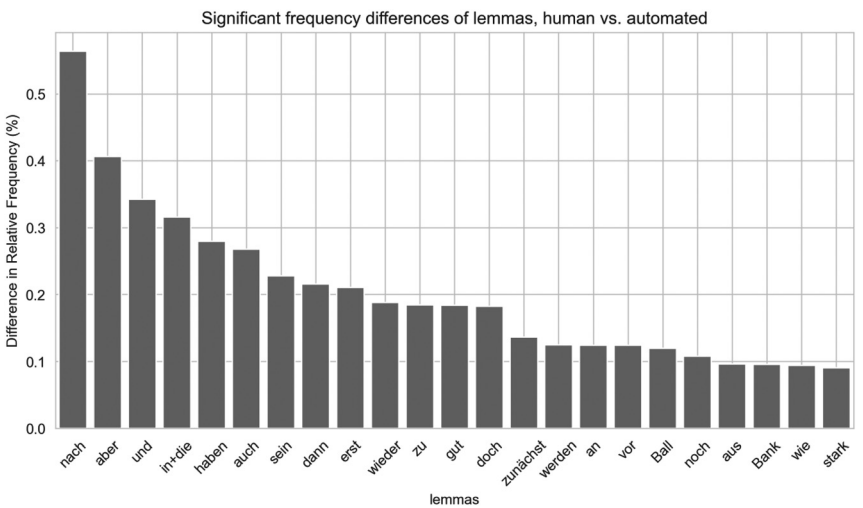

Figure 4: Significant frequency differences of lemmas, human versus automated.

It shows again that adversative connectives such as *aber* and *doch* (both meaning 'but') are typical for the human-written texts. The temporal adverb *zunächst* ('initially') also functions as a connective, since it necessarily requires a continuation and thus indicates a chronological sequence of events. Moreover, the corpus data show that it is typically combined with adversative connectives as in (13):

(13) *Die Borussen hatten zunächst alles im Griff und sahen nach einer 3:0-Führung wie der sichere Sieger aus. Aber die TSG glaubte immer an sich, spielte mutig nach vorne und erzielte ab der 75. Minute durch eine große Moral noch drei Treffer.* (kicker)
'Borussia initially had everything under control and looked like a sure winner after taking a 3-0 lead. But TSG always believed in itself, played courageously forward and scored three more goals from the 75th minute onwards thanks to great morale.'

Similarly, the temporal adverb *dann* ('then') is often combined with *aber* as in (14):

(14) *Die Münchner gingen früh in Front, wurden dann aber zu passiv und kassierten den Ausgleich.* (kicker)
'Munich took an early lead, but then became too passive and conceded the equalizer.'

Such representations of a chronological sequence of events that highlight a surprising change of state seem to be a fundamental narrative principle in human-written match reports. Linguistic constructions of contrast, e.g., by adversative conjunctions and adverbs are a simple but effectful means for this. However, they presuppose conceptual representations of events that can be meaningfully contrasted in the first place, and this goes beyond the capacity of a text generation algorithm.

The same applies to the adverb *auch* ('also'), overrepresented in human-written texts. It requires two comparable events, states or the like, which can then be connected. In (15), *auch* ties the proposition of Stindl's miss back to the Swiss player's miss mentioned before.

(15) *Der 22-jährige Schweizer zögerte allerdings zu lange, Verteidiger Joe Scally (18) sprintete rechtzeitig zur Hilfe. Die Gladbacher dominierten weiter, doch auch ein Kopfball von Lars Stindl (67.) aus kurzer Distanz ging knapp über das Tor.* (sky)
'However, the 22-year-old Swiss hesitated too long, defender Joe Scally (18) sprinted to the rescue in time. Gladbach continued to dominate, but a header by Lars Stindl (67) from close range also went just over the goal.'

In the automated texts, on the contrary, the use of *auch* is limited to a small set of prefixed patterns like *womit es auch weiterhin beim Gleichstand blieb* (literally, 'which kept the score tied also further on').

To sum up, the key item analyses presented in this section show that both the automated and the human-written texts make use of a wide set of textuality cues. These range from grammatical cohesive ties like conjunctions, adverbial connectives and even (complex) anaphora, which establish a high grade of connectivity in the texts, to thematic relations of contrast, temporal sequence or causality. However, a closer look shows that the actual use of these cues differs greatly between the two sorts of texts. In the automated texts, they are often part of prefixed patterns and seem to work without prior conceptual representations, which then would have to be connected in the text production process. In the human-written texts, on the contrary, these textuality cues are used more flexibly, especially for the purpose of narrative elaboration. If the task is not only to describe the course of the game but to present a suspenseful narrative, human writers are clearly superior.

4.3 Qualitative analysis: chains of thematic development

In order to further enrich the corpus linguistic results, a look at a larger snippet is helpful. The following passage (16) is taken from an automated match report:

(16) *Das Match war erst wenige Momente alt, als vor 34.394 Zuschauern bereits der erste Treffer fiel. Yussuf Poulsen war es, der in der zweiten Minute zur Stelle war. RB Leipzig verpasste den Ausbau der Führung, als der Keeper von Fortuna Düsseldorf einen Schuss des 24-jährigen Stürmers entschärfte (2.). Bereits in der neunten Minute baute Ibrahima Konate den Vorsprung von Leipzig aus, nachdem Marcel Halstenberg vorgelegt hatte. Eine Parade nach einem Schuss von Timo Werner verhinderte den nächsten Treffer des Gastes (12.). Das letzte Tor der turbulenten Startphase markierte Poulsen in der 16. Minute nach einer Vorlage von Konrad Laimer.*
'The match was only a few moments old when the first goal was scored in front of 34,394 spectators. It was Yussuf Poulsen who was on target in the second minute. RB Leipzig missed out on extending their lead when the Fortuna Düsseldorf keeper saved a shot from the 24-year-old striker (2'). Ibrahima Konate extended Leipzig's lead as early as the ninth minute after Marcel Halstenberg had laid on. A save after a shot from Timo Werner prevented the next goal from the guest (12'). The last goal of the turbulent opening phase was scored by Poulsen in the 16th minute after an assist from Konrad Laimer.'

Apart from the grammatical cohesive ties discussed above, this snippet shows various means that further establish thematic coherence. For example, the *first goal* reported in the first sentence is taken up and specified by the phrase *to be on target* in the second sentence, just as the vague time information *only a few moments old* is further specified by *in the second minute*. The phrases *extension of lead* and *the next*

goal subtly build on aforementioned information, and the fact that *Leipzig missed out* is specified by the information that the *Düsseldorf keeper saved a shot*. The antonomasia *24-year-old striker*, which reformulates the proper name *Yussuf Poulsen*, instantiates a figure typical of the genre of sports reporting (Burkhardt 2006: 63) and also establishes connectivity while preventing repetitiveness. The phrase *last goal of the turbulent start phase* finally bundles all the events reported so far into an evaluative account. The algorithm is obviously capable of weighting individual pieces of information against the background of certain statistical regularities in such a way that a certain exceptionality of the event is made clear. The phrase *as early as the 9th minute* also points in this direction. All this forms the basis for a coherent thematic development of the report, which goes beyond a mere listing of single events and has narrative qualities.

However, a comparison of this text with its human-written counterpart shows that they differ precisely in their narrative qualities. The report of *kicker.de* reads as follows (17):

(17) *Die Leipziger erwischten einen absoluten Traumstart in die Partie – allerdings auch unter gütiger Mithilfe von Düsseldorfs Keeper Rensing. Nachdem Poulsen zunächst noch an Rensing gescheitert war, kam er kurze Zeit und einige Kopfballduelle später erneut zum Abschluss und markierte diesmal die frühe Führung für die Sachsen (2.). […] Innenverteidiger Konaté schaltete sich in die Angriffsbemühungen ein und erwischte die Düsseldorfer Defensive damit völlig auf dem falschen Fuß. Der Franzose zog an Ayhan, Usami und Morales vorbei und stolperte den Ball dann mit etwas Glück und Unterstützung des Innenpfostens an Rensing vorbei ins Tor – das erste Bundesligator des Abwehrmanns (9.). Werner scheiterte mit seinem Flachschuss noch an Rensing (12.), kurze Zeit später schnürte Poulsen dann nach einem schnellen Angriff über Halstenberg und Laimer den Doppelpack und stellte früh im Spiel bereits auf 0:3 (16.).*
'Leipzig got off to an absolute dream start in the match – albeit with the kind assistance of Düsseldorf keeper Rensing. After Poulsen initially failed to beat Rensing, he finished again a short time and a few aerial duels later, this time giving the Saxons an early lead (2.). […] Central defender Konaté joined in the attacking efforts and caught the Düsseldorf defence completely on the wrong foot. The Frenchman moved past Ayhan, Usami and Morales and then, with a bit of luck and the support of the inside post, stumbled the ball past Rensing into the goal – the defender's first Bundesliga goal (9th). Werner's low shot still failed to beat Rensing (12.), but a short time later Poulsen scored a brace after a quick attack via Halstenberg and Laimer to make it 0:3 early in the game (16.).'

While there are some parallels like the antonomasia *Konaté* > *Frenchman* > *defender* and the emphasis on the extraordinariness of the *0:3 early in the game*, the differences are even more noticeable. First, this report is much richer in evaluative descriptions and sayings like *absolute dream start* or *caught completely on the wrong foot*. Second, the accounts of single events are much more detailed like *stumbled the ball into the goal with the support of the inside post* – information which is not delivered through the match data the algorithm relies on. Finally, a variety of adverbs already discussed in Section 4.2.1 creates additional weightings and contrasts. Twice a figure is used which binds two consecutive events together and creates a contrast between them. Poulsen *initially failed*, then *finished again, this time giving an early lead.* Similarly, Werner *still failed to beat Rensing, but a short time later Poulsen then scored.* In both cases, the success of the goal is described in the light of the previous failure, giving them additional meaning and presenting them as surprising events. In this manner, human writers can structure their narratives even more variably and suspensefully.

5 Discussion and conclusion: automated texts as models of textuality

As argued in Section 2.2, the task of automatic text generation can be described as the task of automatic selection and use of suitable cues of textuality in the sense of Hausendorf et al. (2017), which will lead the recipients to accept the sequences of sentences as coherent texts. Beyond grammatical well-formedness, the sentences must be connected by cohesive ties, which are further enriched by adequate conceptual and thematic relations. While other features of textuality such as delimitability or patternedness are relatively easy to implement due to the schematic structure of football match reports, both cues of connectivity and thematic relatedness pose a major challenge for rule-based algorithms.

The contrastive corpus linguistic analyses of the automated and human-written texts have shown significant differences in the use of these cues. Apart from the greater detail with which human writers can describe the course of the game compared to algorithms that only have basal match data, the automated texts seem to differ from human-written ones in degree and quality of connectivity and thematic relatedness. On the one hand, a broad range of cohesive ties like connective conjunctions and adverbs as well as anaphora is used in the automated texts, allowing for syntactically complex and varied sentences as well as sentence sequences. On the other hand, their use remains somewhat static, since they are part of prefixed patterns as defined in the underlying templates, and they are usually not additionally

underpinned by substantial thematic relations. The analysis has shown, for example, that causal explanations are common in the human-written texts while the automated texts do not seem to provide causal explanations even when causal constructions are used in the grammatical sense. Using the example of 'but', it was also shown that the automated texts make less use of constructions of contrast. Although there are chains of thematic development in the automated texts, which to a certain extent allow for evaluative statements, human writers still show more variable ways of highlighting unexpected events and presenting narrative rather than purely descriptive accounts of the game. In particular, by arranging events into a plot through the construction of contrasts, human writers are better able to exploit the narrative potential of game reports.

All this shows that while the text generation algorithm relies on match data, it still does not conceptually represent the events of the game which it expresses and connects linguistically. As such, this is not surprising, since computers do not think. It is nevertheless revealing to see how a rule-based text generation algorithm can emulate (to a certain degree, successfully) textual connectivity and thematic relatedness through the selection and combination of predefined patterns.

One limitation of this study is due to the corpus linguistic approach, which neglects the multimodal properties of match reports often containing additional statistics and images that contribute to the overall meaning of the text. To capture these properties, a more fine-grained annotation would be needed, which could also trace the thematic progression within the texts in greater detail. In further research, the text-analytical results of this work could also be fed back into the reception studies common in communication studies, e.g., by investigating the effects of different degrees and qualities of connectivity on readers' interpretations and evaluations of the texts.

In the context of this paper, however, a different path shall be proposed. Since rule-based algorithms obviously do not plan and produce texts like humans do and emulate textuality rather than producing it, automated text generation can be described as text modelling (in the sense of dialogue modelling, cf. Jokinen 2009) and automated texts as models of textuality. As Scharloth (2016: 318–320) argues in his discussion of the relevance of models in interactional linguistics, models as purposeful constructions are simplified representations of what they attempt to model. Details that are irrelevant or too complex for the purpose of the model can be omitted as long as the model fulfils its function. What is decisive here is that these models can serve as heuristic tools, since they tell us something about the modelled objects, even if the modelling itself is incomplete and not entirely adequate.

In this way, automated texts – which are to some extent imperfect – can give us valuable information about the nature of texts, be they texts of the specific genre of football match reports or texts in general. Especially in contrast with their human-

written equivalents, they show how grammatical cohesion needs to be underpinned by thematic coherence. They also show that the respective conceptual relations (e.g., of contrast) cannot be derived by the raw data the algorithm relies on but must first be constructed. Finally, they show through which linguistic means descriptive accounts still can acquire narrative qualities. As long as the algorithmic models fail to implement these highly complex features of texts, automated texts will remain *mere* models of textuality, and the cultural technique of text production (Lobin 2014) will not be completely transferred to machines. The future will show whether the advanced AI language models, which are expected to be further developed to be able to deal with real-time data, will be better able to compete with humans in the task of writing.

Acknowledgements: The author would like to thank the company Retresco for providing the texts of the demo software for research purposes. The author has no conflicts of interest to declare.

References

Antos, Gerd. 2017. Wenn Roboter „mitreden". Brauchen wir eine Disruptions-Forschung in der Linguistik? *Zeitschrift für Germanistische Linguistik* 45(3). 392–418.

Baroni, Raphaël. 2014. Tellability. In Peter Hühn, Jan Christoph Meister, John Pier & Wolf Schmid (eds.), *Handbook of narratology*, 836–845. Berlin & Boston: De Gruyter.

Blakemore, Diane. 1989. Denial and contrast: A relevance theoretic analysis of "but". *Linguistics and Philosophy* 12(1). 15–37.

Bondi, Marina. 2010. Perspectives on keywords and keyness: An introduction. In Marina Bondi & Mike Scott (eds.), *Keyness in texts*, 1–18. Amsterdam: Benjamins.

Breindl, Eva, Volodina Anna & Ulrich Hermann Waßner. 2014. *Handbuch der deutschen Konnektoren 2: Semantik der deutschen Satzverknüpfer*. Berlin, München, Boston: De Gruyter.

Brinker, Klaus, Cölfen Hermann & Steffen Pappert. 2018. *Linguistische Textanalyse. Eine Einführung in Grundbegriffe und Methoden*. 9., durchgesehene Auflage. Berlin: Erich Schmidt Verlag.

Brock, Alexander, Pflaeging Jana & Peter Schildhauer (eds.). 2019. *Genre emergence: Developments in print, TV and digital media*. Berlin & New York: Lang.

Burkhardt, Armin. 2006. Sprache und Fußball. Linguistische Annäherung an ein Massenphänomen. *Muttersprache* 2006(1). 53–73.

Carlson, Matt. 2015. The Robotic Reporter: Automated journalism and the redefinition of labor, compositional forms, and journalistic authority. *Digital Journalism* 3(3). 416–431.

Clerwall, Christer. 2014. Enter the robot journalist. *Journalism Practice* 8(5). 519–531.

de Cesare, Anna-Maria. 2021. Répétitions et variations des textes générés: Une analyse linguistique basée sur un corpus d'articles financiers rédigés en français. *Chimera: Revista de Corpus de Lenguas Romances y Estudios Lingüísticos* 8. 79–108.

De Beaugrande, Robert & Wolfgang U. Dressler. 1981. *Introduction to text linguistics*. London & New York: Longman.

Diakopoulos, Nicholas. 2019. *Automating the news: How algorithms are rewriting the media*. New York: Harvard University Press.

Diessel, Holger. 1999. *Demonstratives: Form, function and grammaticalization*. Amsterdam: Benjamins.
Dörr, Konstantin Nicholas. 2016. Mapping the field of algorithmic journalism. *Digital Journalism* 4(6). 700–722.
Floridi, Luciano & Massimo Chiriatti. 2020. GPT-3: Its nature, scope, limits, and consequences. *Minds and Machines* 30(4). 681–694.
Gatt, Albert & Emiel Krahmer. 2018. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research* 61(1). 65–170.
Graefe, Andreas, Mario Haim, Bastian Haarmann & Hans-Bernd Brosius. 2018. Readers' perception of computer-generated news: Credibility, expertise, and readability. *Journalism* 19(5). 595–610.
Gülich, Elisabeth & Heiko Hausendorf. 2000. Vertextungsmuster narration. In Klaus Brinker, Gerd Antos, Wolfgang Heinemann & Sven F. Sager (eds.), *Text- und Gesprächslinguistik*, vol. 1, 369–385. Berlin & New York: De Gruyter.
Haim, Mario & Andreas Graefe. 2017. Automated news: Better than expected? *Digital Journalism* 5(8). 1044–1059.
Haim, Mario & Andreas Graefe. 2018. Automatisierter Journalismus. In Christian Nuernbergk & Christoph Neuberger (eds.), *Journalismus im Internet: Profession – Partizipation – Technisierung*, 139–160. Wiesbaden: Springer Fachmedien.
Halliday, Michael A. K. & Ruqaiya Hasan. 1976. *Cohesion in English*. London: Taylor & Francis.
Hausendorf, Heiko, Wolfgang Kesselheim, Hiloko Kato & Martina Breitholz. 2017. *Textkommunikation: Ein textlinguistischer Neuansatz zur Theorie und Empirie der Kommunikation mit und durch Schrift*. Berlin & Boston: de Gruyter.
Heyd, Theresa. 2016. Digital genres and processes of remediation. In Alexandra Georgakopoulou & Tereza Spilioti (eds.), *The Routledge handbook of language and digital communication*, 87–102. London & New York: Routledge.
Jokinen, Kristiina. 2009. *Constructive dialogue modelling: Speech interaction and rational agents*. Chichester: Wiley.
Juknevičienė, Rita & Paulius Viluckas. 2019. Lexical features of football reports: Computer- vs. human-mediated language. In Marcus Callies & Magnus Levin (eds.), *Corpus approaches to the language of sports: Texts, media, modalities*. London: Bloomsbury.
Kunert, Jessica. 2020. Automation in sports reporting: Strategies of data providers, software providers, and media outlets. *Media and Communication* 8(3). 5–15.
Lermann Henestrosa, Angelica, Hannah Greving & Joachim Kimmerle. 2023. Automated journalism: The effects of AI authorship and evaluative information on the perception of a science journalism article. *Computers in Human Behavior* 138. 107445.
Lobin, Henning. 2014. *Engelbarts Traum: Wie der Computer uns Lesen und Schreiben abnimmt*. Frankfurt a.M.: Campus.
Meier, Simon. 2019. Formulaic language and text routines in football live text commentaries and match reports – a cross- and corpus-linguistic approach. In Marcus Callies & Magnus Levin (eds.), *Corpus approaches to the language of sport*. Texts, media, modalities, 13–35. London: Bloomsbury.
Meier-Vieracker, Simon. 2020. Die Verdatung des Fußballs. Spuren von Algorithmen in der Fußballberichterstattung. *Muttersprache* 130(4/2020). 304–318.
Meier-Vieracker, Simon. 2024. Uncreative Academic Writing: Sprachtheoretische Überlegungen zu Künstlicher Intelligenz in der akademischen Textproduktion. In Gerhard Schreiber & Lukas Ohly (eds.), *KI:Text. Diskurse über KI-Textgeneratoren*, 133–144. Berlin & Boston: De Gruyter.
Mann, William C. & Sandra A. Thompson. 1988. Rhetorical Structure Theory: Toward a functional theory of text organization. *Text – Interdisciplinary Journal for the Study of Discourse* 8(3). 243–281.

Scharloth, Joachim. 2016. Praktiken modellieren: Dialogmodellierung als Methode der Interaktionalen Linguistik. In Arnulf Deppermann, Helmuth Feilke & Angelika Linke (eds.), *Sprachliche und kommunikative Praktiken*, 311–336. Berlin & Boston: De Gruyter.

Scheffler, Tatjana & Manfred Stede. 2016. Adding semantic relations to a large-coverage connective lexicon of German. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 1008–1013. Portorož, Slovenia: European Language Resources Association (ELRA). Available at: https://aclanthology.org/L16-1160.

Schmid, Helmut. 2003. Probabilistic part-of-speech tagging using decision trees. In D. B. Jones & H. Somers (eds.), *New methods in language processing*, 154–164. London: Routledge.

Schmitz, Ulrich. 1994. Automatic generation of texts without using cognitive models: Television news. In Susan Hockey & Nancy Ide (eds.), *Research in humanities computing 2*, 186–192. Oxford: Clarendon.

Schubert, Christoph. 2017. Discourse and cohesion. In Christian Hoffmann & Wolfram Bublitz (eds.), *Pragmatics of social media*, 317–343. Berlin & Boston: De Gruyter.

Schütte, Christian. 2006. *Matchwinner und Pechvögel: Ergebniserklärung in der Fussballberichterstattung in Hörfunk, Internet, Fernsehen und Printmedien*. Münster: Lit.

Schwarz-Friesel, Monika & Manfred Consten. 2011. Reference and anaphora. In Wolfram Bublitz & Neal R. Norrick (eds.), *Foundations of pragmatics*, 347–372. Berlin & New York: De Gruyter Mouton.

Sönning, Lukas. 2024. Evaluation of keyness metrics: Performance and reliability. *Corpus Linguistics and Linguistic Theory* 20(2). 263–288.

Stede, Manfred. 2004. Kontrast im Diskurs. In Hardarik Blühdorn, Eva Breindl & Ulrich H. Waßner (eds.), *Brücken schlagen. Grundlagen der Konnektorensemantik*, 255–286. Berlin & Boston: De Gruyter.

Stede, Manfred. 2018. *Korpusgestützte Textanalyse: Grundzüge der Ebenen-orientierten Textlinguistik*. 2. Aufl. Narr: Tübingen.

Tandoc, Edson C. 2014. Journalism is twerking? How web analytics is changing the process of gatekeeping. *New Media & Society* 16(4). 559–575.

Thurman, Neil. 2019. Computational journalism. In Karin Wahl-Jorgensen & Thomas Hanitzsch (eds.), *The handbook of journalism studies*, 2nd edn., 180–195. London: Routledge.

Thurman, Neil, Konstantin Dörr & Jessica Kunert. 2017. When reporters get hands-on with robo-writing: Professionals consider automated journalism's capabilities and consequences. *Digital Journalism* 5(10). 1240–1259.

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.

Whitener, Chase. 2017. Lingua::Sentence. Perl. Available at: https://metacpan.org/pod/Lingua::Sentence.

Bionote

Simon Meier-Vieracker
Institute of German Studies and Media Cultures, TU Dresden, Dresden, Germany
simon.meier-vieracker@tu-dresden.de
https://orcid.org/0000-0002-0141-9327

Simon Meier-Vieracker is Professor of Applied Linguistics at TU Dresden, Germany, since 2020. He received his PhD in German Linguistics at the University of Bern, Switzerland, in 2012. He is doing research in the field of discourse analysis, media linguistics and corpus linguistics.