Simon Meier-Vieracker

# Von Rohdaten zum Text – Themenentfaltung in automatisierten Fußballspielberichten

Meier-Vieracker, Simon. 2023. Von Rohdaten zum Text – Themenentfaltung in automatisierten Fußballspielberichten. In ThemaTalkers (Julian Engelken | Marc Glund | Jan Hensellek | Lara C. Herford | Saskia Langrock | Sargis Poghosyan | Susanne S. Schmalwieser | Ingo H. Warnke) (eds.), Was ist eigentlich ein Thema? Sieben linguistische Perspektiven, 52–58. OpenAccess U Bremen. https://doi.org/10.26092/elib/2313.

## 1 Einleitung

Vor etwa zehn Jahren trat eine Reihe von Softwareunternehmen im Bereich der Textgenerierung mit einem großen Versprechen an die Öffentlichkeit. *Narrative Science* bot und bietet bis heute Lösungen an, um vollautomatisiert die »stories hidden in your data« zu erzählen, und *Automated Insights* verspricht, mit Technologien der Natural Language Generation Daten in »clear, human-sounding narratives« transformieren zu können. Anwendungsgebiete für solche Softwarelösungen sind Wirtschaftsthemen wie etwa Börsenberichte, Produktbeschreibungen sowie Sportberichte (vgl. Diakopoulos 2019). Überall dort, wo ausreichend Daten vorliegen und Texte nach einem vergleichsweise standardisierten Muster produziert werden müssen, kommen automatisiert generierte Texte zum Einsatz. Ob aber diese Texte tatsächlich als narrative Texte gelten können, bleibt fraglich. Die Textlinguistik mit ihren Begriffen des Vertextungsmusters der Narration oder der narrativen Themenentfaltung bietet hierbei einen geeigneten Rahmen, um diese Frage linguistisch fundiert angehen zu können.

Empirischer Gegenstand meines Beitrags sind automatisierte Fußballspielberichte, die mit einem, vom Berliner Unternehmen *Retresco* entwickelten Algorithmus generiert wurden. Dieser beruht auf dem sog. template-basierten Ansatz, bei dem regelbasiert manuell vorformulierte Formulierungsschablonen auf der Grundlage von strukturierten Daten ausgewählt, gefüllt und zu Texten kombiniert werden (vgl. Haim & Graefe 2018). Die entsprechenden Texte werden tatsächlich von verschiedenen Online-Medien wie etwa *welt.de* publiziert; auf der Webseite des Unternehmens war zudem bis vor kurzem eine Demoversion verfügbar, mit der zu den Spielen des jeweils aktuellen Bundesligaspieltags Berichte in beliebiger Zahl generiert werden konnten. Aus textlinguistischer Sicht sind diese Texte ein interessanter Gegenstand, da sich ihnen die Herausforderung stellt, aus den die reinen Fakten zum Spielverlauf repräsentierenden Rohdaten kohäsive und kohärente Texte zu produzieren, in denen das Thema der Texte – der Verlauf eines Fußballspiels – schrittweise entfaltet wird. Zu fragen ist deshalb, mit welchen sprachlichen Mitteln dies geschieht und ob und wie sich ihr Gebrauch von dem in menschlich verfassten Texten unterscheidet. Dazu wurde ein Parallelkorpus mit jeweils 440 Texten der Anbieter

*kicker.de*, *sky.de* und eben *Retresco* zu dem gleichen Set an Bundesligaspielen aus den Jahren 2019–2021 mit einer Gesamtgröße von knapp 600.000 Tokens erstellt. In diesem Korpus lassen sich also sehr präzise menschliche und automatisierte Vertextungen miteinander vergleichen, sowohl auf Einzeltextebene als auch in einem quantitativen Zugriff auf Korpusebene.

## 2 Theoretische Grundlagen: Textualität und Thematizität

Die hier gestellte Frage nach Vertextungen greift einen Klassiker der Textlinguistik auf, nämlich die Frage nach Textualität als dem Bündel jener Merkmale, die Texte von bloßen Aneinanderreihungen von Sätzen unterscheiden. Beaugrande und Dressler (1981) haben in diesem Zusammenhang die berühmten Textualitätskriterien wie etwa Kohäsion, Kohärenz, Informativität und Intertextualität definiert. In einem neueren Ansatz haben Hausendorf und Kesselheim (2008) Abgrenzbarkeit, Kohäsion/Konnektivität, Thematizität, Funktionalität, Musterhaftigkeit und Intertextualität als konstitutive Eigenschaften von Texten bestimmt. Dieser theoretische Ansatz ist sowohl für die automatisierte Generierung von Texten als auch für deren Analyse einschlägig, da nun gefragt werden kann, wie sich diese Eigenschaften oberflächensprachlich ausprägen. Im Folgenden sollen dabei besonders die Konnektivität (wie werden die einzelnen Sätze und die in ihnen ausgedrückten Propositionen miteinander verbunden?) und die Thematizität (wie wird das Thema entfaltet?) interessieren. Verbleiben die Texte bei der deskriptiven Themenentfaltung oder finden sich auch narrative Elemente?

## 3 Themenentfaltung konkret

Der folgende Auszug aus einem automatisierten Spielbericht soll zunächst einen Eindruck von der Art der Texte vermitteln:

> 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. (retresco #F95RBL)

Man sieht, dass der Algorithmus offenkundig in der Lage ist, einzelne Informationen vor dem Hintergrund gewisser statistischer Erwartbarkeiten so zu gewichten, dass eine gewisse Außergewöhnlichkeit des Geschehenen thematisiert wird. Dass bereits nach wenigen Momenten der erste Treffer fällt und dass insgesamt drei Treffer in den ersten 16 Minuten eine turbulente Startphase bedeuten, geht über einen bloßen Bericht hinaus und setzt das für Erzählungen typische Moment der »tellability« (Norrick 2005) um. Darüber hinaus finden sich auch vielfältige konnektive Mittel. Die im zweiten Satz verwendete Redewendung *zur Stelle sein* etwa greift das vorgenannte *Treffer fallen* präzisierend wieder auf. Im dritten Satz findet sich die anaphorisch gebrauchte Antonomasie *des 24-jährigen Stürmers*, und die Rede vom verpassten *Ausbau der Führung* oder dem *nächsten Treffer* im vorletzten Satz greifen ebenfalls Vorerwähntes wieder auf. Das gesamte bislang beschriebene Spielgeschehen wird schließlich in der bereits erwähnten evaluierenden Wendung der *turbulenten Startphase* wieder aufgegriffen.

Es zeigt sich also, dass der Algorithmus durchaus subtile sprachliche Kunstgriffe einsetzt, um einen kohärenten Text mit narrativen Elementen zu erzeugen. Zum direkten Vergleich sei hier noch die entsprechende Passage aus einem *kicker*-Spielbericht zitiert:

> 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.). […] 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.). (kicker #F95RBL)

Hier zeigen sich noch einmal andere sprachliche Mittel der Konnektivität. Dass Poulsen *zunächst* scheitert, dann aber *erneut* zum Abschluss kommt und *diesmal* trifft, verweist auf eine umfassendere Repräsentation des Spielgeschehens durch den menschlichen Autor, der noch einmal subtiler so etwas wie Überraschungsmomente inszenieren kann. Gleiches gilt für die Adverbien *noch* und *dann*, mit denen die Aktionen von Werner und Poulsen auf eine Weise zueinander ins Verhältnis gesetzt werden, die Spannung erzeugen kann.

## 4 Korpuslinguistische Befunde

Um diese qualitativen Befunde zu Unterschieden zwischen automatisierten und menschlichen Texten weiter anzureichern, bietet sich eine korpuslinguistische Analyse von sogenannten Keywords an. Keywords sind Wörter, deren Frequenzen in einem Untersuchungskorpus sich signifikant von denen in einem Referenzkorpus unterscheiden (vgl. Culpeper & Demmen 2015). Als Untersuchungskorpus dienen hier alle automatisierten Texte, die mit allen menschlichen Texten verglichen werden. Die Keywords lassen sich auf Lemmaebene (Grundformen) wie auch auf Wortartenebene berechnen.

Dabei zeigt sich, dass in den menschlichen Texten u.a. die folgenden Lemmata häufiger verwendet werden als in den automatisierten Berichten:

> *aber, dann, zunächst, wieder, auch, doch, danach, erneut, zwar, fast, weil*

Dies bestätigt die Beobachtung aus der qualitativen Analyse. Insbesondere adversative Konnektoren (vgl. Breindl, Volodina & Waßner 2014) wie *aber, doch* und *zwar*, die Aussagen in Opposition zueinander setzen und mithin ein Spannungsverhältnis inszenieren können, aber auch temporale Ausdrücke wie *zunächst* und *wieder*, die Spielverläufe umfassend repräsentieren können und ihrerseits adversative Lesarten haben können, sind typisch für menschliche Texte. Aufschlussreich ist auch das kontrafaktische Adverb *fast*, das sich in den automatisierten Texten kaum findet.

Auf Wortartenebene zeigt sich, dass subordinierende Konjunktionen in den automatisierten Texten häufiger sind als in den menschlichen. Das mag zunächst überraschen, da Nebensätze ein typisches Merkmal syntaktischer wie auch satzsemantischer Komplexität sind. Bei einer Auswertung der häufigsten Konjunktionen in den beiden Teilkorpora zeigt sich jedoch, dass in den automatisierten Texten die temporale Konjunktion *als* hochfrequent ist, die indes auf sehr eigentümliche Weise gebraucht wird:

> 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 #B04FCB)

Es zeigt sich, dass in Haupt- und Nebensatz jeweils die gleiche Information vermittelt, also eigentlich gar keine temporale Verknüpfung vorgenommen wird. Besonders häufig in den menschlichen Texten wird die kausale Konjunktion *weil* verwendet, um bestimmte Aspekte des Spiels erklären zu können:

> In der Schlussphase wurde es noch einmal turbulent, weil beide Teams den Sieg wollten. (kicker #VFBFCN)

In den automatisierten Texten wird *weil* insgesamt seltener verwendet, und dann in ebenfalls eigentümlicher Weise:

> Die verbleibende Zeit der ersten Halbzeit blieb ohne weitere Treffer, weil die Chancen durch Fink, Gießelmann und Usami ohne Erfolg blieben. (retresco #F95RBL)

Auch hier wird im *weil*-Satz keine echte Begründung geliefert, sondern eigentlich nur die im Hauptsatz bereits vermittelte Information weiter elaboriert. Die automatisierten Texte weisen also eine gewisse syntaktische Komplexität auf, die jedoch nicht der semantischen Komplexität, wie sie für menschliche Texte üblich ist, entspricht.

## 5 Fazit

Die text- und korpuslinguistische Analyse zeigt, dass die automatisierten Fußballspielberichte als Modellierungen (vgl. Scharloth 2016) und manchmal auch bloße Simulationen von narrativer und zumeist nur deskriptiver Textualität beschrieben werden können. Der Verlauf des Spiels wird im fortlaufenden Text in Ansätzen als schrittweise Themenentfaltung repräsentiert, für die textlinguistisch gut beschreibbare Mittel der syntaktischen und semantischen Konnektivität zum Einsatz kommen. Die narrative Ausgestaltung der Ereignisdokumentationen scheint aber dennoch der menschlichen (Nach)Bearbeitung vorbehalten zu bleiben.

Die nähere Zukunft wird zeigen, ob neuere, KI-basierte Technologien der Textgenerierung, die nicht mehr regelbasiert, sondern auf der Grundlage statistischer Sprachmodelle funktionieren, die Texte in dieser Hinsicht werden verbessern können. Noch scheitert deren Einsatz in der konkreten Berichterstattung über realweltliche Ereignisse daran, dass Sprachmodelle wie etwa GPT-3 letztlich nur Formulierungsmuster reproduzieren und rekombinieren können, die

basierend auf dem Trainingskorpus gelernt wurden. Denkbar wäre aber eine Kombination aus beiden Technologien, in der das Grundgerüst der Texte regelbasiert erstellt und dann mithilfe eines Sprachmodells stilistisch angereichert wird. In jedem Falle wird es spannend zu beobachten sein, wie sich Texte und auch unser Verständnis von Texten und den in ihnen verhandelten Themen in Zukunft entwickeln werden.

## 6 Literatur

Beaugrande, Robert de & Wolfgang Dressler. 1981. *Introduction to Text Linguistics*. London / New York: Routledge.

Breindl, Eva, Anna Volodina & Ulrich Hermann Waßner. 2014. *Handbuch der deutschen Konnektoren 2: Semantik der deutschen Satzverknüpfer.* Berlin/ München/Boston: de Gruyter. https://doi.org/10.1515/9783110341447.

Culpeper, Jonathan & Jane Demmen. 2015. Keywords. In Douglas Biber & Randi Reppen (eds.), *The Cambridge Handbook of English Corpus Linguistics*, 90–105. Cambridge: Cambridge University Press. https:// doi.org/10.1017/CBO9781139764377.006.

Diakopoulos, Nicholas. 2019. *Automating the News: How Algorithms Are Rewriting the Media*. Harvard: University Press.

Haim, Mario & Andreas Graefe. 2018. Automatisierter Journalismus. In Christian Nuernbergk & Christoph Neuberger (eds.), *Journalismus im Internet: Profession – Partizipation – Technisierung*, 139–160. Wiesbaden: Springer Fachmedien. https://doi.org/10.1007/978-3-531-93284-2_5.

Hausendorf, Heiko & Wolfgang Kesselheim. 2008. *Textlinguistik fürs Examen (Linguistik fürs Examen 5)*. Göttingen: Vandenhoeck & Ruprecht.

Norrick, Neal R. 2005. The dark side of tellability. *Narrative Inquiry* 15(2). 323–343. https://doi.org/10.1075/ni.15.2.07nor.

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. https://doiorg/10.1515/9783110451542-013.