Florian Duda

„Was war oder ist Ihre schönste, tollste und angenehmste Kindheitserinnerung?“ – Ein sprachwissenschaftlicher Ansatz zur Machine-Learning-Datengenerierung

Abstract Deutsch

Das Training von Systemen, die auf Maschinenlernen basieren, kann ein herausforderndes Unterfangen darstellen. Diese Herausforderungen sind oftmals eng mit Fragen nach der Quantität und Qualität der verwendeten Datensätze verbunden, um zu gewährleisten, dass solche Systeme auch nach den Testphasen im freien Feld verlässliche Ergebnisse liefern (Sessions/Valtorta 2006). Betrachtet man Chatbots bzw. kognitive virtuelle Assistenten, die auf der Verarbeitung von natürlicher Sprache basieren und für Interaktionsaufgaben mit zufälligen Nutzer*innen konzipiert wurden, wird dieses Problem umso deutlicher, da von signifikanten Abweichungen zwischen bereits vorhandenen Testdaten und neuen Nutzereingaben in den Bereichen der Sprachqualität als auch des transportierten Inhalts ausgegangen werden kann.

Im Sommer 2019 entwickelte der Autor ein systematisches, sprachwissenschaftliches Modell, das es einem sogenannten interaktiven digitalen Zeugnis ermöglichen soll, eine möglichst hohe Trefferquote bei der Zuordnung von neuen Nutzereingaben zu bereits vorhandenen Datensätzen zu erzielen. In dem folgenden deskriptiven Forschungsbericht soll die Notwendigkeit der vorgenommenen Datengenerierung bzw. -variation anhand des Korpus von Fragen, die im Dezember 2018 dem jüdischen Zeitzeugen Abba Naor gestellt wurden, beleuchtet werden, um daran anschließend auf die Entwicklung des zu diesem Zweck konzipierten Modells einzugehen. Abschließend sollen einige mit diesem Ansatz einhergehende ethische und sprachwissenschaftliche Schwierigkeiten im Licht der interaktiven digitalen Zeugnisse diskutiert werden.

- Eine allgemeine Einführung in die Grundprinzipien des Maschinenlernens und die Anwendung von Maschinenlernen in den interaktiven digitalen Zeugnissen des LediZ-Projekts.
- Die Beleuchtung der mit der Datengenerierung und -variation einhergehenden ethischen und sprachwissenschaftlichen Fragen.
- Die Beschreibung des sprachwissenschaftlichen Modells zur Datengenerierung bzw. -variation.

Abstract English

Training machine learning systems can be challenging, and these challenges are often closely linked to the quality and quantity of the datasets required to ensure that such systems continue to deliver reliable results when the test phase is complete (Sessions/Valtorta 2006). When it comes to setting up chatbots or cognitive virtual assistants, which process natural language and are designed to interact with random users, that problem becomes even more evident as user input can deviate enormously from test input, both in terms of speech quality and content.

In summer 2019, the author developed a systematic linguistic model that would enable an interactive digital testimony to achieve the highest possible hit rate when assigning new user input to existing data sets. This descriptive research report outlines the importance of data generation and variation, specifically based on the corpus of questions put to the Jewish contemporary witness Abba Naor in December 2018. The report then describes the development of the model conceived for this purpose before finally discussing the ethical and linguistic issues of the described methodology for the field of digital testimonies.

This paper will cover the following aspects:

- An introduction to the general principles of machine learning and its application in the creation of digital interactive testimonies within the LediZ-Project.
- A discussion of the ethical and linguistic issues associated with this kind of data generation and variation.
- A description of the linguistic model used for data generation and variation.

Zur Bedeutung von kognitiven virtuellen Assistenten in Alltag und Forschung

In vielen Bereichen des Internets, die darauf abzielen, bestimmte Produkte oder Dienstleistungen zu verkaufen, werden heutzutage sogenannte kognitive virtuelle Assistenten (engl. *cognitive virtual assistants*) eingesetzt, die den Kund*innen generelle Fragen beantworten, aber auch Bestell- oder Buchungsvorgänge autonom und ohne menschliche Hilfe durchführen können (Sabharwal/Agrawal 2020: 1). Ein Beispiel für einen solchen kognitiven virtuellen Assistenten (KVA) ist der Service-Chatbot „Kai" der Deutschen Bahn, der für Nutzer*innen zu fest vorgegebenen Themenoptionen oder aber freien Suchanfragen allgemeine Informationen rund um das Thema „Bahnreisen" bereitstellt. Folgt man während des Suchprozesses den vorgegebenen Auswahloptionen, funktioniert dieser KVA sehr präzise und man erhält binnen weniger Mausklicks die gesuchte Information, verwendet man hingegen die freie Eingabe und stellt speziellere Fragen, werden schnell die Grenzen des Assistenten deutlich und Kai verweist den/die Fragensteller*in in seiner Antwort auf seine menschlichen Kolleg*innen. Doch die Einsatzmöglichkeiten von KVA beschränken sich nicht allein auf den Bereich des Kundenservices, die Anwendungsbereiche können vielmehr ebenso mannigfaltig sein wie die konkreten Funktionsweisen der KVA (Stucki et al. 2020: 3f.).

Im Münchener Projekt LediZ (Lernen mit digitalen Zeugnissen) wird ein KVA zur Erkennung von Fragen, die einem sogenannten interaktiven digitalen Zeugnis gestellt werden, und für die anschließende Zuordnung der gestellten Fragen zu passenden Antworten eingesetzt. Das LediZ-Team entschied sich dabei gegen die Vorgabe von festen Frageoptionen und für einen freien Frage-Antwort-Prozess, was es den Fragenden ermöglichen soll, selbst zu entscheiden, welche Frage sie stellen und wie sie ihre Formulierung syntaktisch und lexikalisch gestalten. Ferner sollen die Kommunikant*innen dadurch befähigt werden, interessengeleitet individuelle Fragen zu stellen und das Zeugnis auf diese Weise selbstständig zu erschließen (vgl. Ballis et al. 2019: 428). Eine Herausforderung, die mit dieser Entscheidung einherging, war es, den auf ursprünglich 1.000 Fragen basierenden KVA so zu trainieren, dass er möglichst viele Varianten der im Zeitzeugeninterview gestellten Fragen den passenden Videoantworten zuordnen konnte. Dabei mussten die als Trainingsdaten generierten Fragenvarianten von einer solchen Qualität sein, dass sie zwar insgesamt eine möglichst breite Formulierungsvielfalt von Fragen abdeckten, sich andererseits aber der Wahrheitsgehalt der Aussagen des Zeugnisses nicht änderte.

Im Nachfolgenden soll zunächst der Prozess der Kommunikation mit einem interaktiven digitalen 3D-Zeugnis konzise dargestellt werden, um daran anschließend auf die grundsätzliche Funktionsweise von kognitiven virtuellen Assistenten einzugehen und die Bedeutung von Trainingsdaten herzuleiten. Der letzte Teil wird sich der Beschreibung eines sprachwissenschaftlichen Modells zur Trainingsdatengenerierung unter Einbezug ethischer Standards widmen.

Die Kommunikation mit einem interaktiven digitalen Zeugnis

Die beiden zum Zeitpunkt dieses Aufsatzes im Projekt LediZ entwickelten interaktiven digitalen Zeugnisse der zwei Zeitzeug*innen Abba Naor und Eva Umlauf beruhen auf Interviews, die im Dezember 2018 bzw. Januar 2019 im Pollen Studio in England durchgeführt wurden. Bei diesen Interviews wurden sowohl die Lebensgeschichte der Zeitzeug*innen festgehalten als auch jeweils ca. 1.000 Fragen zu verschiedenen biografischen Themenbereichen gestellt und stereoskopisch aufgenommen. Die Frage-Antwort-Paare bilden die Grundlage für den späteren Kommunikationsprozess mit dem interaktiven 3D-Zeugnis. Hierbei ist anzumerken, dass im Gegensatz zu den Fragen, die von den Fragensteller*innen flexibel ausformuliert werden können, die Antwort des digitalen Zeugnisses stets unverändert bleibt und somit der Antwort entspricht, die zum Zeitpunkt des Interviews von dem Zeitzeugen oder der Zeitzeugin gegeben wurde.

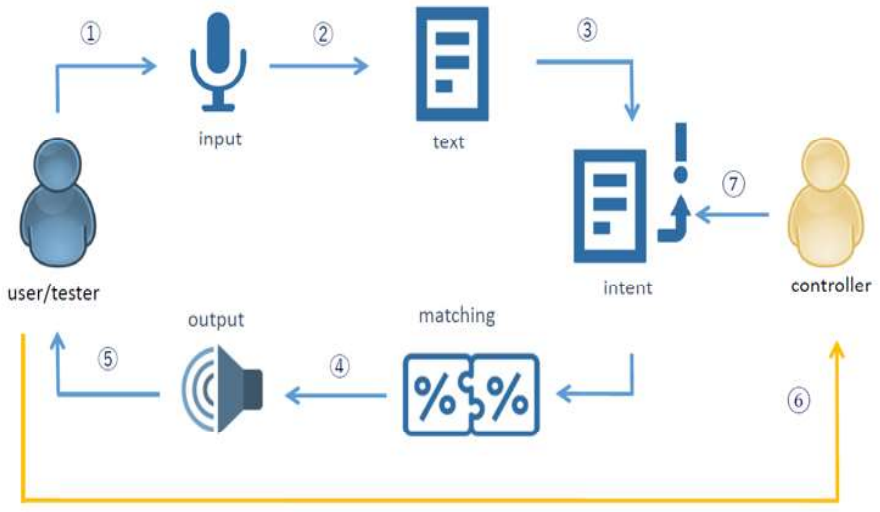

Abbildung 1 – Prozess des Fragenstellens und die technische Verarbeitung der Frage nach Ballis und Kolb (Ballis/Barricelli 2019)

Das kognitive Zentrum des interaktiven digitalen 3D-Zeugnisses, das sowohl für die Spracherkennung als auch für die Sprachverarbeitung im Prozess des Fragenstellens zuständig ist, bildet der KVA *Google Dialogflow*. Der KVA muss innerhalb der Anwendung im ersten Schritt zunächst die in gesprochener Sprache mithilfe eines Mikrofons an das Zeugnis gestellte Frage des Nutzers/der Nutzerin erkennen und diese anschließend in einen segmentierten Text transkribieren (1&2). Diese Transkription wird dann mit den in der Datenbank des KVA vorhandenen Ursprungsfragen bzw. Fragenvarianten abgeglichen, um die mögliche Absicht des Fragestellers/der Fragenstellerin („Intent") zu ermitteln (3). Im Falle einer signifikanten Übereinstimmung („Match") zwischen gestellter Frage und der in der Datenbank vorhandenen Frage bzw. Fragenvariation wird eine Anfrage an den Server gestellt, in dem die Videoclips mit den Antworten aus den Interviews gespeichert sind, und eine Videodatei, die auf die Frage antwortet, wird abgespielt (4&5). Sollte der Abgleich der gestellten Frage und der Datenbank zu keiner Übereinstimmung führen, wird die Frage einem sogenannten „Fallback Intent" zugeordnet, der eine Videodatei aufruft, die den Nutzer/die Nutzerin entweder dazu auffordert, die gestellte Frage zu reformulieren, oder aber mitteilt, dass in der Datenbank keine passende Antwort vorhanden ist.

Da die Fehlerquote der Zuordnung von Frage und Antwort besonders zu Beginn des Trainings, wenn nur wenige Trainingsdaten bzw. nur die ursprünglich gestellten Fragen in der Datenbank vorhanden sind, besonders hoch ist, muss ein menschlicher Prüfer/eine menschliche Prüferin („controller") die zusammengeführten Frage-Antwort-Paare nochmals überprüfen und gegebenenfalls händische Korrekturen an der Zuordnung der Datensätze vornehmen (6).

Beispiel der Funktionsweise eines KVA-Systems

Dieser Teil soll die Funktionsweise der Sprachverarbeitung und des Sprachverstehens des KVA *Google Dialogflow* explizieren und darauf aufbauend den Einfluss von Trainingsdaten auf die Trefferquote im Zuordnungsprozess von Frage und Antwort eruieren.

Sabharwal und Agrawal (2020) definieren einen KVA als einen Software-Agenten (engl. *software agent*), der bestimmte Aufgaben für Menschen oder Maschinen auf Grundlage von Text-, Sprach- oder Bildeingaben durchführt (Sabharwal/Agrawal 2020: 2). Häufig wird der Ausdruck „KVA" in der Forschungsliteratur synonym zu dem Begriff „Chatbot" verwendet, wobei einige Autor*innen darauf verweisen, dass Chatbots nicht

immer eine bestimmte funktionale Aufgabe außerhalb der reinen Kommunikation erfüllen müssen (vgl. Stucki et al. 2020: 3). Im Projekt LediZ liegen die Hauptaufgaben des KVA *Google Dialogflow*, wie bereits dargelegt, in der Verarbeitung (engl. *natural language processing*) und dem Verstehen (engl. *natural language understanding*) natürlicher Sprache sowie der daran anknüpfenden Zuordnung von Frage-Antwort-Paaren. Um zu verstehen, wie *Google Dialogflow* Sprache verarbeitet und die Intention hinter einer Spracheingabe erkennt, möchte ich zunächst den Unterschied zwischen datenbasierten und streng[1] regelbasierten Systemen zur Sprachverarbeitung anhand eines Beispiels explizieren:

In der Beispielkommunikation wollen wir von beiden Systemen wissen, ob sie über spezielle Namen verfügen, weshalb es naheliegend wäre, den zwei Systemen die Frage „Wie heißt du?“ entweder in gesprochener oder in geschriebener Sprache zu stellen.

Ein streng regelbasierter KVA würde auf diese Frage nur dann eine Ausgabe in Form einer Antwort produzieren, wenn die Frage in einem durch den/die Programmier*in integrierten Muster bzw. einer Regel zuvor erfasst wurde. Beispielsweise in Form einer einfachen Wenn-Dann-Funktion wie: „Wenn Input ‚Wie heißt du?‘, dann Output ‚NAME‘.“ Wurde der KVA für einen förmlichen kommunikativen Kontext entwickelt und würde daher nur die Frage „Wie heißen Sie?“ berücksichtigen, würden wir trotz derselben Intention der beiden Fragen auf die von uns gestellte Frage mit dem Personalpronomen „du“ keinen Output generieren. Es muss folglich bereits im Entwicklungsprozess eines streng regelbasierten KVA überlegt werden, welche möglichen Fragen dem System gestellt sowie in welcher syntaktischen, morphologischen und lexikalischen Form diese womöglich realisiert werden (vgl. Stucki et al. 2020: 7).

Ein datenbasierter KVA könnte im Gegensatz dazu die frei formulierte Frage nach seinem Namen auch dann verstehen, wenn das der Frage zugrundeliegende sprachliche Muster bzw. die Zeichenfolge (engl. *string*) nicht explizit im Entwicklungsprozess berücksichtigt wurde. Grundlegend für diese Funktion ist die Kompetenz, aus Eingaben von Nutzer*innen nachträglich definierte Nutzerabsichten (engl. *intents*) abzuleiten. Diese Fähigkeit des nachträglichen Lernens basierend auf Komponenten mit freien Variablen und Parametern wird Machine Learning oder maschinelles Lernen genannt (Serban et al. 2015). In unserem Beispiel könnte der KVA folglich die Frage nach seinem

[1] In meinem Beispiel nehme ich einen Algorithmus an, der nur identische Zeichenfolgen als Übereinstimmungen erkennt, etwaige Teilübereinstimmungen der Zeichenfolgen bleiben dabei unberücksichtigt bzw. werden vom Algorithmus als Nichtübereinstimmung gewertet.

Namen auch dann noch erlernen, wenn die hinter dieser Frage stehende Absicht von dem System zunächst nicht erkannt wurde. Im Falle des interaktiven digitalen Zeugnisses würden Prüfer*innen (vgl. (6) auf Seite 4) die Frage „Wie heißt du?“ entweder, sofern ein Antwortvideo im Server vorhanden ist, als Intent (Fragenabsicht) neu erstellen oder – wie im Falle der beiden Zeugnisse des LediZ-Projektes – die Frage der Ursprungsfrage „Wie heißen Sie?“ zuordnen. Das System würde demnach lernen, dass Fragensteller*innen mit den beiden Fragen „Wie heißt du?“ und „Wie heißen Sie?“ dieselbe Absicht, nämlich das In-Erfahrung-bringen des Namens des Zeitzeugen bzw. der Zeitzeugin, verfolgt.

Bereits anhand dieses einfachen Beispiels wird der Vorteil eines datenbasierten KVA wie *Google Dialogflow* und der Einfluss von Varianten auf einzelnen Intents deutlich: Je mehr Varianten eines Intents in der Datenbank des KVA vorhanden sind, desto höher ist die Wahrscheinlichkeit, dass das System die Frage eines zufälligen Nutzers oder einer zufälligen Nutzerin korrekt verarbeitet und einer passenden Antwort zuordnet, sofern die Absicht mit einer bereits vorhandenen Frage übereinstimmt (vgl. auch Sabharwal/Agrawal 2020: 38).

Verwendete Zuordnungsalgorithmen in *Google Dialogflow*

Um zu verstehen, wie die in *Google Dialogflow* verwendeten Zuordnungsalgorithmen funktionieren, werden diese nachfolgend anhand der semantischen Erschließung von Interrogativsätzen näher beleuchtet. Letztere ermöglicht es dem KVA, die Bedeutung von Nutzer*inneninputs zu erschließen und von den Ursprungsfragen abweichende Fragen der jeweiligen Fragenabsicht korrekt zuzuordnen. *Google Dialogflow* folgt im Erkennungs-, Verarbeitungs- und Verstehensprozess einem sogenannten Pipelinemodell (vgl. Kersting et al. 2019: 75f.): Ausgehend von Schallinformationen wird ein segmentierter Text generiert. In der morphologischen Analyse werden Flexions- und Deklinationsformen analysiert und ihre Grundformen wiederhergestellt. Daran anschließend wird in der syntaktischen Analyse der Aufbau des Satzes ausgewertet. Die folgende semantische Analyse dient der Bedeutungs- und Intentionserschließung der eingegebenen Sätze. Abschließend können je nach Kontext, in dem der KVA verwendet wird, in der Diskursanalyse mögliche (weitere) Absichten, Zwecke und Intentionen der Eingabe abgeleitet werden.

Die in *Google Dialogflow* konkret durchgeführte Zuordnung von Nutzer*innen-eingaben zu vorhandenen Intents basiert dabei immer auf einem regelbasierten

Grammatikabgleich und *kann* gleichzeitig auch auf einem ML(Machine Learning)-Abgleich fußen (Google 2020). Der letztgenannte Algorithmus[2] verfolgt dabei das Ziel, auch Inputs zu erkennen, die nicht 1:1 mit den in der Datenbank von *Google Dialog* vorhandenen Intents und Intentvariationen übereinstimmen. Der ML-Abgleich ist deswegen wichtig, da dieser jene in der gesprochenen Sprache häufiger auftretenden Fehler oder Abweichungen von der Standardsprache berücksichtigt, die in der rein menschlichen Kommunikation – je nach Schwere des Fehlers – nicht zwangsläufig zu Sprachverstehensproblemen führen würden. Eine Maschine, die mit streng regelbasierten Algorithmen arbeitet, könnte diese Abweichungen ohne ML-Abgleiche allerdings nicht korrekt verarbeiten. Hierbei greift *Google Dialogflow* auf stochastische Verfahren zurück, die es erlauben, auch von den vorhandenen Datensätzen abweichende *leicht* fehlerhafte[3] Spracheingaben zu erkennen und der entsprechenden Antwort zuzuordnen (vgl. Lobin 2010: 16). Ein Beispiel für eine leicht fehlerhafte Eingabe wäre die segmentierte Transkription „we heiß du?“, die ohne stochastische Verfahren dem Default Fallback Intent „Hierauf habe ich leider keine Antwort.“ zugeordnet werden würde, obwohl die phonetisch ähnliche Frage „Wie heißt du?“ durchaus in der Datenbank vorhanden ist.

Der regelbasierte Grammatikabgleich geht historisch auf die von Noam Chomsky entwickelte Theorie der Generativen Grammatiken zurück und stellt im Vergleich zum ML-Abgleich einen Ansatz dar, der rein auf Basis der Eingaben und der in der Datenbank vorhandenen Daten arbeitet (Chomsky 1957). Chomsky geht dabei davon aus, dass Sprachen durch feste grammatische Regeln generiert werden. Das bedeutet auf einer mikrosprachlichen Ebene, dass ein Satz S einer bestimmten Sprache L nur dann als grammatikalisch korrekter Satz erkannt wird, wenn er den bestimmten Regeln der jeweiligen Grammatik folgt und sich den Buchstaben bzw. Zeichen eines festgelegten Alphabets bedient. Was zunächst trivial erscheinen mag, ist von fundamentaler Bedeutung für die Erkennung und Verarbeitung von Sprache durch Maschinen. Denn eine Maschine kann, sobald sie die grammatischen Regeln und somit die Struktur einer Sprache kennt, die Menge aller grammatikalisch korrekten Sätze einer Sprache bewerten und ebenfalls solche Sätze erzeugen (vgl. Schlobinski 2003: 90).

---

[2] *Google Dialogflow* arbeitet sowohl mit Algorithmen, sprich eindeutigen Handlungsvorschriften zur Problemlösung, als auch Lernalgorithmen, also Regeln, die auf den eingespeisten Daten basieren (vgl. Kersting et al. 2019: 11, 22).

[3] „Fehlerhaft“ bezieht sich auf Fehler in den Bereichen Syntax, Morphologie und Lexik, die von einem grammatikbasierten Algorithmus nicht korrekt zugeordnet werden würden.

Da es im Folgenden auch um die Ableitung der Bedeutung aus Ergänzungsfragen, also solche Fragen, die nicht allein mit „ja“ oder „nein“ beantwortet werden können, gehen wird, führe ich an dieser Stelle das Beispiel „Was ist Ihre schönste Kindheitserinnerung?“ an. Diese Ergänzungsfrage folgt im Deutschen dem syntaktischen Schema einer Komplementierer-Phrase (CP), die für unser Beispiel wie folgt in einem einfachen Phrasenstrukturbaum dargestellt werden kann:

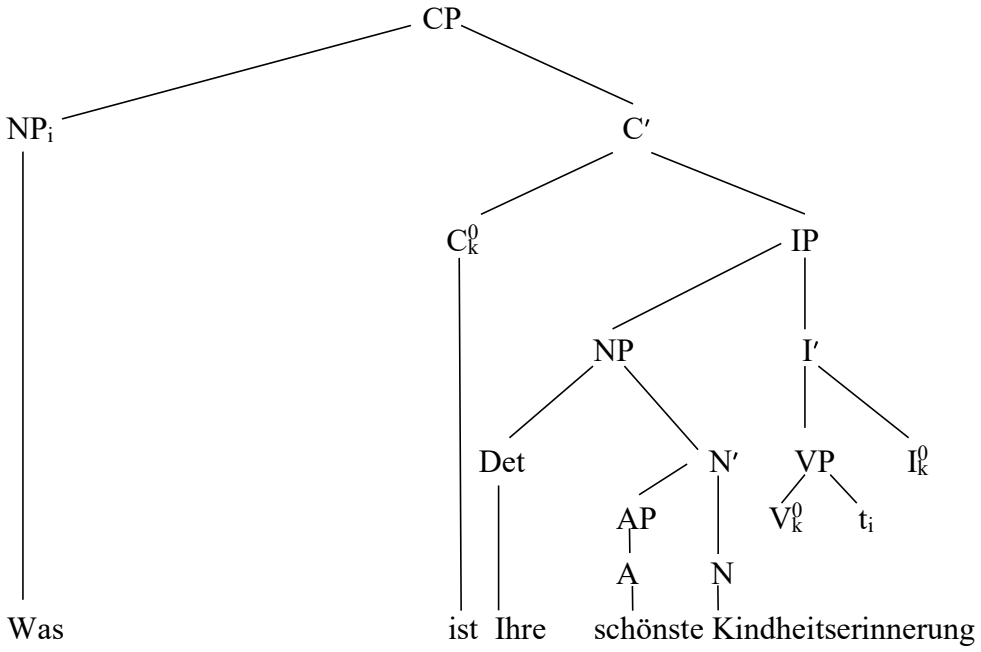

Abbildung 2 – Phrasenstrukturmodell der Ergänzungsfrage „Was ist Ihre schönste Kindheitserinnerung?“

Deutlich wird in dieser strukturierten Schreibweise der Ergänzungsfrage mindestens zweierlei: Zum einen besteht unser Beispielsatz aus verschiedenen Phrasen: Nominalphrasen (NP), Komplementierer (C), Flexionsphrase (IP), Adjektivphrase (AP) und Verbalphrase (VP). Ferner sind die Positionen der einzelnen Wörter nicht zufällig, sondern unterliegen syntagmatischen Regeln. So besteht eine Nominalphrase immer zumindest aus einem Nomen (N). In unserem Beispiel besteht die Nominalphrase aus einem Determinativ (Det), einer Adjektivphrase bzw. einem Adjektiv und einem Nomen. Die Anordnung der einzelnen Bestandteile dieser Nominalphrase folgt hierbei wiederum festen Regeln, weswegen beispielsweise eine AP-Det-N-Anordnung („schönste Ihre Kindheitserinnerung“) als ungrammatisch gewertet werden würde. Aus dem Phrasenstrukturmodell lassen sich jedoch nicht nur syntagmatische Muster ablesen. Das Modell gibt ebenfalls Aufschluss darüber, welche Struktur eine Antwort auf diese Fragen

haben könnte. Für die erfragte Nominalphrase (NP) kommen dabei zwei Stellen in Frage: Entweder die Fokusstelle NP_i (bspw. „Mein achter Geburtstag ist meine schönste Kindheitserinnerung.“) oder aber die Stelle der Spur t_i als Teil der Verbalphrase (VP) (bspw. „Meine schönste Kindheitserinnerung ist mein achter Geburtstag.“). Ferner wird u. a. bereits in der Frage innerhalb der Verbalphrase (VP) der Kasus der erfragten Nominalphrase NP_i durch das Verb ($V^{()}_k$) festgelegt.

Ein regelbasierter Grammatikabgleich dient folglich nicht nur dazu, zu bewerten, ob ein von der Spracherkennung transkribierter Text den Regeln einer bestimmten Grammatik folgt oder ob die grammatische Struktur mit der Struktur vorhandener Daten übereinstimmt, er lässt ferner Voraussagen über die *mögliche* Syntax und Kasusmerkmale des Antwortsatzes zu. Fragenvarianten geben *Google Dialogflow* hierbei etwaige Syntaxmuster vor, die dem im freien Feld erwartbaren syntaktischen Aufbau der gestellten Fragen entsprechen können.[4]

Durch das Wissen über die Tiefenstruktur von Sätzen kann ein KVA jedoch nicht nur syntaktische Symbolfolgen bewerten, er kann daraus auch Bedeutungsmöglichkeiten ableiten (vgl. Mainzer 2019: 68). Wie anhand des Phrasenstrukturmodells gezeigt, besteht ein Satz aus mehreren Phrasen, die durch feste Regeln weiterzerlegt werden können, bis man zu einzelnen Wörtern einer Sprache gelangt (vgl. Mainzer 2019: 68). So kann sich in einem Satz wie „Die Frau schlug den Mann mit dem Lexikon.“ die Präpositionalphrase „mit dem Lexikon“ entweder auf die Tätigkeit des Schlagens beziehen oder aber ein Attribut des geschlagenen Mannes darstellen. Ob ein und derselbe Satz mehrere Bedeutungen besitzt, wird für die Maschine erst durch die Analyse der Tiefenstruktur der Sätze sichtbar.

Am Beispiel von Ergänzungsfragen wie „Was ist Ihre schönste Kindheitserinnerung?“ und „Was ist Ihre schlimmste Kindheitserinnerung?“, die aus Sicht der Syntax identisch sind, wird deutlich, weswegen die Bedeutung eines Satzes nicht allein von der Art der Verknüpfung seiner Phrasen abhängig ist. Vielmehr sind auch die Bedeutungen der einzelnen Wörter, die diese Phrasen bilden, zu berücksichtigen. Man spricht hierbei vom sogenannten Frege-Prinzip (vgl. Lobin 2010: 85). Ein KVA wie *Google Dialogflow* muss folglich zunächst lernen, dass es einen semantischen Unterschied zwischen den Adjektiven „schlimm“ und „schön“ gibt, bevor er die entsprechenden Fragen korrekt weiterverarbeiten kann. Auf Ebene der Aussagenlogik impliziert man dabei mit der Frage

[4] Die Auseinandersetzung mit dem ML-Abgleich wird zeigen, wieso es i. d. R. nicht notwendig ist, alle möglichen syntaktischen Muster in Fragenvariationen abzubilden.

„Was ist Ihre schönste Kindheitserinnerung?“, dass etwas existiert, das Teil der Menge aller Kindheitserinnerungen ist. Gleichzeitig aber auch, dass dieses Existierende das Attribut bzw. die Eigenschaft „schön“ besitzt ($\exists x \in M : A(x)$), und nicht nur, dass eine „irgendwie“ attribuierte Kindheitserinnerung existiert ($\exists x \in M$).

Um nun die Bedeutung von Sätzen zu erschließen, könnte daher parallel zum Grammatikabgleich ein Semantikabgleich ablaufen, bei dem ein Satz einer natürlichen Sprache in einen aussagenlogischen Satz überführt und anschließend die gewonnene logische Struktur der Eingabe mit den vorhandenen logischen Strukturen der Datenbank verglichen wird. Da die Wörter einer Sprache jedoch häufig polysem und kontextsensitiv sind, kann ein KVA mit diesem Verfahren die Bedeutung eines Satzes aus seinen einzelnen Bestandteilen immer nur bis zu einem gewissen Grad[5] mathematisch berechnen (vgl. Lobin 2010: 85).

Um Bedeutungsunterschiede zwischen verschiedenen Sätzen trotz Ambiguität und sprachlichen Ungenauigkeiten zu erkennen, nutzt *Google Dialogflow* zusätzlich zu dem regelbasierten Grammatikabgleich einen auf Machine Learning basierenden zweiten Abgleich (ML-Abgleich) (vgl. Bitter et al. 2010: 152). Dabei verwendet der KVA ein Sprachmodell (engl. *language model*), das aus wiederkehrenden Mustern der Ursprungsfrage und der zugeordneten Trainingsdaten bzw. Fragevariationen in der Datenbank erzeugt sowie fortlaufend optimiert wird (vgl. Aggarwal 2018: 4). Ein Sprachmodell stellt einen Ansatz dar, auf Grundlage von Daten eine Repräsentation von Text zu generieren und Muster aus diesen Daten abzuleiten (vgl. Aggarwal 2018: 4). Ferner lernt der KVA zwischen bedeutungstragenden Wörtern und Phrasen und weniger bedeutungsrelevanten Bestandteilen eines Satzes (z. B. Füllwörtern) zu unterscheiden. Darüber hinaus kann das System dahingehend trainiert werden, dass es Signifikanzen zwischen einzelnen Wörtern erkennt. So kann ein KVA derartig modifiziert werden, dass bspw. das Artikelpaar „ein“ und „der“ weniger Signifikanz für die Erschließung der Bedeutung besitzt als die antonymen Adjektive „schön“ und „schlimm“ (vgl. Aggarwal 2018: 5f.). Ein wesentlicher Nachteil des ML-Abgleichs ist die große Ungenauigkeit der Zuordnung von gestellten Fragen zu vorhandenen Fragen in der Datenbank zu Beginn des Trainings, da das System zu diesem Zeitpunkt noch nicht die bedeutungsrelevanten Muster der Ursprungsfragen erschlossen hat. Aus diesem Grund eignet sich der ML-Abgleich bei wenigen Datensätzen nur bedingt (vgl. Google 2020).

[5] Beispielsweise treten im Bereich von idiomatischen Ausdrücken (z. B. „ich platze gleich“) und Metaphern häufig Fehler bei der aussagenlogischen Analyse des KVA auf.

Variationen der ursprünglich vorhandenen Textdaten sind besonders für ML-Abgleiche erforderlich, da sie nicht nur die Wahrscheinlichkeit, Übereinstimmungen zwischen Eingabe und vorhandenen Daten zu erkennen, erhöhen, sondern auch dazu beitragen, bedeutungsähnliche und bedeutungsrelevante Textelemente zu ermitteln (vgl. Reyes Ochoa et al. 2019: 5). Dies ist auch dann relevant, wenn man mit syntaktisch ähnlichen Ursprungsfragen arbeitet, die sich allein durch einzelne Lexeme (z. B. „schönste Kindheitserinnerung“ vs. „schlimmste Kindheitserinnerung“) voneinander unterscheiden. Daten können folglich als „Grundzutat“ für jeden Lernalgorithmus angesehen werden, ohne die die Bildung eines spezifischen Sprachmodells nicht möglich wäre (Kersting et al. 2019). Je mehr unterschiedliche Varianten einer Ursprungsfrage in der Datenbank vorhanden sind, desto höher ist die Wahrscheinlichkeit, dass der KVA die Frage dem richtigen Intent zuordnet (vgl. Sabharwal/Agrawal 2020: 38).

Abschließend ist darauf hinzuweisen, dass *Google Dialogflow* bei dem ML-Abgleich mit einer sogenannten Intent-Erkennungskonfidenz arbeitet, die zwischen dem Wert 0,0 (völlig unsicher) und 1,0 (vollkommen sicher) schwanken kann (vgl. Google 2020). Hierbei kann es bei einem niedrig eingestellten Schwellenwert für die Konfidenz (ML Classification Threshold) passieren, dass der KVA zwar gestellte Fragen, die einen hohen Korrelationswert zu den in der Datenbank vorhandenen Fragen und Fragenvariationen aufweisen, dem korrekten Intent zuordnet, jedoch darüber hinaus auch Fragen als Übereinstimmung wertet, die lediglich in einem Schlagwort mit Variationen und Ursprungsfrage übereinstimmen. Folglich muss der Wert der Intent-Konfidenz so angepasst werden, dass der KVA nur Eingaben zulässt, die in möglichst vielen Merkmalen mit den Datensätzen in der Datenbank übereinstimmen.

Forschungsethische Überlegungen auf dem Weg zu einem Datengenerierungsmodell

Im Vorangegangenen wurde im Kontext von Syntax (regelbasierter Grammatikabgleich) und Semantik (Semantische Logik, ML-Abgleich) der Einfluss von Daten auf die Funktionalität eines KVA erörtert. Der nun folgende Abschnitt soll sich kritisch mit der Verwendung eines KVA innerhalb eines interaktiven digitalen Zeugnisses auseinandersetzen und darlegen, welche Konsequenzen und qualitativen Anforderungen sich aus der forschungsethischen Diskussion für die Entwicklung eines Modells zur Datengenerierung ergeben. Dieser kurze Exkurs ist notwendig, um den exklusiven Charakter des im darauf folgenden Teil beschriebenen Modells zu verstehen, aufgrund dessen zwar viele der prinzipiell möglichen Fragenvariationen zugelassen, einige

Variationen aufgrund fehlender ethischer Anforderungen dagegen ausgeschlossen werden.

Wieso sind ethische Standards für die Generierung von Fragevariationen notwendig? Es ist gerade die interaktive Form der Erschließung von Geschichte und persönlicher Erinnerung durch ein interaktives digitales Zeugnis, die antithetische Perspektiven in sich vereint. So kann einerseits aus kognitionswissenschaftlicher Sicht ein positiver Effekt durch die aktivere Beteiligung der Lernenden und die interessengeleitete Kommunikation mit dem Zeugnis auf die Speicherung von Information im Langzeitgedächtnis angenommen werden (vgl. Challenor/Ma 2019: 2). Andererseits droht aus medienphilosophischer Perspektive die notwendige Grenze zwischen Produzent*in der zu vermittelnden Botschaft und Medium durch die Verwendung eines KVA zu verschwimmen. Dies kann geschehen, sobald die Maschine bzw. der KVA Entscheidungsgewalt darüber erlangt, nach welchen Regeln die Botschaft zu übermitteln ist, und folglich ihre bzw. seine diaphane Eigenschaft verliert (vgl. Krämer 2008: 30). Wieso es einem solchen Verschwimmen zwischen Medium und Botschaft gerade im Kontext von digitalen interaktiven Zeugnissen vorzubeugen gilt, belegen die Ergebnisse einer Studie von Christina Brüning aus dem Jahr 2019. Hierbei zeigte es sich, dass manche Lernenden die digitalen Zeitzeug*innen im Prozess der Interaktion zu vergessen beginnen und aufhören, das vom Medium Übermittelte zu hinterfragen, es vielmehr unreflektiert zur wahren Geschichte werden lassen (vgl. Brüning 2019: 398). Da der KVA eine Entscheidungsträgerfunktion im Kommunikationsprozess erfüllt, erlangt er tendenziell die Befähigung, den Wahrheitsgehalt der aufgerufenen Antwort je nach gestellter Frage zu beeinflussen, ferner Glaubwürdigkeit und Vertrauen in das Zeugnis wankend zu machen und in letzter Konsequenz einen Verlust der Authentizität des Berichteten herbeizuführen (vgl. Byford 2014: 65; Krämer 2008: 227).

Die Aufgabe von Fragevariationen muss es daher sein, das System so zu trainieren, dass nur diejenigen Fragen zu einem Aufruf einer Antwort führen, die in möglichst vielen semantischen Aspekten mit den ursprünglich gestellten Fragen übereinstimmen. Dies ist immer dann der Fall, wenn die konkret gestellten Fragen und die aus den vorhandenen Daten abgeleiteten Muster bedeutungsähnlich sind (z. B. „Wie heißt du?“ und „Wie+heißen+Sie/du“[6]). Ferner müssen die generierten, neuen Variationen von einer solchen Qualität sein, dass sie trotz ihrer Abweichung von der Ursprungsfrage zwar selbst

[6] Dieses Beispiel dient lediglich der Veranschaulichung und entspricht keinem der tatsächlich verwendeten Muster.

zu keiner *radikalen*[7] Veränderung des Wahrheitsgehaltes der Botschaft führen, dennoch in einer solchen Menge vorhanden sind, dass sie die Bildung des Sprachmodells bzw. das Sprachverstehen des KVA maßgeblich positiv beeinflussen können. Dies bedeutet, dass einem Muster der beiden Fragen „Wie heißen Sie?“ und „Wie heißt du?“ zwar auch die Frage „Wie ist Ihr Name?“ zugeordnet werden soll, nicht aber eine Frage wie „Wie heißt Ihre Frau?“. Das System muss folglich lernen, was die exkludierenden Kriterien sind, um bestimmte Fragen trotz vorhandener Teilüberstimmungen einer Antwort nicht zuzuordnen.

Ein Modell zur Generierung von Fragenvariationen bzw. Trainingsdaten aus linguistischer Sicht

Das im Juni 2019 entwickelte Modell ist das Resultat aus einem ersten, explorativen Versuch, für die ca. 1.000 Ursprungsfragen von Abba Naor Trainingsvarianten zu generieren. Hierbei wurde den für die Variationen zuständigen Mitarbeiter*innen keinerlei Vorgaben gemacht, auf welche Art die Ursprungsfragen zu variieren seien, was zu einem mehr oder minder intuitiven Vorgehen bei der Erstellung der Variationen führte. Bei diesem ersten, wenig systematischen Generierungsversuch wurden vor allem zwei Aspekte deutlich: Einerseits wurden oftmals Kombinationsmöglichkeiten zur Generierung von neuen sprachlichen Mustern außer Acht gelassen (vor allem die Kombination von bedeutungsähnlichen Phrasen untereinander vgl. Tabelle 1), andererseits ließen sich Unterschiede im Grad der Abweichung von der Ursprungsfrage feststellen. Angereichert wurden diese Erkenntnisse aus der explorativen Phase anschließend durch Analysen der Thema-Rhema-Gliederung unter Miteinbezug der gegebenen Antwort des Zeitzeugen oder der Zeitzeugin sowie der bereits beschriebenen ethischen Standards. Das Hauptziel, das damit hauptsächlich verfolgt wurde, war die Generierung einer möglichst hohen Anzahl an Fragenvarianten, die allerdings noch deutlichen Bezug zur ursprünglichen Frage aufweisen sollten.

Ferner ist darauf hinzuweisen, dass das entwickelte Modell weniger darauf abzielt, sämtliche syntaktische Strukturen der erwartbaren Fragen vorherzusagen als den KVA auf die möglichen Wortkombinationen und die Ähnlichkeit bestimmter sprachlicher Muster im Bereich der Semantik zu trainieren. Daneben war es ein Anliegen, dass sich

[7] In einem strengen Sinne verändern sprachliche Variationen immer (zumindest) leicht den Wahrheitsgehalt der ursprünglich gegebenen Aussage, mögen die generierten Variationen auch über noch so eine große Bedeutungsähnlichkeit verfügen.

die erzeugten Varianten nach Möglichkeit ihrer Modalität nach im Gegensatz zu den teils langen und konzeptionell schriftlich gestalteten Ursprungsfragen eher an den Charakteristika der konzeptionellen Mündlichkeit orientierten.

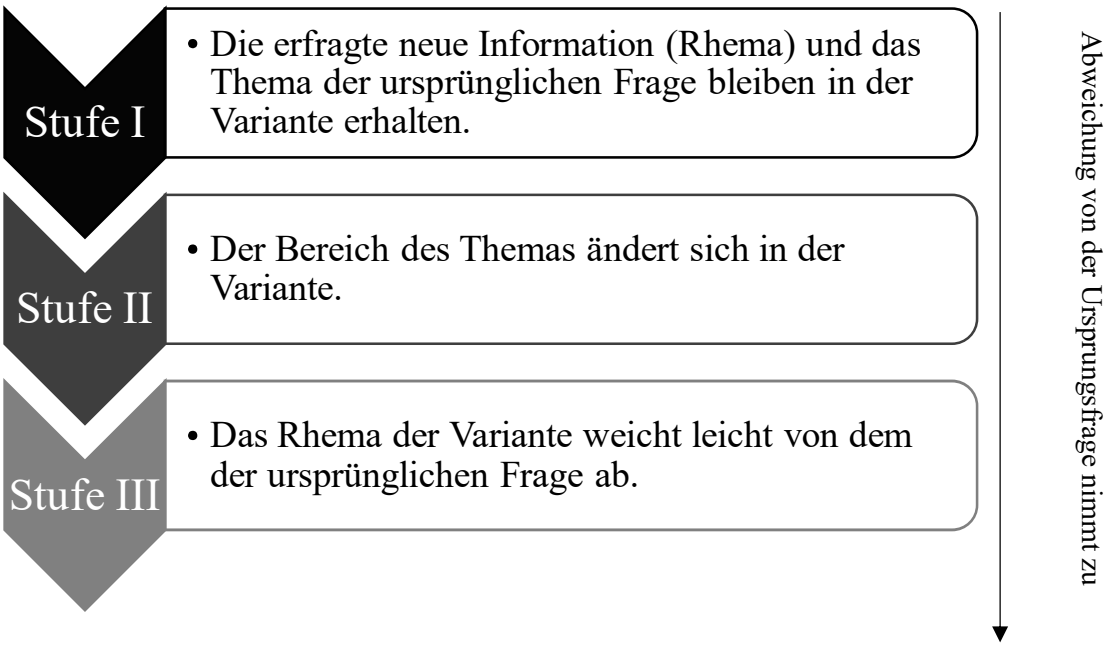

Abbildung 3 – Dreistufiges Modell zur Generierung von Fragenvarianten.

Das entwickelte Modell bewertet Fragenvarianten nach dem Grad der Abweichung vom ursprünglichen Thema und vom Rhema der zu erwartenden Antwort. Dabei werden die Ursprungsfragen ab Stufe II hinsichtlich ihrer Informationsstruktur (nach Thema und Rhema) bewertet. „Thema“ und „Rhema“ meinen dabei in Bezug auf Ammann die Unterscheidung zwischen dem Gegenstand einer Frage und der diskursneuen Information, die mit dieser Frage erfragt werden soll (vgl. Ammann 1928: 3). Hierbei ließe sich diese dichotomische Unterscheidung nach heutigem sprachwissenschaftlichen Kenntnisstand noch spezifizieren und ausdifferenzieren (vgl. bspw. Groot 1981; Krifka 2008), jedoch würde damit aus pragmatischer Sicht in Hinblick auf die Generierung von Fragenvariationen kein Nutzen einhergehen, da die Fragen bereits in ihrer Ursprungsform aus einem fest definierten thematischen Feld entstammen und somit die Themenfelder der Fragen bereits bekannt waren und nicht erst noch individuell erschlossen werden mussten (vgl. Ballis et al. 2019: 429f.). An dieser Stelle ist ebenfalls hervorzuheben, dass die Zeitzeug*innen die ihnen gestellten Fragen nie allein mit „ja“ oder „nein“, sondern stets kontextgebunden beantworten sollten.

Stufe I des Modells fokussiert die Findung von bedeutungsähnlichen Wörtern und Phrasen sowie das Training des KVA auf bedeutungsneutrale Nebensätze. Hierzu wurden die nachfolgenden Operationen durchgeführt, wobei der Hauptfokus auf das Finden von synonymen[8] Ausdrücken gelegt wurde:

I a) Umstellung der Wortfolge/Ergänzung der Frage durch neutrale Phrasen, die nicht zu einer Änderung des Fragegehaltes führen (z. B. „Können Sie uns sagen …“)
I b) Änderung der Sprachebene: „Hat man Ihnen bei der Arbeit in Utting **Gewalt angetan**?“ → „Wurden Sie bei der Arbeit in Utting **vermöbelt**?“
I c) Tempuswechsel: „**Konnten** Sie **sich** im Ghetto Kaunas frei **bewegen**? “ → „**Haben** Sie **sich** im Ghetto Kaunas frei **bewegen gekonnt**?“
I d) Ersetzung des Verbs durch eine bedeutungsähnliche Variante (z. B. „fliehen“ durch „Flucht ergreifen“)
I e) Ersetzung der Substantive durch bedeutungsähnliche Varianten (z. B. „Ehefrau“ durch „Frau“)
I f) Ersetzung des Interrogativpronomens/Interrogativadverbs durch eine bedeutungsähnliche Variante (z. B. „Wieso“ durch „Weshalb“)

In Stufe II wurde der thematische Bereich der Frage verändert, d. h. der Themenbereich der Ursprungsfrage wurde entweder vergrößert oder verkleinert:

II a) Der erfragte thematische Bereich vergrößert sich:
Bsp.: „Hatten Sie **Haustiere**?“ (bezieht sich nur auf als Haustiere gehaltene Tiere) → „Hatten Sie **Tiere**?“ (umfasst ebenso Nutztiere wie bspw. Hühner oder Kühe)
II b) Der erfragte thematische Bereich verkleinert sich:
Bsp.: „Hatten Sie **Haustiere**?“ → „Hatten Sie einen **Hund**?“ (Es wird nicht mehr nach dem Hyperonym, sondern nach einem Hyponym bzw. Token dieses ursprünglichen Hyperonyms gefragt)

In Stufe III ändert sich die erfragte Information leicht; dieses neue Rhema ist jedoch häufig auch bereits in dem ursprünglich Erfragten enthalten:

[8] Wird im Nachfolgenden das Wort „synonym“ verwendet, so meint dies stets „bedeutungsähnlich“ und nicht „bedeutungsgleich“.

Bsp.: „Wie war es in Stutthof?“ (Rhema: Beschreibung der äußeren und/oder inneren Zustände in Stutthof) → „Wie fühlten Sie sich in Stutthof?“ (Rhema: Beschreibung des inneren Zustandes in Stutthof)
Ob eine Variante der Stufe III möglich ist, lässt sich nur unter Berücksichtigung der gegebenen Antwort entscheiden. So wäre es wenig sinnvoll, die Frage nach den Gefühlen des Zeitzeugen in Stutthof einer Antwort zuzuordnen, in der lediglich die Beschaffenheit des Lagers beschrieben wird.

Mit diesem Modell der systematischen Generierung konnten im Durchschnitt pro Ursprungsfrage ca. 263 Fragenvarianten generiert werden. Dieser Mittelwert ist stark von der Länge der Ursprungsfrage abhängig: So kann der Wert bei kurzen Fragen wie bei „Wie heißen Sie?“ im niedrigen ein- bis zweistelligen Bereich liegen, wohingegen aus längeren Ursprungsfragen (z. B. „Hat man Ihnen gesagt, was mit den Kindern, die im Lager von Kaunas zusammengetrieben worden waren, passierte?“) zwischen 700 und 1.000 Varianten generiert werden können. Letzterer Wert ergibt sich aus einem sprunghaften Anstieg, der daraus folgt, dass alle bedeutungsähnlichen Wörter einer Phrase a aus einer hypothetischen Frage ab wiederum mit allen Varianten der Phrase b inklusive a verbunden wurden, um auf diese Weise die Gesamtheit aller möglichen Kombinationsmöglichkeiten dieser Varianten zu erfassen:

Ursprungsfrage	**Variationen nach dem ersten Austausch bei jeweils vier gefundenen Synonymen für a**	**Variationen nach dem zweiten Austausch bei jeweils vier gefundenen Synonymen für b**
ab	a1b, a2b, a3b, a4b ∑ **4**	ab1, a1b1, a2b1, a3b1, a4b1 ab2, a1b2, a2b2, a3b2, a4b2 ab3, a1b3, a2b3, a3b3, a4b3 ab4, a1b4, a2b4, a3b4, a4b4 ∑ **20**

Tabelle 1 – Darstellung des Anstiegs bei der Generierung von Fragevarianten der Stufe I.

Fazit – Besondere Anforderungen an die Datengenerierung für digitale interaktive Zeugnisse

Im Vorangegangenen wurde die Funktionsweise der Sprachverarbeitung und des Sprachverstehens des KVA *Google Dialogflow* dargelegt. Im Zuge dessen wurde deutlich, dass es sich bei einer lernenden Maschine zugleich um eine zu unterrichtende

Maschine handelt, die Datensätze von einer bestimmten Qualität benötigt, um die bedeutungstragenden Teile einer Frage zu erkennen und aus diesen die Regeln für eine Zuordnung abzuleiten. Die Notwendigkeit der Datengenerierung in Form von Fragevariationen ergab sich dabei aus der Zielsetzung des Projektes LediZ, in der eine freie Interaktion ohne Vorgabe der zu stellenden Fragen angestrebt wurde. Ohne diese Zielvorstellung wären Fragevarianten für die Zuordnung von Spracheingabe zu Videoaussage nicht notwendig gewesen und es könnte ebenso anstelle eines KVA ein streng regelbasierter Algorithmus verwendet werden, der allein darauf ausgelegt wäre, die Ursprungsfragen wiederzuerkennen.

Das entwickelte Modell stellt einen Vorschlag dar, wie eine möglichst große Anzahl an Datensätzen für ein digitales interaktives Zeugnis generiert werden kann, ohne sich dabei zu sehr von der ursprünglich an die Zeitzeug*innen gestellten Frage zu entfernen. Mit diesem Ansatz geht der Wunsch einher, die „Authentizität“ des Zeugnisses und den Wahrheitsgehalt der Antwort zu erhalten. Ein Element, das sich bei *Google Dialogflow* neben der Datensatzmenge maßgeblich auf die korrekte Zuordnung von Frage und Antwort auswirkt, ist der Schwellenwert der Konfidenz des Zeugnisses. Dieser Wert kann, sofern er zu niedrig eingestellt ist, selbst bei einer großen Menge an Variationen dazu führen, dass Frage-Antwort-Paare allein nach einzelnen Schlagworten zugeordnet werden. Fragenvariationen führen folglich zunächst nur dazu, einen Pool an Kombinationsmöglichkeiten von sprachlichen Mustern anzubieten und das System auf diese Musterverbindungen zu trainieren. Wie groß die Übereinstimmung zwischen Eingabe und den Mustern sein muss, wird erst durch den Schwellenwert festgelegt. Demnach sind die Fragenvariationen im Münchener Projekt nur ein Bestandteil des Trainings des auf Machine Learning basierenden KVA, den es mit dem Schwellenwert der Konfidenz auf Grundlage empirischer Ergebnisse in Einklang zu bringen gilt, um auf diese Weise das digitale Zeugnis so zu trainieren, dass es auf möglichst viele der gestellten Fragen eine in der Biografie der Zeitzeug*innen verbürgte Antwort gibt.

Literatur

Aggarwal, Charu C. *Machine learning for text*, Cham: Springer, 2018.

Ammann, Hermann. *Die menschliche Rede. Sprachphilosophische Untersuchungen. Teil II (Der Satz)* (Vol. 1928), Lahr: Schauenburg, 1928.

Ballis, Anja und Michele Barricelli. „Educational Issues on 3D-Testimonies in the German Language – A Research Report“, Workshop, Konferenz „The Holocaust in Europe. Research Trends, Pedagogical Approaches, and Political Challenges“, München, 6. November 2019.

Ballis, Anja, Michele Barricelli und Markus Gloe. „Interaktive digitale 3-D-Zeugnisse und Holocaust-Education – Entwicklung, Präsentation und Erforschung“, in: *Holocaust Education Revisited. Wahrnehmung und Vermittlung. Fiktion und Fakten. Medialität und Digitalität*, Anja Ballis und Markus Gloe (Hg.), Wiesbaden: Springer VS, 2019, 403–436.

Bitter, Christian, David Elizondo und Yingjie Yang. *Natural language processing: a prolog perspective*, Dordrecht: Springer Netherlands, 2010.

Brüning, Christina Isabel. „Holocaust Education in Multicultural Classrooms. Some Insights into an Empirical Study on the Use of Digital Survivor Testimonies“, in: *Holocaust Education Revisited. Wahrnehmung und Vermittlung. Fiktion und Fakten. Medialität und Digitalität*, Anja Ballis und Markus Gloe (Hg.), Wiesbaden: Springer VS, 2019, 391-402.

Byford, Jovan. „Remembering Jasenovac: Survivor Testimonies and the Cultural Dimension of Bearing Witness“, in: *Holocaust and Genocide Studies* 28, 1 (2014), 58-84.

Challenor, Jennifer und Minhua Ma. „A Review of Augmented Reality Applications for History Education and Heritage Visualisation”, in: *Multimodal Technologies and Interaction* 3, 2 (2019), 39.

Chomsky, Noam. *Syntactic structures*, S'-Gravenhage: Mouton, 1957.

Google. „Intent-Zuordnung“, https://cloud.google.com/dialogflow/es/docs/intents-matching, zuletzt geprüft am 23. Juni 2020.

Groot, Casper de. „On Theme in Functional Grammar“, in: *Perspectives on functional grammar*, H. Teun Hoekstra et al. (Hg.), Dordrecht: Foris, 1981, 75-88.

Kersting, Kristian, Christoph Lampert und Constantin Rothkopf. *Wie Maschinen lernen*, Wiesbaden: Springer, 2019.

Krämer, Sybille. *Medium, Bote, Übertragung*, Frankfurt a.M.: Suhrkamp, 2008.

Krifka, Manfred. „Basic notions of information structure“, in: *Acta Linguistica Hungarica* 55, 3–4 (2008), 243–276.

Lobin, Henning. *Computerlinguistik und Texttechnologie*, Stuttgart: UTB GmbH, 2010.

Mainzer, Klaus. *Künstliche Intelligenz – Wann übernehmen die Maschinen?*, Berlin: Springer, 2019.

Reyes Ochoa, Charlos Roberto, David Garza, Leonardo Garrido, Victor De la Cueva, Jorge Adolfo Ramirez Uresti. „Methodology for the Implementation of Virtual Assistants for Education Using Google Dialogflow“, 2019, https://www.researchgate.net/publication/336827160_Methodology_for_the_Implementation_of_Virtual_Assistants_for_Education_Using_Google_Dialogflow, zuletzt geprüft am 23. Juni 2020.

Sabharwal, Navin und Amit Agrawal. *Cognitive Virtual Assistants Using Google Dialogflow*, Neu-Delhi: Apress, 2020.

Schlobinski, Peter. *Grammatikmodelle*, Wiesbaden: Westdeutscher Verlag, 2003.

Serban, Iulian Vlad, Ryan Lowe und Peter Henderson. *A Survey of Available Corpora for Building Data-Driven Dialogue Systems*, 2015.

Sessions, Valerie und Marco Valtorta. „The Effects of Data Quality on Machine Learning Algorithms, ACM Journal of Data and Information Quality“, 1, 3, Artikel 14, (Dezember 2009).

Stucki, Toni, Sara D'Onofrio und Edy Portmann. *Chatbots gestalten mit Praxisbeispielen der Schweizerischen Post*, Wiesbaden: Springer Vieweg, 2020.