Laura Dreessen
im Interview mit Katharina Makosch

Die derzeitige Voice-Technologie-Branche aus Sicht einer Linguistin

In Ihrer Arbeit in der Voice-Technologie-Branche beschäftigen Sie sich unter anderem mit der Konzeption von Sprachdialogsystemen. Wie sind Sie dazu gekommen?

Ich bin ursprünglich promovierte Linguistin. Ich habe Deutsch, Englisch und Italienisch studiert und mich im Rahmen meiner Promotion in englischer Sprachwissenschaft mit Abstraktion in der Sprache, also mit theoretischer Computerlinguistik, beschäftigt. Computerlinguistik ist ein Anwendungsgebiet der Linguistik. Sie beschäftigt sich damit, wie man große Mengen an Sprachdaten sammeln, abstrahieren und beschreiben kann.

Nach meiner Promotion im Jahr 2015 bin ich in der Automobilindustrie gelandet und habe dort mit meiner Arbeit im Bereich der Voice-Technologie angefangen: die natürlichen Sprachdaten von realen Nutzenden aufzubereiten und sie so für die Maschine vorzusortieren, dass sie sie verstehen kann. Das ist die Grundlage für jegliches Dialogdesign, weil ein Dialog nicht stattfinden kann, wenn das Gegenüber nicht verstanden wird.

In weiteren freiberuflichen Projekten kam ich schließlich in den Bereich Voice-UX-Design, der sich damit beschäftigt, die eigentlichen Dialoge aufgrund der technologischen Basis und der Merkmale der jeweiligen Sprache zu gestalten. Weil Sprachassistenzsysteme immer prominenter wurden, bin ich aus der Automobilbranche gewechselt und baue jetzt mit VUI.agency eigene Assistenzsysteme. Wir sorgen also nicht nur dafür, die bestehenden großen KI wie Google und Alexa um Fähigkeiten zu erweitern und dabei die Marke eines Kunden zu repräsentieren. Viel häufiger arbeiten wir an eigenständigen Assistenzsystemen in Europa, die mehrere Sprachen sprechen und verstehen können. Dadurch habe ich einen ganzheitlichen Blick auf verschiedene Technologien mit Themen wie Datenschutz und Sound Branding gewonnen und arbeite jetzt bei VUI.agency mit 35 anderen Linguist*innen und UX-Designer*innen zusammen. Gemeinsam sind wir ein Team aus sehr vielseitig orientierten Expert*innen, die für eine ganzheitliche Experience mit der Maschine sorgen.

Welche Anforderungen gibt es an einen guten Dialog mit einer Maschine? Fällt nonverbale Kommunikation dabei weg?

Für uns fällt die nonverbale Kommunikation nicht wirklich weg, weil wir uns immer an menschlicher Konversation mit all ihren Facetten orientieren. Bei VUI.agency haben wir zum Beispiel eine bestimmte Idee geprägt, die sich „auditives Charisma" nennt. Der Gedanke dahinter ist: Wenn eine Person, mit der ich gerne interagieren möchte, in einen Raum kommt, dann hat diese Person eine ganz bestimmte Präsenz. Nicht nur durch den Klang ihrer Stimme und durch ihre Art, sich auszudrücken, sondern eben auch durch nonverbale Aspekte. Und genau deshalb sind wir nicht ausschließlich auf Voice fokussiert – es geht um eine sogenannte multimodale Experience. Das bedeutet, dass wir versuchen, technologisch nachzuempfinden, wie man solche nonverbalen Signale erkennen und umsetzen kann. Das Nonverbale wird zum Beispiel durch bestimmte Sounds, die ganz klar mit dieser Marke oder einem bestimmten Anwendungsfall assoziiert sind, umgesetzt. Wir haben also auch außerhalb der Sprache bestimmte technische Möglichkeiten oder verschiedene Ausgabemedien zusätzlich zur Stimme.

Wenn ich etwas zeigen muss, designe ich ergänzend auch für den Screen. In den Bereichen Entertainment und Smart Home ist der größte Screen beispielweise der Fernseher im Wohnzimmer, der mit verschiedenen Geräten kombiniert wird. Das Assistenzsystem ist dabei in allen Geräten präsent, wodurch sein abstrakter Charakter über verschiedene Ausgangsmedien erhalten bleibt. Die Persona kann also entscheiden, wann sie etwas anzeigt und wann sie etwas per Sprache ausgibt; das heißt, während ich auf dem Screen etwas abbilde, das klar visuell ausgedrückt werden muss, habe ich zusätzlich noch eine Sprachrepräsentanz, die zum Beispiel auch nur ein einfacher Bestätigungs-Sound sein kann. So entsteht eine ganzheitliche Interaktion, die auf multimodalem Design basiert.

Ist ein guter Dialog also vom Anwendungsfall abhängig?

Auf jeden Fall. Wir sprechen hier von Use-Cases und denken das jeweilige Szenario mit: Sitze ich zuhause oder bin ich unterwegs und habe mehrere Menschen um mich herum? Es gibt natürlich Anwendungsfälle, die ich nicht mit den Menschen um mich herum teilen möchte. Oder kann ich davon ausgehen, dass das Gerät, über das ich das Assistenzsystem erreiche, nur für mich verfügbar ist? Wir denken immer die Konversationssituation mit, so wie ich als Mensch auch entscheiden würde, in

welcher Situation ich mich befinde und welche Art der Interaktion ich dafür wählen möchte.

Die Kommunikation soll also möglichst natürlich gestaltet werden. Wie lässt sich das umsetzen? Und gibt es, zum Beispiel auf technischer Seite, Einschränkungen?

Selbstverständlich gibt es noch viele Einschränkungen. Wir orientieren uns an natürlicher Interaktion, weil wir davon ausgehen, dass der Paradigmenwechsel vom letzten Interface zu Voice ein großer Schritt in Richtung intuitive Mensch-Maschine-Konversation ist. Die Spracheingabe ist sehr intuitiv, läuft vielfach schneller ab als manuelles Tippen und ist ein direktes Adressieren eines Wunsches gegenüber einer Maschine. An Natürlichkeit orientieren wir uns in dem Sinne, dass Sprechen als extrem intuitive Handlung technisch bestmöglich umgesetzt werden sollte.

Eine Limitation befindet sich eindeutig – und das auch völlig zu Recht – im Bereich Datenschutz. Die Spracherkennung und damit auch die Assistenzsysteme laufen immer dann am besten, wenn möglichst viele Daten zur Verfügung stehen. Deshalb sind uns amerikanische Konzerne auch so weit voraus, weil sie andere Datenschutzbestimmungen haben, weil Englisch nun einmal Weltsprache ist und weil sie auch schon seit vielen Jahren Daten sammeln. Das funktioniert bei uns Menschen genauso: Je mehr jemand liest oder mit Sprache konfrontiert wird, desto eloquenter wird diese Person auch. Technisch ist es im Moment noch so, dass ein System umso besser funktioniert, je mehr Daten gesammelt werden, und das beeinflusst auch das Design.

Beim sprachlichen Design eines Alexa-Skills haben sich beispielsweise bestimmte Sprachmuster durchgesetzt, die wir nicht so einfach ändern können. Seit ein paar Jahren werden Alexa-Skills immer mit einer sogenannten Utterance „Alexa, öffne XY“ geöffnet. „Öffne“ ist aber eigentlich nichts, was ich persönlich am Anfang eines Gesprächs sagen würde. Seit Jahren übernehmen Designer*innen dieses Interaktionsmuster sprachlich vom visuellen Interface, wo ich Dateien und Ordner auf dem Screen öffne. Wenn ich mich davon wegbewegen möchte, muss ich natürlich mein Design so anpassen, dass die Nutzenden nicht mehr „Öffne“ sagen.

Bei den großen Plattformen kommt noch dazu, dass ein besonders erfolgreicher Skill ein sogenannter „Native Skill“ werden kann, den ich nicht mehr separat öffnen muss, sondern direkt ansprechen kann. Wir haben beispielsweise zusammen mit dem Carlsen-Verlag einen Gute-Nacht-Geschichten-Skill zu den Pixi-Geschichten umgesetzt. Wenn ich also diesen Skill initial konzipiere, muss ich sagen: „Alexa, öff-

ne Pixi und erzähle mir eine Gute-Nacht-Geschichte“, was nicht besonders natürlich ist. Wenn dieser Skill aber erfolgreich ist und ein Native Skill wird, bekommt man eine sogenannte goldene Utterance, mit der ich einfach sagen kann „Alexa, erzähl mir eine Gute-Nacht-Geschichte“, und dann wird automatisch mein Skill aufgerufen. Das zeigt, wie wir einerseits technisch und andererseits auch von den größten Anbietern durch die Masse an Sprachdaten limitiert werden. Bei der Arbeit an eigenen Assistenzsystemen haben wir linguistisch gesehen die Möglichkeit, diese Interaktion von vornherein natürlicher zu gestalten. Das ist auch das Spannende daran, an eigenen Assistenzsystemen zu arbeiten. Damit kommt aber auch eine gewisse Verantwortung, das Ganze natürlich, intuitiv und trotzdem datenschutzkonform zu gestalten.

Sprachassistenzsysteme werden immer häufiger im Bezug darauf kritisiert, dass sie oft weiblich gegendert sind und damit die Implikation einer weiblichen Servicekraft einhergeht. Für wie wichtig halten Sie Diversität im Fokus auf Ihre Arbeit?

Diversität ist ein Thema, das uns ständig begleitet – auch, weil wir bei VUI.agency hauptsächlich Mitarbeiterinnen haben. Wir sind Teil einer internationalen Community namens „Women in Voice“, welche sich mit der Frage beschäftigt, wie wir in der Voice-Technologie-Branche für Gleichberechtigung sorgen können. Im Bereich der Technologie sind wir mit Diversität insofern konfrontiert, dass wir Assistenzcharaktere erschaffen. Meist geht es hier um die Repräsentation einer Marke. Im Endeffekt sitzen wir mit Menschen aus der Brand- und Marketingabteilung zusammen und überlegen: „Wie ist der Charakter der Marke?“ Wenn also dort definiert wird, dass der Charakter der Marke weiblich oder männlich ist, entsteht schon dort eine Diskussion. Dann würde ich mit dem Brand- und Marketingteam diskutieren, was an diesem Charakter denn genau männlich und weiblich ist.

Technologisch gesehen haben wir aber vielfältige Möglichkeiten, Assistenzsysteme gender-divers zu gestalten, zumal bestimmte Merkmale in einer Stimme oder Konversation absolut geschlechterunabhängig sind. Ein guter Gesprächspartner ist erst einmal ein Gesprächspartner, der versteht, eine gewisse Bildung hat, gut zuhört, sich auszudrücken weiß und sein Gegenüber in Erwägung zieht, bevor er etwas sagt. Das alles sind Parameter, die geschlechterunabhängig sind.

Darüber hinaus können wir die Stimmwahl technisch unabhängig gestalten, auch wenn sie meist von der Marke abgeleitet wird. Es gibt mittlerweile beispielsweise Ansätze dafür, neutrale synthetische Stimmen zu nutzen, die aber trotzdem von den

Hörenden immer wieder mehr mit der einen oder der anderen Kategorie assoziiert werden. Das heißt, es gibt durchaus eine kulturelle Grundlage dafür, dass wir in diesen Kategorien denken und nicht neutral bleiben. Wir haben aber durch unser Design und unser Bewusstsein für das Thema die Möglichkeit, unsere Kunden davon zu überzeugen, dass natürliche und intuitive Konversation auch geschlechterunabhängig laufen kann. So könnte man für eine Marke auch einfach einen kleinen Roboter gestalten, der kein Geschlecht hat und in Bezug auf Vorurteile völlig neutral bleibt, und dieser könnte trotzdem ein großartiges Interaktions- und Kommunikationserlebnis bieten. Aus technischer Perspektive gibt es also keinen Grund, einem Assistenzsystem ein bestimmtes Geschlecht zuzuordnen. Wir versuchen da Stück für Stück jeden Tag etwas zu verändern, und das geht nun einmal im Persona-Design besonders gut.

Noch einmal zurück zu „Women in Voice“: Dabei handelt es sich um ein Netzwerk, welches Frauen und Minderheiten in der Voice-Branche vernetzen und ihnen mehr Sichtbarkeit verleihen möchte. Weshalb besteht dafür eine Notwendigkeit?

Wir alle haben einen sehr persönlichen Ansatz dazu. Ich selbst komme wie gesagt aus der Automobilbranche, wo ich oft als einzige Frau für die Voice-Technologie zuständig war. Bevor ich zu „Women in Voice“ kam, habe ich mir selbst lange gar nicht bewusst gemacht, dass es in der Technologiebranche Unterschiede zwischen den Geschlechtern gibt. Durch die Webinare und (momentan virtuellen) Treffen im Rahmen dieser Community wird uns allen die Existenz solcher Vorurteile überhaupt erst bewusst. Ich hatte dann später in meiner Karriere auch den umgekehrten Fall, dass ich mit Frauen in bestimmten Situationen oder Positionen viel mehr aneinandergeraten bin und mich gefragt habe, wo das überhaupt herkommt. Wir haben aufgehört, Unsicherheiten zu leugnen, um uns behaupten zu können, und versuchen nun, das Beste aus der Situation zu machen und zu überlegen, was uns eigentlich stört, worin wir uns schwach fühlen oder worin wir uns unterscheiden.

So geht es zum Beispiel auch um Dinge wie „Wie gut bin ich darin, spontan meine Meinung über Social Media zu präsentieren?“ Wir haben festgestellt, dass wir einen gewissen Anlauf und ein bisschen Mut dafür brauchen, den wir uns gegenseitig zusprechen können. Das wird dank unserer Community mittlerweile besser, weil sie weltweit agiert und immer weiter wächst, weil sich immer mehr Menschen, mehr Frauen, aber auch mehr Männer, zugehörig fühlen und diese Themen offen ansprechen. Ich denke, wir müssen zeigen, dass diese Probleme existieren und dass es keinen Grund gibt, aus Angst für sich allein zu bleiben, so wie es eigentlich überall

ist. Man findet Stärke vor allem dadurch, dass man die Themen zusammen angeht. Das gilt auch für die Voice Community.

Diversität tritt ja noch in verschiedenen weiteren Formen auf. Bei vielen Sprachen gibt es eine „Standardsprache", wie zum Beispiel Hochdeutsch. Sprachassistenzsysteme sprechen tendenziell eine solche Standardsprache. Gibt es dafür einen bestimmten Grund?

Von Standardsprachen gibt es die meisten Datensammlungen. Zudem sollen Sprachassistenzsysteme ja für eine möglichst breite Masse nutzbar sein. Wir hören häufig die Frage, warum das Assistenzsystem keine Dialekte erkenne. Wir bei VUI.agency hatten die Möglichkeit, hier etwas zu verändern, da wir das erste Assistenzsystem mitkonzipiert haben, das Schweizerdeutsch versteht. Der aus der Zusammenarbeit entstandene „Hey Swisscom"-Assistant beherrscht fünf verschiedene Sprachen, unter anderem auch Schweizer Dialekte, und war das erste unabhängige Schweizer Assistenzsystem. Es musste also eine kleine Nutzer*innengruppe, die entsprechend für Sprachdaten sorgt, priorisiert werden, was nur möglich ist, wenn man nicht von einem rein quantitativen Ansatz ausgeht. Das ist allerdings, wie gesagt, im Moment noch nicht so verbreitet.

Wenn man das Konzept Dialekt weiterdenkt, kommt man aus linguistischer Sicht zum Idiolekt, das ist meine ganz persönliche Art, mich auszudrücken, zu sprechen und zu formulieren. Alexa versteht mittlerweile auch ein paar Abweichungen, zumindest in der Aussprache. Aber was ist auf dem deutschen Markt zum Beispiel mit Menschen, die mit einem anderen Akzent oder einer akzentuellen Färbung sprechen? Die Spracherkennung kann auch da noch nicht alles. Es wäre mein Wunsch, dass man da in Zukunft breiter denkt.

Mit zunehmenden technischen Möglichkeiten werden synthetische Stimmen immer menschenähnlicher. Zum Beispiel im Fall von Google Duplex gab es Kritik an der Entwicklung, weil Google Duplex in Testläufen bei Restaurants und bei Friseurgeschäften angerufen hat und die angerufenen Menschen nicht erkannt haben, dass sie mit einem Assistenzsystem sprechen. Stellt eine so große Menschenähnlichkeit eine Gefahr dar?

Synthetische Stimmen werden heutzutage immer besser. Audio-Fakes und Stimmfakes stellen eine gewisse Gefahr dar, aber aus meiner Designperspektive heraus

kann ich nur sagen: Alle, die an solchen Systemen arbeiten, sollten dafür sorgen, dass, auch wenn es eine menschliche Stimme auf der anderen Seite gibt, anhand der konkreten Interaktion deutlich wird, dass es sich nicht um einen Menschen handelt. Es wäre aus Designperspektive kein Problem, an irgendeiner Stelle zu sagen „Hallo, ich bin ein digitales Assistenzsystem." Wenn wir also mit sehr menschlichen Stimmen arbeiten, würde ich dazu tendieren, zusätzliche Hinweise in der Konversation zu hinterlassen, die eine Unterscheidung möglich machen. Ich weiß, dass Audio-Fakes an vielen Stellen zum Einsatz kommen – klar, die Möglichkeiten sind erschreckend. Wenn ich aber von meinem eigenen Anwendungsgebiet spreche, und vom Design solcher Assistenzinteraktionen, bin ich der Meinung, dass wir eine Verantwortung haben, entsprechende Hinweise zu hinterlassen.

Richard David Precht sagt, was künstliche Intelligenz von menschlicher unterscheide, sei, dass wir Menschen immer in der Lage sind, aufgrund unserer Erfahrung und unserer Emotionen in kleinsten Details in jeglicher Situation eine neue Entscheidung zu treffen, einen neuen Kontext anzuwenden. Eine Maschine kann das zwar bis zu einem gewissen Grad, denn wir arbeiten ja auch damit, dass sie den Kontext der Konversation kennen muss und auf die Situation eingeht. Sie wird aber niemals in der Lage sein, aufgrund von Erfahrung und Gefühl entscheiden zu können, was sie antwortet. Je mehr ich über mein Gegenüber, nämlich die Maschine, im Allgemeinen weiß, desto eher werde ich sie erkennen können. Und das ist meine Verantwortung als Designerin, diese kleinen Hinweise in der Interaktion zu hinterlassen. Wenn ich der Maschine also eine menschliche Stimme gebe, dann muss ich an anderer Stelle klarstellen, dass sie kein Mensch ist.

*Gibt es neben der Verantwortung von Entwickler*innen in der Gesellschaft weitere Möglichkeiten, Aufklärung zu leisten?*

Das ist der große Kontext der Frage, wie unsere reale Welt im Digitalen abgebildet wird. Das ist natürlich durch Situationen wie die momentane Pandemie noch einmal viel aktueller geworden. Aus unserer Perspektive weiß ich, dass jegliche Technologie in unserem Bereich nur dadurch funktioniert, dass Sprachdaten gesammelt werden. Insgesamt muss es also Aufklärung oder zumindest ein Bewusstsein dafür geben, wie viele Daten wir wo hinterlassen und dass ich selbst entscheiden muss, wie und in welcher Form ich sie hinterlasse. Ob ich sie vielleicht auch selbst manipuliere, ob sie immer Auskunft darüber geben, wer genau ich bin, ob sie immer mein Dasein abbilden, ob sie immer alles Echte ins Digitale übersetzen, oder ob ich ab und zu auch darauf verzichte, sie zu hinterlassen.

Wie oft klickt man auf Geschäftsbedingungen, die man nie durchgelesen hat, und wie oft erhält man auffällig passende Werbeanfragen und wundert sich, woher sie kommen? Ich würde gerne dafür sorgen, Menschen diese Zusammenhänge bewusster zu machen, und versuche das auch in meinem Kontext. Den Menschen muss bewusst sein, dass sie dadurch zwar einerseits mehr Komfort in der Nutzung erleben, aber andererseits durch vermehrte Datensammlung eben auch z.B. Audio-Fakes entstehen können, weil alles, was lernbar ist, heutzutage gesammelt wird. Ich möchte Nutzenden immer wieder ermöglichen, den Unterschied zwischen uns emotionalen menschlichen Individuen und unseren Werkzeugen, den Maschinen, zu erkennen.

Der vorliegende Aufsatz entstammt der Publikation

Marcus Erbe / Aycha Riffi / Wolfgang Zielinski (Hrsg.)
Mediale Stimmentwürfe
Perspectives of Media Voice Designs

Schriftenreihe Digitale Gesellschaft NRW, Bd. 7
Kopaed Verlag, 2022

ISBN 978-3-96848-642-0