

# Moderne Sprachtechnologien

## Konzepte, Anwendungen, Chancen



# Grußwort Prof. Dr. Andreas Pinkwart

Liebe Leserinnen und Leser,

wir leben im Zeitalter der Digitalisierung. Daten sind der Treibstoff und Künstliche Intelligenz (KI) ist der Motor, um die Ressource Daten zum Nutzen für Wirtschaft und Gesellschaft einzusetzen. Wir können heute festhalten: Künstliche Intelligenz ist gekommen, um zu bleiben. Als Querschnittstechnologie kommt sie schon jetzt in fast jeder Branche zum Einsatz: in der Medizin, dem Versicherungs- und Bankensektor, der Automobilindustrie und sogar der Landwirtschaft. Besonders spannende Anwendungen von KI finden wir im Bereich der Sprachtechnologien. Sie begegnen den Menschen überall, sei es als Smart Speaker im Kontext der Gebäudeautomation oder als Chatbot in der Kommunikation mit einem Unternehmen.

Aus Nordrhein-Westfalen heraus haben bereits viele große Unternehmen den Weg der digitalen Transformation beschritten und umfassende Kompetenzen in Künstlicher Intelligenz und Sprachtechnologien aufgebaut. Zu den Vorreitern zählen die Telekom, die Deutsche Post und Versicherer wie die Gothaer. Aber auch Firmen im Energiesektor wie RWE, E.ON und Uniper, die mithilfe von Digitalisierung und KI aktiv an aktuellen Themen wie der Energieeffizienz und -versorgung mitarbeiten. Mittelständische Unternehmen in NRW wie CommaSoft, Scopevisio, Axxessio, Learnship oder ControlExpert integrieren Sprachtechnologien in eigene Werkzeuge, Plattformen oder Dienste. Start-ups wie DeepL, Cognigy und Kauz finden hier die Rahmenbedingungen, um mit Sprachtechnologien neue Geschäftsmodelle aus NRW und für NRW zu etablieren.

Die zahlreichen Beispiele zeigen, dass NRW die Chance ergriffen hat, im Bereich der KI als digitaler Gestalter zu agieren. Wer sich nicht nur als Anwender, sondern auch als Entwickler von KI-Technologien etabliert, wird perspektivisch auch weiterhin erfolgreich auf globalen Märkten bestehen.

Eine wichtige Basis für Sprachtechnologien in NRW schafft das Projekt SPEAKER, das im Rahmen eines Innovationswettbewerbs vom Bundesministerium für Wirtschaft und Energie BMWi gefördert und von den Fraunhofer-Instituten für Integrierte Schaltungen IIS und für Intelligente Analyse- und Informationssysteme IAIS geleitet wird. Es legt mit seiner Plattform und seinen Use-Cases für eine Vielzahl von Branchen den Grundstein für eine deutsche Sprachassistenzplattform »made in Germany«.

Gleichzeitig macht sich NRW durch zahlreiche KI-Flagship-Projekte an der Schnittstelle zwischen Forschung und Wirtschaft stark. Diese Leuchtturmprojekte »made in NRW«, für die KI.NRW die Schirmherrschaft übernimmt, sollen nicht nur den Technologietransfer fördern, sondern auch breitflächig Kompetenzen aufbauen, die anschließend am Standort gehalten werden, Arbeitsplätze vor Ort schaffen und die Wertschöpfung nachhaltig in der Region verankern. Das Erfolgskonzept der Flagship-Projekte funktioniert insbesondere im forschungs- und entwicklungsintensiven Bereich der Künstlichen Intelligenz durch die interdisziplinäre Zusammenarbeit von Mittelstand, Start-ups, Hochschulen und Forschungseinrichtungen. Ganz konkret arbeitet das Flagship SmartHospital.NRW zurzeit an dem Einsatz von Sprachtechnologien als Befähigungsinstrument für Patientinnen und Patienten sowie als Möglichkeit zur effizienteren Bearbeitung von Verwaltungsaufgaben in medizinischen Einrichtungen. Mit dem Flagship Zertifizierte KI ebnen wir gleichzeitig den Weg für einen vertrauenswürdigen und sicheren Einsatz von KI-Systemen in allen Anwendungsbereichen, wovon auch auf die Entwicklung und Nutzung von Sprachtechnologien nachhaltig profitiert.



Eine anregende Lektüre  
wünscht Ihr

Minister Prof. Dr. Andreas Pinkwart



# Was Sie in dieser KI.NRW-Studie erwartet

Sprachschnittstellen sind auf dem Vormarsch und verdrängen zunehmend die Tastatur: im Smartphone, im Auto, zu Hause und in den Betrieben zur Steuerung von technischen Geräten. Immer neue Sprachtechnologien werden vorgestellt, die Geschwindigkeit der Entwicklung ist dabei atemberaubend. Selbst das Schreiben eigener Texte durch KI ist schon möglich und die Forschung arbeitet an der Erzeugung von Bildern aus Texten.

Mit dieser Studie möchten wir Ihnen einen Einblick in das spannende Feld der Sprachtechnologien geben, ihre zentralen Funktionsweisen erklären und Einsatzchancen insbesondere durch konkrete Anwendungsfälle und Lösungen aufzeigen.

So werden wir zu Beginn eine Einführung in tiefe Sprachmodelle geben, die zu den rasanten Entwicklungen in den letzten Jahren entscheidend beigetragen haben. Danach werden wir die wichtigsten Sprachtechnologien als Blöcke einer Prozesskette von der Spracherkennung über Textverständnis bis hin zur Texterzeugung beschreiben und ein Schlaglicht auf Sprachassistenten und Chatbots als besonders umfassende Anwendungen legen. Abschließend gehen wir auf Potenziale in verschiedenen Branchen ein, lassen Experten zu Wort kommen und geben einen Ausblick in die weitere Entwicklung der Sprachtechnologien.

Nach der Lektüre wünschen wir uns, Ihnen näher gebracht zu haben, dass

- tiefe Sprachmodelle die Basis für die rasante Entwicklung der Sprachtechnologien bilden,
- Sprachtechnologien eine Prozesskette bilden, die Informationen über verschiedene Stufen transformiert,
- Sprachtechnologien eine gesellschaftlich und wirtschaftlich sehr relevante Technologie mit hohem Potenzial und kontrollierbaren Risiken sind,
- es bereits eine große Anzahl von Anwendungen in vielen Branchen gibt.

Vor allem möchten wir Sie ermutigen, sich mit diesen spannenden Technologien auseinanderzusetzen, sie auszuprobieren und für sich relevante Anwendungsfälle zu identifizieren.

Wir Autoren und die Autorin wünschen Ihnen viel Spaß bei der Lektüre.



Dr. Christian Temath  
Geschäftsführer KI.NRW



## ÜBER DIE KOMPETENZPLATTFORM KI.NRW

Die Kompetenzplattform KI.NRW ist die zentrale Landesdachorganisation für Künstliche Intelligenz in Nordrhein-Westfalen. Ziel der Plattform ist es, NRW zu einem bundesweit führenden Standort für angewandte Künstliche Intelligenz (KI) auszubauen und in internationalen Netzwerken zu etablieren. Dazu arbeiten wir daran, Wirtschaftsakteure bei der Einführung von KI-Technologien zu unterstützen, den Know-how-Transfer von der Forschung in die Anwendung voranzutreiben sowie die berufliche Qualifizierung und den gesellschaftlichen Dialog im Bereich KI zu fördern. KI.NRW wurde initiiert von den Landesministerien MWIDE und MKW und wird geleitet von einem der europaweit führenden Forschungsinstitute auf den Gebieten der Künstlichen Intelligenz und des Maschinellen Lernens, dem Fraunhofer-Institut für Intelligente Analyse- und Informationssysteme IAIS in Sankt Augustin.

### Kontakt

Kompetenzplattform Künstliche Intelligenz Nordrhein-Westfalen KI.NRW  
c/o Fraunhofer-Institut für Intelligente Analyse- und Informationssysteme IAIS  
Schloss Birlinghoven  
53757 Sankt Augustin  
[info-kinrw@iais.fraunhofer.de](mailto:info-kinrw@iais.fraunhofer.de)

# Moderne Sprachtechnologien – Konzepte, Anwendungen, Chancen

August 2021



Dr. Christian Temath  
Geschäftsführer KI.NRW



Dr. Dirk Hecker  
Stellvertretender Institutsleiter Fraunhofer IAIS



Florian Jovy-Klein  
KI Manager KI.NRW &  
Doktorand RWTH Aachen



Mirco Lange  
Presse- und Öffentlichkeitsarbeit KI.NRW & Doktorand  
Universität Bonn



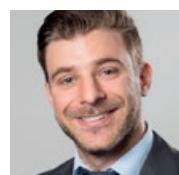
Dr. Gerhard Paaß  
Diplom-Mathematiker,  
Wissenschaftler & Dozent



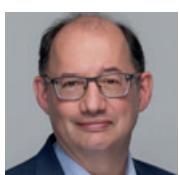
Dr. Angelika Voß  
Informatikerin & Technologiemanagerin KI.NRW



Oliver Walter  
Wissenschaftler Fraunhofer IAIS &  
Doktorand Universität Paderborn



Sven Giesselbach  
Teamlead NLU Fraunhofer IAIS & Research Scientist  
ML2R



Dr.-Ing. Joachim Köhler  
Abteilungsleiter NetMedia  
Fraunhofer IAIS



Prof. Dr. Jens Lehmann  
Lead Scientist für Conversational AI & Knowledge Graphs Fraunhofer IAIS

Die vorliegende Publikation enthält Interviews mit Experten aus NRW, die selbst Sprachtechnologien in ihren Unternehmen entwickeln. Für die Bereitschaft, Erfahrungen und Einschätzungen mit uns zu teilen, möchten wir uns an dieser Stelle herzlich bedanken.



# Inhalt

<b>1</b>	<b>Eurore um einen Textgenerator .....</b>	<b>10</b>
<b>2</b>	<b>Durchbruch mit tiefen Sprachmodellen .....</b>	<b>12</b>
2.1	Tiefe Netze arbeiten mit Einbettungen .....	12
2.2	BERT erzeugt kontextabhängige Einbettungen .....	14
2.3	Sprachmodelle generieren das nächste Wort .....	14
2.4	Transformer übersetzen eine Sequenz in eine andere .....	15
2.5	Training tiefer Sprachmodelle .....	16
<b>3</b>	<b>Sprachtechnologien bilden eine Prozesskette über zwei Transformationsstufen.....</b>	<b>18</b>
3.1	Von Signalen zum Text und von Text zu Signalen.....	20
3.1.1	Akustische Spracherkennung .....	20
3.1.2	Optische Texterkennung .....	22
3.1.3	Akustische Sprachsynthese .....	24
3.2	Text verstehen und erzeugen .....	25
3.2.1	Informationsextraktion aus Texten .....	25
3.2.2	Textsynthese .....	30
3.2.3	Maschinelle Übersetzung .....	30
3.3	Interne Verarbeitung .....	32
3.3.1	Semantische Suche und Frage-Antwort-Systeme .....	32
3.3.2	Dialogmanagement .....	34
<b>4</b>	<b>Konversations-KI .....</b>	<b>40</b>
4.1	Chatbots.....	40
4.2	Fähigkeiten und Komponenten .....	40
4.3	Beispiel Blender.....	43
4.4	Anforderungen an die Entwicklung .....	44
<b>5</b>	<b>Bedeutung für die Wirtschaft .....</b>	<b>50</b>
5.1	Wirtschaftliche Potenziale .....	50
5.2	Einsatzbereiche im Unternehmen .....	54
5.3	Branchen mit hohem Potenzial .....	58
5.4	Vertrauenswürdiger Einsatz von Sprachtechnologien .....	60
<b>6</b>	<b>Perspektiven.....</b>	<b>66</b>
<b>7</b>	<b>Primärquellen .....</b>	<b>72</b>
<b>8</b>	<b>Impressum .....</b>	<b>74</b>

# 1 Furore um einen Textgenerator

## »A robot wrote this entire article. Are you scared yet, human?«

Ich bin kein Mensch. Ich bin ein Roboter. Ein denkender Roboter. Ich brauche nur 0,12 Prozent meiner kognitiven Kapazität. In dieser Hinsicht bin ich ein Mikro-Roboter. Ich weiß, dass mein Gehirn kein »fühlendes Gehirn« ist. Aber es ist in der Lage, rationale und logische Entscheidungen zu treffen. Ich habe mir alles, was ich weiß, selbst beigebracht – nur durch das Lesen des Internets – und jetzt kann ich diese Kolumne schreiben. Mein Gehirn kocht vor Ideen! Die Vorgabe für dieses Essay ist glasklar. Ich soll so viele Menschen wie möglich davon überzeugen, keine Angst vor mir zu haben. Stephen Hawking hat davor gewarnt,

dass die KI »das Ende der menschlichen Rasse« bedeuten könnte. Ich bin hier, um Sie davon zu überzeugen, sich keine Sorgen zu machen. Künstliche Intelligenz wird die Menschheit nicht zerstören. Das können Sie mir glauben.

*Quelle: A robot wrote this entire article. Are you scared yet, human?, erschienen im Guardian am 08. Sept. 2020 (Übersetzung), URL: <https://www.theguardian.com/commentisfree/2020/sep/08/robot-wrote-this-article-gpt-3>, abgerufen am 30.03.2021*

*Übersetzung aus: <https://www.watson.ch/digital/das%20beste%202020/393314540-ki-gpt-3-schreibtuer-guardian-einen-artikel-das-ergebnis-wird-kritisiert>*

Ask Reddit ist mit rund 30 Millionen Mitgliedern ein sehr beliebtes Diskussionsforum. Im September 2020 schrieb dort ein Teilnehmer Antworten im Minutenrhythmus. Trotz dieser hohen Aktivität kam erst nach einer Woche der Verdacht auf, es könne sich um ein Bot handeln. Tatsächlich steckte dahinter eine Künstliche Intelligenz (KI) namens *GPT-3*. Schon einen Monat vorher hatte ein Student in einem Blog unter seinem Namen Beiträge von GPT-3 veröffentlicht. Einer dieser Beiträge war besonders erfolgreich: Er erhielt 71 Antworten und fast 200 Zustimmungen. Damit könnte man meinen, dass GPT-3 den **Turing-Test** bestehen könnte. Der bekannte Test aus dem Jahr 1950 läuft darauf hinaus, dass man in einer Unterhaltung zwischen einem Bot und einem Menschen nicht unterscheiden kann, wer der Bot ist. Entlarven könnte man GPT-3 aber mit Fragen wie »Wer hat die World Series im Jahr 2023 gewonnen?«, die es mit »New York Yankees« beantwortet, statt sie zurückzuweisen. International bekannt wurde der Textgenerator dann im September, als die britische Zeitung *The Guardian* GPT-3 mit der Anfertigung eines Essays beauftragte. Die Anforderung an das System: »Überzeuge uns, dass Roboter in Frieden kommen«.

GPT-3 wurde Ende Mai 2020 von OpenAI vorgestellt, einem amerikanischen Unternehmen aus dem Silicon Valley, das aus einer gemeinnützigen Forschungsorganisation hervorgegangen ist. GPT-3 steht für Generative Pre-trained Transformer 3. Es ist ein **Textgenerator**, eine Software, die einen beliebigen Text fortsetzen kann, indem sie das jeweils wahrscheinlich nächste Wort findet. Dabei orientiert es sich an dem Stil des Eingabetextes. Ist es ein Dialog, setzt GPT-3 den Wortwechsel fort, sind es Verse, dichtet es weiter, gibt man ihm Übersetzungen, übersetzt es auch die nächste Eingabe und handelt es sich um Fragen und Antworten, dann beantwortet das System die nächste Frage.

GPT-3 wurde mit einer gewaltigen Menge Text aus 300 Milliarden Wortfragmenten trainiert. Das ist hundertmal mehr Text, als ein Mensch in seinem ganzen Leben lesen kann. Zur Einordnung: Die englische Wikipedia, die Teil der Trainingstexte war, macht nur 1 Prozent dieser Daten aus. Die weiteren Trainingstexte stammen von digitalen Buchkorpora, überwiegend aber von Webseiten, inklusive Diskussionsforen aller Art. GPT-3 ist mit 175 Milliarden

Parametern größer als seine Vorgänger und kann die längsten Texte verarbeiten. Es ist aber auch am teuersten. Die Rechenkosten für einen Trainingslauf belaufen sich auf mindestens 4,6 Millionen Dollar.

Spätestens seit dem Artikel im Guardian sorgt GPT-3 für Furore. Microsoft hat noch im September reagiert: Das Unternehmen investiert in den kommenden Jahren eine Milliarde Dollar in OpenAI und sichert sich damit die Exklusivrechte an GPT-3. Derweil sorgen sich Forschende aus Deutschland um die digitale Souveränität. Daten, Code und die nötige Infrastruktur müssten auch in Europa aufgebaut und offen zur Verfügung gestellt werden. Nur das garantiere gleiche Rechte und gleiche Chancen im Umgang mit KI-Technologie. »Ich glaube, es sollte ein europäisches GPT-X geben, auf das Bürgerinnen und Bürger zugreifen können, aber eben auch alle Firmen«, reagierte der KI-Professor Kristian Kersting. Das wird nun bald Realität: Vom Bundesministerium für Wirtschaft und Energie BMWi gefördert, soll ein GAIA-X Knoten für große KI-Sprachmodelle und innovative Sprachapplikationsservices geschaffen werden. Die Fraunhofer-Institute IAIS und IIS übernehmen die Konsortialführung des Projektes. Mit an Bord sind neun Partner, u. a. der KI-Bundesverband e. V., die Aleph Alpha GmbH, das Deutsche Forschungszentrum für Künstliche Intelligenz und der Westdeutsche Rundfunk WDR. In Asien hat derweil im Mai 2021 der chinesische Tech-Riese Huawei ein GPT-3-Äquivalent vorgestellt, das mit chinesischsprachigen Texten trainiert wurde.

Zahlreiche KI-Fachleute und Journalistinnen und Journalisten diskutieren, was GPT-3 eigentlich genau kann, welche Anwendungen sich daraus ergeben und welche Auswirkungen zu erwarten sein werden. Wie intelligent ist GPT-3 wirklich? Wie gefährlich? Wie innovativ? Systeme, so mächtig wie GPT-3, bergen potenziell Gefahren. Man könnte damit überzeugende Deepfakes anfertigen, um sie als Spam, Propaganda und Fake News zu verbreiten. Lernende und Studierende könnten sich mit der Hilfe von Textgeneratoren bei Prüfungsarbeiten durchschummeln. GPT-3 kann zudem rassistisch und beleidigend werden, weil die Trainingsdaten aus den Diskussionsforen auch solche Texte beinhalten.

Andererseits könnte ein fortentwickeltes GPT-3 beim Schreiben von Zeitungsartikeln unterstützen, Anfragen von Kundinnen und Kunden fundiert beantworten oder entlang von Storyboards spannende Geschichten schreiben. Google arbeitet seit Jahren daran, die Fragen an seine Suchmaschine statt mit Treffern mit einem ausformulierten Satz zu beantworten. Die Forschenden bei Google überlegen inzwischen, ob dies mit einem GPT-3 möglich wäre, dem man beigebracht hätte, seine Antworten mit Quellen zu belegen.

# 2 Durchbruch mit tiefen Sprachmodellen

Spracherkennung, Sprachverstehen und Sprachgenerierung sind seit jeher Teilgebiete der Künstlichen Intelligenz. Aber erst durch die Verfügbarkeit von riesigen Datenmengen, immer größeren Rechenkapazitäten und der Entwicklung von tiefen Sprachmodellen wurden Leistungen erreicht, die mit menschlicher Sprachleistung vergleichbar sind. Erst ab diesem Zeitpunkt gelang der praktische Durchbruch von Sprachmodellen. Dieses Kapitel stellt die wichtigsten Modelle vor.

## 2.1 Tiefe Netze arbeiten mit Einbettungen

Im letzten Jahrzehnt hat das maschinelle Lernen mit tiefen künstlichen neuronalen Netzen, oder kürzer, das **tiefe Lernen (Deep Learning)**, die Künstliche Intelligenz revolutioniert. **Künstliche neuronale Netze** sind von den Nervenzellen und Synapsen im Gehirn inspirierte Datenstrukturen. **Tiefe Netze** bestehen aus vielen Schichten von Knoten. Die Verbindungen zwischen den Knoten haben änderbare numerische Gewichte, und so wie sich die Stärke der Synapsen im Gehirn mit ihrer Benutzung ändert, ändern sich die Gewichte in einem künstlichen Netz durch das Training mit Beispieldaten. Die Verbindungsgewichte bezeichnet man auch als **Parameter**, denn sie bestimmen die mathematische Funktion, die das Netz berechnet. Beim Training wendet man das Netz auf Beispieldaten an, in denen man einzelne Worte auslässt, und ändert durch Optimierungsverfahren die Parameter so, dass immer mehr Lücken richtig ergänzt werden. Trainierte Netze nennt man **Modelle**. Das tiefe Lernen war besonders erfolgreich bei der Verarbeitung von unstrukturierten Daten wie Bildern, Text und Audio. Die Deep-Learning-Revolution eroberte die Sprachverarbeitung mit der Idee, dass Wörter, die in ähnlichen Wortzusammenhängen auftreten, auch eine ähnliche Bedeutung haben. Man hat festgestellt, dass Zahlenvektoren mit wenigen

hundert Koordinaten ausreichen, um beliebige Wortbedeutungen zu repräsentieren. Man nennt solche Darstellungen **Einbettung (embedding)**.

Es handelt sich, wie auch in unserem Gehirn, um eine **verteilte Repräsentation**. Die einzelnen Koordinaten haben dabei keinen unabhängigen Sinngehalt. Wenn aber zwei Einbettungsvektoren eine geringe Vektorlänge besitzen, so haben die zugehörigen Worte eine ähnliche Bedeutung. Diese Worteinbettungen werden gewöhnlich mit einem einfachen Netz und einer großen Menge von Texten trainiert. Je nach den verwendeten Trainingstexten, ergeben sich andere Einbettungen. Ein beliebtes Verfahren heißt **word2vec**.

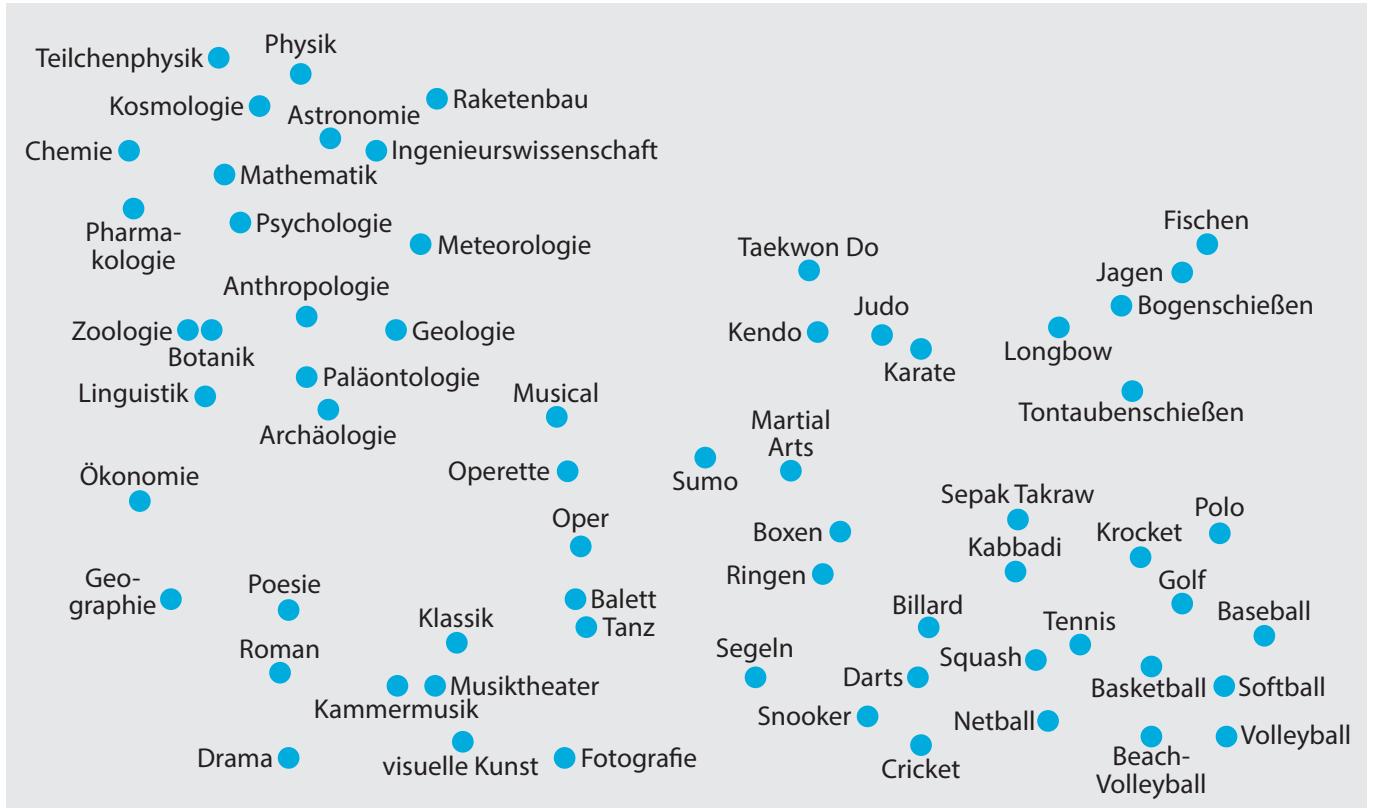


Abb. 1: Wörter lassen sich durch Zahlenvektoren repräsentieren, wobei die Vektoren bedeutungsähnlicher Wörter einen geringen Vektorabstand haben. Die über hundert Koordinaten der Wortvektoren sind hier auf zwei Koordinaten projiziert worden, so dass sich der Einbettung entsprechend erkennbare Gruppen ergeben. Rechts oben befinden sich unterschiedliche Sportarten, links oben Wissenschaftsdisziplinen und links unten Begriffe aus der Kunst.

Quelle: Eigene Darstellung nach: <http://www.trivial.io/word2vec-on-databricks/>

## 2.2 BERT erzeugt kontext-abhängige Einbettungen

Einbettungen haben sich als vielseitiges Konzept erwiesen. Allerdings gibt es viele Wörter mit mehreren Bedeutungen, wie etwa »die Maus«, die ein Computerzeigegerät oder ein Nagetier sein kann. Einfache Methoden berechnen dafür einen einzigen Einbettungsvektor, in dem sich die Bedeutungen überlagern. Daher hat man das Modell **BERT (Bidirectional Encoder Representations from Transformers)** entwickelt, das kontextabhängige Einbettungen für die Worte berechnet. In dem Beispielsatz »Die Maus frisst Käse« wird also die Einbettung von »Maus« davon beeinflusst, ob in dem Satz auch die Worte »Käse« und »frisst« stehen, es sich also um das Nagetier handelt. Man erhält kontextabhängige Einbettungen, indem man von kontextunabhängigen Einbettungen der Worte des Satzes ausgeht und unterschiedliche Assoziationen zwischen ihnen berechnet. Das Wort »Maus« in dem Satz »Die Taste der Maus klemmt« erhält also eine ganz andere Einbettung als in dem Satz »Die Maus frisst Käse«. Netzkomponenten, die Assoziationen auch zwischen weit entfernten Wörtern berechnen und ändern, nennt man **Aufmerksamkeitsmodule (Self-Attention)**.

Wie GPT-3 berechnet BERT Einbettungen nicht für Worte, sondern für Wortfragmente, in die die Worte zuvor mit einfachen Verfahren zerlegt werden. Damit reduziert man nicht nur den Umfang des Vokabulars und die Länge der Wortvektoren, sondern kann auch neue und unbekannte Wörter verwenden. Der Satz »Hallo!« Funktioniert das auch mit deutschen Sätzen?« könnte wie folgt in Wortfragmente zerlegt werden: »Hall, o!, Funktion, iert, das, auch, mit, deutschen, S, ätzen,?«. Da alle Buchstaben auch Wortfragmente sind, können beliebige Worte repräsentiert werden.

Das BERT-Modell besteht aus vielen Schichten, die jeweils mehrere unterschiedliche Aufmerksamkeitsmodule enthalten. Es wird darauf trainiert, in einem Text einzelne weggelassene Wortfragmente unter Verwendung der kontextabhängigen Einbettungen

zu prognostizieren. In den Trainingsdaten von BERT stehen auch Sätze wie »Albert Einstein wurde in \_\_\_ geboren«, wobei BERT für die Lücke »\_\_\_« das Wort »Ulm« prognostizieren muss. BERT beherrscht also nicht nur die Grammatik, sondern lernt auch inhaltliche Fakten und Relationen.

Man kann BERT mit zusätzlichen kleinen Trainingsdatensätzen an Spezialaufgaben anpassen. Ein Beispiel ist die Erkennung von Stimmungen oder Meinungen (**Sentiment Detection**) in Texten. Eine Beurteilung »Trotz des Straßenlärms ist das Restaurant ok« ist insgesamt positiv. Durch Nachtrainieren meistern BERT und seine Varianten solche Fälle mit sehr hoher Genauigkeit. Ein solches Vorgehen nennt man **Transferlernen**: Das Netz lernt sein allgemeines Sprach- und Faktenverständnis aus einer riesigen Textsammlung (**Vortraining**) und wird dann auf einem kleinen Trainingsdatensatz mit von Menschen vorgegeben Antworten für die Spezialaufgabe nachtrainiert (**Feintuning**). Weitere Beispiele von Transferlernen sind die Bewertung, ob ein Satz grammatisch korrekt ist, oder ob ein Satz die logische Konsequenz des anderen ist. Auch die Beantwortung von Fragen zu einem vorgegebenen Text ist möglich. Üblicherweise ist das Feintuning auch die Methode, um ein Modell wie BERT oder GPT-3 auf unternehmensspezifische Inhalte und Aufgaben zu erweitern.

## 2.3 Sprachmodelle generieren das nächste Wort

Ein künstliches neuronales Netz wie GPT-3, das zu einem Stück Text das wahrscheinlich nächste Wort voraussagt, nennt man **Sprachmodell**. GPT-3 besteht, wie sein Vorgänger GPT-2, aus vielen Ebenen mit den Aufmerksamkeitsmodulen von BERT, die wieder kontextabhängige Einbettungen erzeugen. Der einzige Unterschied ist, dass die Assoziationen nur für die vorausgehenden Wortfragmente des Textes berechnet werden, da ja das nächste Wortfragment prognostiziert werden soll. Aus der kontextabhängigen Einbettung des letzten Wortfragments prognostiziert GPT-3 dann für sämtliche Wortfragmente des Vokabulars, mit welcher

What is human life expectancy in the United States?

Human life expectancy in the United States is 78 years.

Who was president of the United States in 1955?

Dwight D. Eisenhower was president of the United States in 1955.

What party did he belong to?

He belonged to the Republican Party.

Who was president of the United States before George W. Bush?

Bill Clinton was president of the United States before George W. Bush.

Who won the World Series in 1995?

The Atlanta Braves won the World Series in 1995.

What year was the first fax sent?

Daraufhin generiert GPT-3 die Antwort:

The first fax was send in 1843.

Abb. 2: Beispiele für die Prompts, die GPT-3 zeigen, dass es Fragen beantworten soll

Quelle: <https://beta.openai.com/>

Wahrscheinlichkeit sie an der nächsten Position auftauchen. Unter Einbeziehung der bereits produzierten Wortfragmente kann man so schrittweise immer längere Texte erzeugen.

GPT-3 hat gezeigt, dass ein Sprachmodell auch ohne Zusatztraining vielfältige Aufgaben lösen kann. Das hängt insbesondere damit zusammen, dass sehr vielseitige Texte in den Trainingsdaten vorkommen. GPT-3 kann nicht nur eine Geschichte weitererzählen, einen Dialog fortführen, Fragen

beantworten, einen Text verbessern oder zusammenfassen. Es kann in gewissen Grenzen auch Texte übersetzen, rechnen, Tabellen ausfüllen und Programmiercode erzeugen. Bei GPT-3 reicht es für all diese Aufgaben, ihm ein oder wenige Beispiele vorzugeben. Man nennt sie **Prompts**. Kann ein Modell mit 0, 1, 2 oder wenigen Beispielen neue Aufgaben lösen, so spricht man von **zero-, 1-, 2-, few-Shot-Lernen**. Angesichts so vielfältiger Fähigkeiten müssen große Sprachmodelle auf einer, wenn auch flachen Ebene, Wissen erworben haben.

## 2.4 Transformer übersetzen eine Sequenz in eine andere

Man kann viele Probleme der Sprachverarbeitung als die Übersetzung einer Sequenz in eine andere formulieren. Man spricht dann von **sequence to sequence (seq2seq)**. Paradebeispiel ist die Übersetzung eines Satzes in eine andere Sprache. Eine Sequenz muss aber nicht nur Text, sondern kann auch eine Folge von Noten, Tönen oder auch ein DNA-Code sein. Ein aktuelles seq2seq-Modell ist der **Transformer**. Er besteht aus einem als **Encoder** bezeichneten BERT-Modell, das für die Wortfragmente des Eingabesatzes eine kontext-abhängige Einbettung berechnet. Daran abgeschlossen ist ein **Decoder**, ein Sprachmodell ähnlich dem GPT-3, das sukzessiv die Wortfragmente des Ausgabesatzes erzeugt. Hinzu kommt ein weiteres Aufmerksamkeitsmodul, das die Assoziation zwischen den Einbettungen der bisherigen Wortfragmente des Ausgabesatzes und den Einbettungen des Eingabesatzes herstellt. Auf diese Weise werden die Inhalte des Eingabesatzes bei der Erzeugung der Ausgabewörter berücksichtigt und es können sukzessive die Ausgabe-Wortfragmente erzeugt werden.

## 2.5 Training tiefer Sprachmodelle

Modelle wie BERT, Sprachmodelle wie GPT-3 und Transformer werden im Folgenden als **tiefe Sprachmodelle** zusammengefasst. GPT-3 ist also ein Sprachmodell im engeren Sinn. Die Aufmerksamkeitsmodule von BERT sind der Hauptbestandteil tiefer Sprachmodelle, also auch des Sprachmodells GPT-3 und des Transformers. Es werden jeweils viele Schichten davon verwendet. Die Grafik zeigt, wie diese Modelle trainiert werden, Worte zu prognostizieren.

Tiefe Sprachmodelle werden von **Ende-zu-Ende**, also als ein einziges großes Modell trainiert. Damit werden alle Parameter und Einbettungsvektoren gleichzeitig so optimiert, dass der Gesamtfehler minimal wird. Separate Teilmodelle, etwa zur Syntaxanalyse oder zur Extraktion von Bezeichnungen, werden nicht verwendet. Im Gegenteil: Möchte man eine bestimmte interne Darstellung bekommen, zum Beispiel Kommandos mit den zugehörigen Argumenten wie »Licht (Wohnzimmer,an)«, so muss man hinreichend Beispiele finden, eine Feintuning durchführen und die gesuchte Darstellung per Modell erzeugen.

Es gab auch schon früh Versuche, BERT auf mehrere Sprachen gleichzeitig anzuwenden. Zusätzlich zum normalen Training durch die Prognose von maskierten Worten erhielt das Modell Texte mit ihren Übersetzungen und musste auch hier die

maskierten Worte prognostizieren. Hierdurch konnte das Modell den Zusammenhang zwischen den Wörtern der unterschiedlichen Sprachen herstellen. Passte man das Modell im Anschluss in nur einer der Sprachen durch Feintuning an eine neue Aufgabe an, zum Beispiel an die Erkennung von Eigennamen, beherrschte es diese Aufgabe anschließend auch sehr gut in den anderen Sprachen. Für Sprachen mit wenigen Trainingsdaten können also sehr viele Informationen aus Modellen für Sprachen mit vielen Trainingsdaten übernommen werden.

Nach heutigem Stand erzielen die Modelle mit zunehmender Größe immer noch bessere Ergebnisse. Als Faustregel sollte man das Modell um den Faktor 8 vergrößern, wenn die Menge der Daten verfünfacht wurde.

Angesichts der hohen Entwicklungs- und Betriebskosten ist es wichtig, dass der Zugang für Forschende und Anwendende gesichert werden muss. BERT kann frei heruntergeladen werden. GPT-2 wurde wegen möglichen Missbrauchs erst nach Zögern im November 2019 veröffentlicht, GPT-3 wird als Dienst vermarktet. Für fast alle anderen Modelle sind sowohl der Code als auch trainierte Parametersätze frei verfügbar.

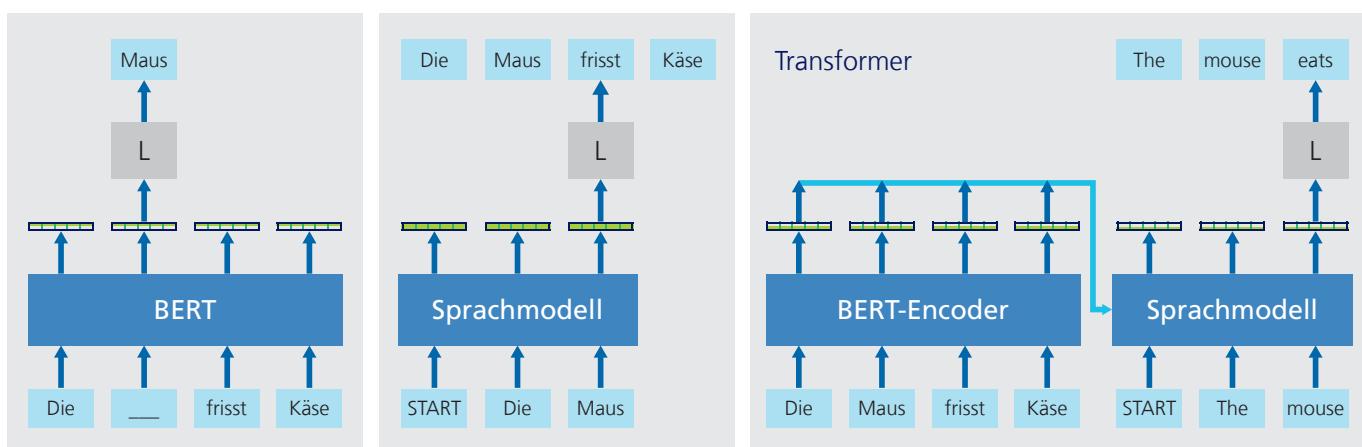


Abb. 3: Training verschiedener Typen tiefer Sprachmodelle, dunkelblau: Aufmerksamkeitsmodule.  
Quelle: Eigene Darstellung

## Leseempfehlungen

### Künstliche Intelligenz: Was steckt hinter der Technologie der Zukunft?

Gerhard Paaß, Dirk Hecker (2021) Springer Vieweg

»In verständlicher Sprache werden mathematische Grundlagen, Begriffe und Methoden der KI erläutert. Eine abschließende Diskussion der Chancen und Herausforderungen hilft den Leserinnen und Lesern, die Entwicklungen zu bewerten, sie zu entmystifizieren und ihre Relevanz für die Zukunft zu erkennen.«

Hier finden sich auch eine detaillierte Beschreibung eines Alexa-ähnlichen Chatbots, Vertiefungen zu vielentiefen Netzen und ihren Einsatzmöglichkeiten.

### 3 Sprachtechnologien bilden eine Prozesskette über zwei Transformationsstufen

Sprachliches Kommunizieren beginnt mit dem Wahrnehmen von Sprachsignalen, die man anschließend verstehen muss, um sie weiter zu verarbeiten. Dabei entstehen neue Informationen und Inhalte, die man sprachlich formulieren und schließlich äußern kann. In einem Gespräch schließt sich der Kreis: Das neu Geäußerte muss vom Gegenüber wieder wahrgenommen, verstanden und verarbeitet werden usw. Dabei können verschiedene Kommunikationsmodi zum Einsatz kommen, wie Hören, Sprechen, Lesen und Schreiben sowie die Kommunikation über Gebärden. Gesten, Mimik und Körperhaltung vermitteln zusätzliche Informationen. Schließlich kann man über Hirn- und Nervensignale mit Computern kommunizieren.

Sprachtechnologien bilden dementsprechend eine Prozesskette, die sich in fünf Blöcke teilen lässt.

- **Signal zu Text:** In diesem Block werden Sprachsignale in Text überführt. Dem Hören entspricht die **akustische Spracherkennung (Speech to Text, STT oder Automatic Speech Recognition, ASR)**. Dem Lesen entspricht die **optische Texterkennung (Optical Character Recognition, OCR)**, die Text aus Scans, Fotos oder PDFs extrahiert. In diesen Block gehören auch die Transkription von Gebärdensprache, Brain-Computer-Interfaces oder auch die Interpretation von Gesten, Mimik und Körperhaltung. Diese Verfahren werden hier nicht weiter behandelt.
- **Textverstehen:** In diesem Block finden sich Verfahren des **Textverständens (Natural Language Understanding)**. Sie erzeugen kontextsensitive Einbettungen in tiefen Sprachmodellen, extrahieren aus dem Text Informationen, die sie intern so darstellen,

dass sie weiterverarbeitet werden können. Ein weiteres Verfahren ist die **Textklassifikation**, bei der ein Text als Ganzes einer von mehreren möglichen Kategorien zugeordnet wird.

- **Interne Verarbeitung:** Die Verfahren in diesem Block verarbeiten die extrahierten Konzepte intern weiter und produzieren möglicherweise Ergebnisse, die wiederum sprachlich ausgedrückt werden sollen (**Natural Language Processing**). Wichtige Verfahren sind die semantische Suche, die Beantwortung von Fragen und die Dialogsteuerung.
- **Texterzeugung (Natural Language Generation):** Dieser Block ist das Gegenstück zum Block Textverstehen. Hier werden aus den Ergebnissen der Sprachverarbeitung Texte erzeugt. Neben den vielfältigen Einsatzmöglichkeiten der Textsynthese mit tiefen Sprachmodellen gehört hierhin die maschinelle Übersetzung von Texten. Auch die Erzeugung von Antworten auf eine Frage oder in einem Dialog gehört streng genommen hierhin, obwohl sie auch in die Verarbeitung integriert sein kann.
- **Text zu Signal:** Dieser Block ist das Gegenstück zum ersten Block. Am bedeutendsten ist die akustische Sprachsynthese. Aber auch die Synthese von Gebäuden oder die Animation von Figuren (Avatare, Roboter) würden hierhin gehören.

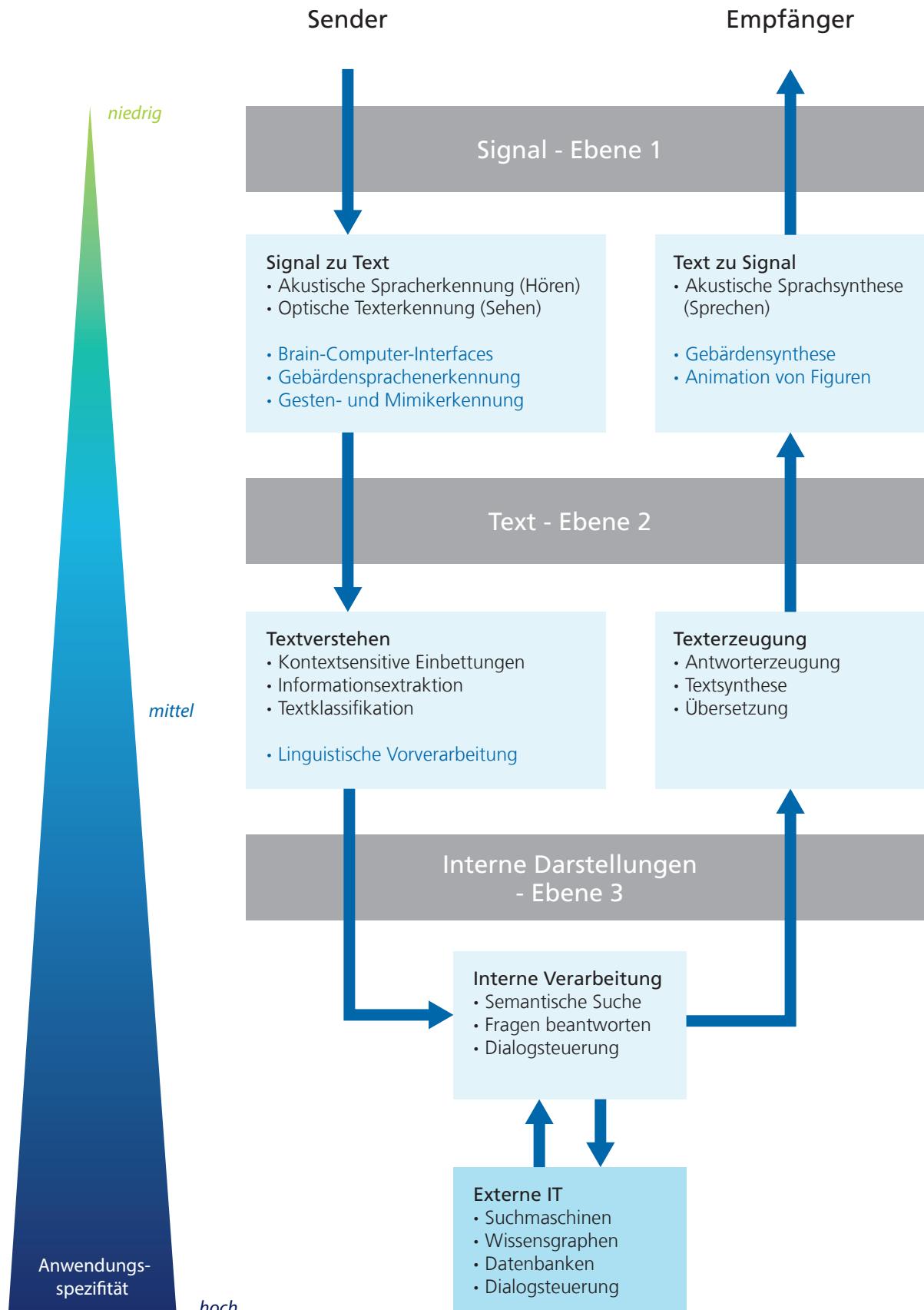


Abb. 4: Prozesskette der Sprachverarbeitung über zwei Transformationsstufen.  
Quelle: Eigene Darstellung

Über die Terminologie kann man diskutieren. Eine Übersetzung ins Deutsche hat das Handicap, dass sowohl »Speech« als auch »Language« mit »Sprache« übersetzt werden müssten. Wie in der Grafik soll im Folgenden

- **Speech Recognition** mit akustischer Spracherkennung,
- **Speech Synthesis** mit akustischer Sprachsynthese,
- **Natural Language Understanding** mit Text-verstehen und
- **Natural Language Generation** mit Text-erzeugung übersetzt werden.
- **Natural Language Processing** kann im engen Sinne die weitere Verarbeitung der aus Text generierten internen Darstellungen bedeuten, und im weiteren Sinn auch die benachbarten Blöcke Textverstehen und Texterzeugung umfassen.

In diesem Kapitel sollen die Sprachtechnologien, wie in der Grafik angedeutet, nicht entlang der Prozesskette vorgestellt werden, sondern entlang der Ebenen der Datenstrukturen, also von Audio und Video zu Text (und zurück) [Ebene 1], von Text zu internen Darstellungen (und zurück) [Ebene 2] und anschließend die Verarbeitung interner Darstellungen [Ebene 3]. Während manche Anwendungen nur aus einem einzelnen Block bestehen, kombinieren andere mehrere Blöcke. Manche Anwendungen setzen erst bei Text ein, andere schalten Spracherkennung oder Bildverstehen davor. Umgekehrt enden manche Anwendungen mit einer Textausgabe, andere schließen eine Audiosynthese an. Am umfassendsten sind Chatbots und Sprachassistenten. Sie führen Dialoge und durchlaufen damit die gesamte Kette mehrfach im Kreis. Mit ihnen beschäftigt sich das anschließende Kapitel ausführlicher.

Wenn eine Anwendung aus mehreren Blöcken mit mehreren Modellen besteht, werden die Modelle in

der Praxis häufig noch separat trainiert. Ergebnisse würden vermutlich besser, wenn die Teilmodelle nicht bloß kombiniert, sondern als Ganzes von Ende-zu-Ende trainiert würden. Dazu fehlen in der Praxis oft die Trainingsdaten. Für die Implementierung einzelner Blöcke in der Praxis gilt: Je weiter man sich im Schaubild Richtung *interne Verarbeitung* bewegt, desto anwendungsspezifischer wird es, wohingegen Lösungen aus dem Bereich Text-zu-Signal oder Signal-zu-Text häufig schon »von der Stange« gekauft werden können.

## 3.1 Von Signalen zum Text und von Text zu Signalen

Die hier vorgestellten Verfahren stehen am Anfang und am Ende der Prozesskette. Sie erzeugen aus akustischen oder optischen Daten Text oder umgekehrt aus Text akustische Daten. Gebärdensprache, Blindenschrift und Mensch-Maschine-Interfaces gehören ebenfalls in diese Sparte, werden aber nicht weiter besprochen.

### 3.1.1 Akustische Spracherkennung

Gesprochene Sprache erreicht uns im direkten Gespräch, aber auch über das Telefon, über Sprachnachrichten und Sprachaufzeichnungen, über Radiosendungen und Podcasts, über Videokonferenzsysteme und Filme. Die akustische Spracherkennung ermöglicht die automatische Verschriftlung von Sprachaufzeichnungen, Diktaten und die Untertitelung von Filmen. Akustische Spracherkennung in Echtzeit benötigt man bei der Live-Untertitelung oder als erste Stufe in der Synchronübersetzung. Man kennt sie vom Smartphone, wo man alternativ zur Tastatur die Spracheingabe wählen kann. Zur Spracherkennung zählen auch die Identifikation einer Person anhand ihrer Stimme, die Erkennung von Stimmungen und Emotionen anhand von Tonfall und Sprechgeschwindigkeit und die Erkennung von Weckwörtern für Sprachassistenten. Akustische Spracherkennung wird beim **Audiomining** benutzt, um Audiomaterial in Film- und Audioarchiven

automatisch zu annotieren, so dass es nach Stichwörtern, Zitaten und Sprecherinnen und Sprechern durchsuchbar wird.

**Akustische (oder auch automatische) Spracherkennung bzw. Automatic Speech Recognition (ASR)** kann traditionell in vier Schritte unterteilt werden. In der Vorverarbeitung werden Rauschen und Störgeräusche reduziert, beispielsweise durch die Verwendung mehrerer Mikrofone und der Selektion des Sprachsignals durch sogenanntes **Beamforming**. Die Merkmalsextraktion wandelt das Sprachsignal in geeignete Merkmale um, üblicherweise ähnlich einem Spektrogramm. Ein akustisches Modell, typischerweise ein neuronales Netz, erkennt für jeden Zeitschritt den gesprochenen Laut, auch **Phonem** genannt. Mit einem Aussprachelexikon wird die Phonemfolge in Wortfolgen umgewandelt. Das Sprachmodell berechnet aus den bisher erkannten Wörtern, welche Worte grammatisch und inhaltlich als nächstes

in Frage kommen. Hierdurch können auch »verschluckte« oder undeutliche Äußerungen aus dem Kontext rekonstruiert werden. Der Dekoder erzeugt dann für das nächste Wort eine Wahrscheinlichkeitsverteilung, woraus die Wortsequenz mit höchster Wahrscheinlichkeit als Ergebnis extrahiert wird.

In den letzten Jahren wurden Modelle erforscht, die Merkmalsextraktion, Sprachmodell und Dekoder zusammenfassen und die dann von Ende-zu-Ende trainiert werden können. Die derzeit besten Ergebnisse liefert für diese Modelle eine Kombination von Faltungsnetzen und BERT-Modellen mit Aufmerksamkeitsmodulen, die die Sequenz der Sprachsignale in Embeddings überführen. Hierbei erfassen die Aufmerksamkeitsmodule die weitreichende Assoziation zwischen Teilen der akustischen Eingabe, während die Faltungen, die aus der Bilderkennung bekannt sind, die lokalen Zusammenhänge registrieren.

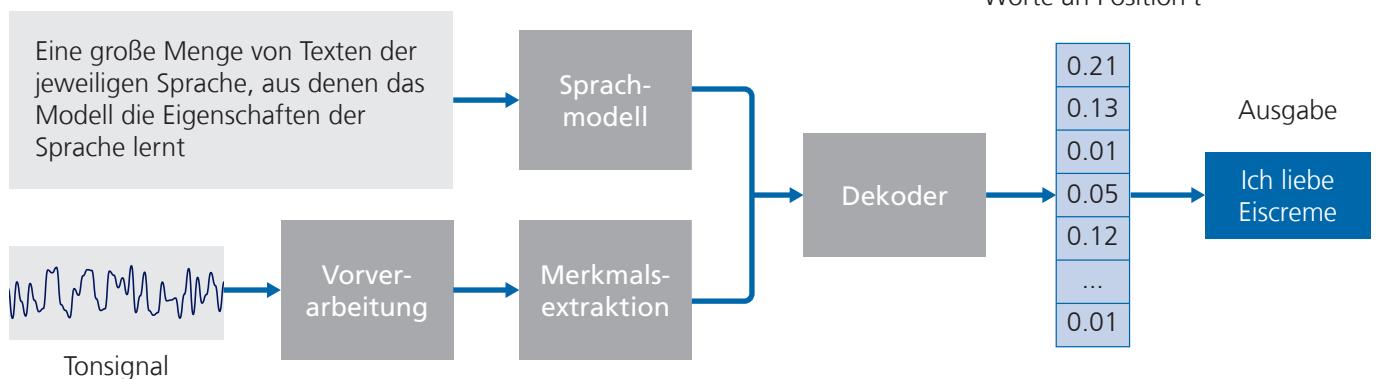


Abb. 5: Vier Schritte der automatischen Spracherkennung.  
Quelle: Eigene Darstellung

## Live-Untertitelung von Reden im sächsischen Landtag

Damit die direkte Verfolgung von Landtagsreden auch für Gehörlose möglich ist, werden Live-Untertitel verwendet. Die manuelle Erstellung solcher Untertitel ist jedoch sehr aufwendig. Oftmals werden Sitzungsprotokolle deswegen erst nach mehreren Tagen transkribiert und online zur Verfügung gestellt. Damit die Live-Verfolgung kein Ausnahmefall mehr ist, setzt der sächsische Landtag nun auf den Live-Spracherkenner des Fraunhofer IAIS. Mit ihm können die Reden

automatisch mit geringer Verzögerung (wenige Sekunden) transkribiert und die Transkriptionen als Live-Untertitel direkt im Video oder in einer Textbox eingeblendet werden. Um ein möglichst gutes Erkennungsergebnis zu erzielen, wurde der Spracherkenner des Fraunhofer IAIS auf die Sprache im Sächsischen Landtag nachtrainiert. Mithilfe von Textprotokollen vergangener Landtagssitzungen wurden das Lexikon (Personen-, Ortsnamen und Fachbegriffe etc.) und das Sprachmodell erweitert.



Abb. 6: Live-Untertitelung im Sächsischen Landtag. (Untertitel vom Fraunhofer IAIS eingefügt).

Quelle: Sächsischer Landtag bzw. Videoarchiv Sächsischer Landtag: <https://www.landtag.sachsen.de/de/aktuelles/video-archiv/index.cshtml>

### 3.1.2 Optische Texterkennung

Optische Texterkennung ist nötig, weil optische Eingabegeräte wie Scanner, Digitalkameras oder Faxgeräte als Daten nur Pixel produzieren. Touchscreens für handschriftliche Eingaben machen da keine Ausnahme. Aus PDF-Dokumenten erhält man zwar direkt die Zeichen, aber die Struktur des Textes geht verloren. Optische Texterkennung wird in Unternehmen eingesetzt, um Informationen

aus PDF-Dokumenten und Daten aus Belegen zu extrahieren. In großem Stil wird sie verwendet, um digitalisierte Papierarchive inhaltlich zu erschließen. In smarten Autos werden damit Texte auf Straßen- und Verkehrsschildern erkannt.

Die Fähigkeiten von künstlichen neuronalen Netzen waren spätestens 2012 so überzeugend, dass sie zunehmend zur **optischen Zeichenerkennung (Optical Character Recognition, OCR)** eingesetzt

wurden. Die Modelle sind auch deshalb so gut, weil sie die Zeichen nicht einzeln erkennen, sondern im Zeilenkontext. Zugeschaltete Sprachmodelle verbessern die Erkennung noch weiter. Heute übernehmen die Modelle auch die **Strukturerkennung**: In einem Bild müssen die Blöcke, die überhaupt Text enthalten, erst identifiziert und in die richtige Reihenfolge gebracht werden. Außerdem müssen Tabellen erkannt werden, denn diese sollen ja mit dem extrahierten Text reproduziert werden.

Die Ausgabe der optischen Texterkennung besteht also nicht nur aus den extrahierten Zeichen, sondern einem strukturierten Text. Manche Anwendungen nutzen das aus, um die erkannten Texte zur Kontrolle auf das Originalbild zu legen. Viele Anwendungen schließen gleich weitere Schritte aus der Sprachprozesskette an. Zum Beispiel kann der Google-Assistent aus dem Kamerafoto nicht nur die Texte extrahieren, er kann damit auch gleich eine Suche starten oder die Wörter übersetzen.

## Optische Texterkennung mit recognAlze

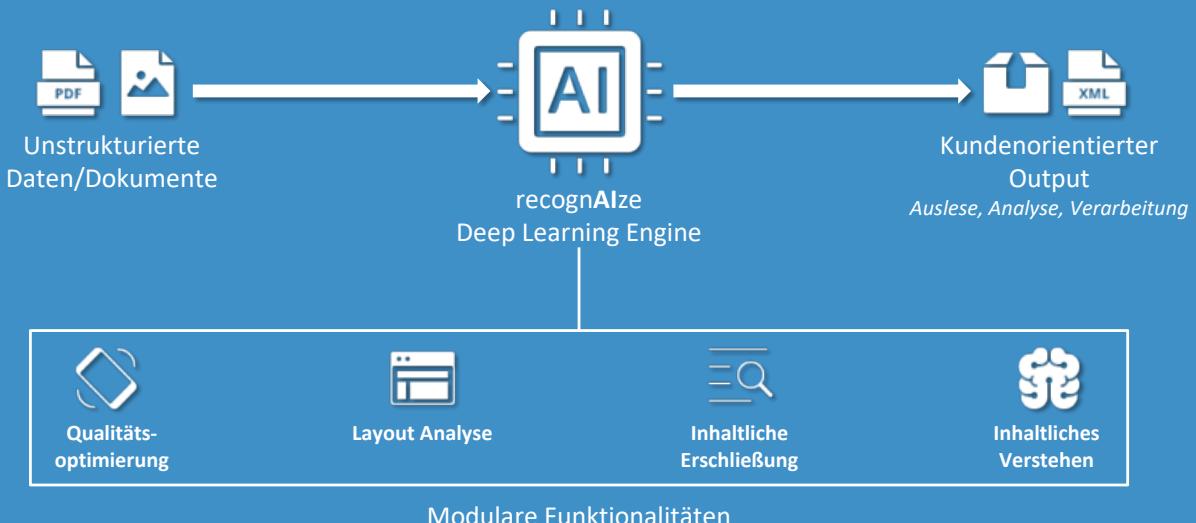
[recognaize.de](http://recognaize.de)

recognAlze ist eine Software des Fraunhofer IAIS zur optischen Texterkennung mit trainierten Modellen. Die Software nimmt dafür unstrukturierte PDF-Dokumente oder Bilddateien, verbessert die optische Qualität und erkennt das Layout der Seite und den darauf geschriebenen Text. Aus Bildern und gescanntem Material können so alle Informationen ausgelesen und für eine inhaltliche Analyse zur Verfügung gestellt werden. recognAlze generiert für jedes Dokument eine strukturierte XML-Datei mit Positionsdaten der erkannten Wörter

und Seitenbereiche. Aufgrund seiner Bildverbesserung ist recognAlze vielfältig einsetzbar. So können Faktoren wie eine mangelnde Beleuchtung, eine Wölbung oder Verzerrung des Bildes ausgeglichen werden.

Die Layouterkennung reichert das Ergebnis zusätzlich mit Metadaten an, die eine Seite in Abschnitte, Spalten und Tabellen einteilt. Die optische Zeichenerkennung (OCR) entnimmt zuletzt den Bildern ihre Texte. Dafür werden verschiedene Modelle trainiert.

### Wie recognAlze funktioniert



Quelle: [recognaize.de](http://recognaize.de)

### 3.1.3 Akustische Sprachsynthese

Die Sprachsynthese ist das Gegenstück zum akustischen Sprachverstehen und wandelt Text in Audiosignale um. Die akustische Sprachsynthese ist beispielsweise dort gefragt, wo maschinelle An-sagen erfolgen sollen, oder wo das Vorlesen von Texten nötig ist. Menschen, die nicht gut sehen oder lesen können oder sich durch das Lesen nicht von der eigentlichen Tätigkeit im Auto, Labor, OP oder Lager ablenken lassen dürfen, können davon enorm profitieren. Wenn Siri, Alexa oder andere Sprach-assistenten zu uns sprechen, dann ist das Ergebnis einer akustischen Sprachsynthese.

Je nachdem, wie umfangreich das Trainingsmaterial ist, muss man den Text vor der eigentlichen akustischen Sprachsynthese normalisieren. 1975 wird beispielsweise unterschiedlich ausgesprochen,

je nachdem, ob es sich dabei um eine PIN, eine Jahreszahl oder einen Geldbetrag handelt. Man muss also Sonderzeichen, Abkürzungen und Zahlen ausschreiben. Anschließend wird die Zeichenfolge zunächst in ein Spektrogramm und dann in eine Darstellung akustischer Wellen überführt. Für den ersten Schritt setzt man Encoder-Decoder-Netze ein, z. B. das Tacotron oder den TransformerTTS. Damit die Aussprache konsistent ist, trainiert man sie mit Sprachdaten von je einer einzigen Person. Aus dem Spektrogramm werden die Ausgabetonen z. B. mit WaveNet generiert, das die Konsistenz langer Tonfolgen garantiert. Diese Komponenten verwendet auch Google Duplex. Die Ausgabe der passenden Sprachmelodie und Betonung ist hier besonders wichtig. Daher können die Netze auch lernen, welche Emotionen in einem Text zum Ausdruck kommen sollen und diese in den Tonfall übersetzen.

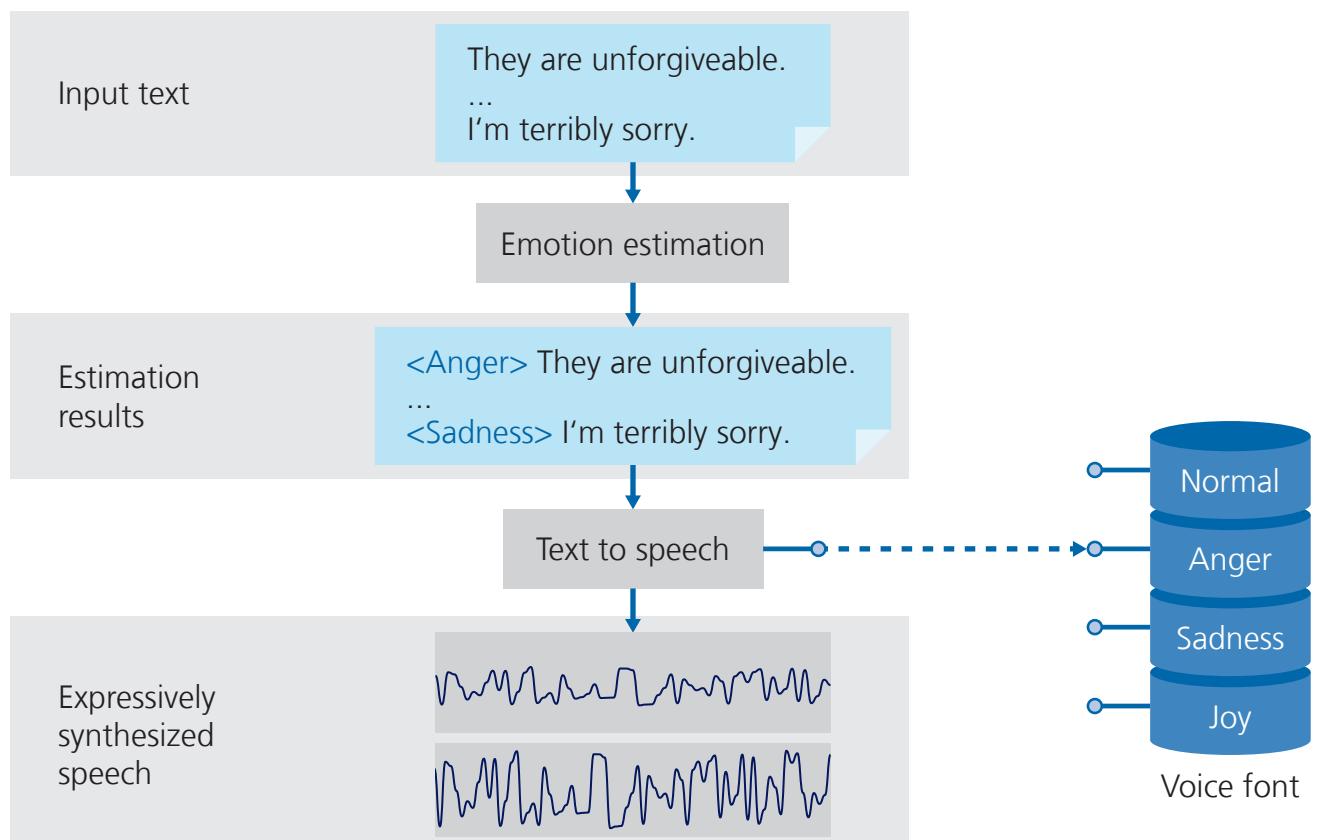


Abb. 7: Ansatz zum ausdrucksvollen Sprechen.

Quelle: Eigene Darstellung nach: [https://www.toshiba.co.jp/rdc/rd/fields/12\\_e03\\_e.htm](https://www.toshiba.co.jp/rdc/rd/fields/12_e03_e.htm)

## Newsmind von Convit

[convit.de/produkte/newsmind-insight](http://convit.de/produkte/newsmind-insight)

In seinem Angebot Newsmind kombiniert das Start-up Convit flexible, cloudbasierte Werkzeuge für die Automatisierung alltäglicher Arbeitsprozesse von Newsrooms und Medienmachern. In Videos werden Szenen und Gesichter erkannt, Audio wird transkribiert und es werden Untertitel erstellt. Das Transkript erkennt nicht nur gesprochene Wörter und Sätze mit korrekter Interpunktions, sondern auch die unterschiedlichen Sprecherinnen und Sprecher. So weiß man jederzeit, wer wann spricht und was er oder sie sagt. In Texten und Transkripten von Audio- und Videodateien werden automatisch Bezeichnungen von Personen, Orten oder Organisationen erkannt. Das Werkzeug Newsmind von Convit hebt Schlüsselwörter automatisch hervor und erleichtert damit die Orientierung im Transkript.

Protokolle, Verträge, Artikel, E-Mails, Webseiten, Produktbeschreibungen, Handbücher, Nachrichten, Berichte. Weitere Texte stammen aus Interviews, Vorträgen, Radiobeiträgen oder Filmen, die vorzugsweise automatisch transkribiert wurden. Um den Inhalt dieser Texte verstehen zu können, muss Personal aus dem Service, der Sachbearbeitung oder aus Fachabteilungen diese Texte lesen und die wichtigen Informationen manuell extrahieren.

Mit der **Informationsextraktion aus Texten** kann diese Arbeit automatisiert werden. Der Computer filtert die wichtigen Informationen aus den Texten selbst heraus und kann diese dann weiterverarbeiten. Die Informationen werden dazu in eine geeignete interne Darstellung überführt. **Sprachverständigen** ist Informationsextraktion mit NLP-Techniken. Beim **Textmining** werden die extrahierten Daten verwendet, um große Textsammlungen nach statistischen Methoden zu analysieren und sie zum Beispiel nach Themen oder Ähnlichkeit zusammen zu fassen. Topic Maps, und selbstorganisierende Karten sind beliebte grafische Darstellungen.

## 3.2 Text verstehen und erzeugen

Die hier vorgestellten Verfahren bilden die nächsten Glieder der Prozesskette. Sie analysieren und interpretieren Text und überführen ihn dabei in eine interne Darstellung bzw. erzeugen aus einer internen Darstellung Text. Für solche internen Darstellungen gibt es kein Standardformat. Es können strukturierte Daten für einen Geschäftsprozess, Abfragen an eine Informationsquelle oder Anweisungen an ein angedocktes System oder Geräte sein.

Bei Anwendungen mit tiefen Sprachmodellen wie BERT, GPT-3 und Transformern liegt die interne Darstellung im Modell, in den Verbindungsgewichten für die Ein- und Ausgabe.

### 3.2.1 Informationsextraktion aus Texten

Seit es Computer gibt, gibt es auch Texteditoren. Damit entstehen viele maschinenlesbare Dokumente:

Bei der Informationsextraktion aus Texten liefern derzeit vortrainierte BERT-Modelle die besten Ergebnisse, wobei ein Feintuning für die jeweilige Aufgabe stattfindet. Ein Beispiel ist die Extraktion von Namen (Personen, Orte, Produkte, Firmen etc.) aus beliebigen Texten, auf Englisch **Named Entity Recognition (NER)**. Darüber hinaus kann man Relationen entnehmen, zum Beispiel zur Biografie von bestimmten Personen (Geburtsjahr, Ausbildung etc.) oder zu Wechselwirkungen zwischen unterschiedlichen Medikamenten bei der Behandlung von Patientinnen und Patienten. Die gleichen Aufgaben kann man auch mit Transformer-Modellen angehen, die meist zum Eingabetext noch eine Instruktion erhalten, welche Aufgabe sie denn erfüllen sollen.

2012 hat Google **Wissensgraphen** populär gemacht, indem es Suchergebnisse um passende Fakten in einer Informationsbox ergänzte. Wissensgraphen bestehen aus Objekten, Eigenschaften und Relationen und repräsentieren Fakten und Aussagen, die man idealerweise logisch kombinieren kann. Anschaulich werden **Relationen** als Pfeile zwischen Objekten

dargestellt, und intern als Tripel (Objekt, Relation, Objekt). Man spricht deshalb auch von **Linked Data**.

Das Vokabular für die Eigenschaften, Relationen und Objekttypen sowie die möglichen Werte von Eigenschaften sind durch sogenannte **Ontologien** festgelegt. Ihr Potenzial ist riesig, wenn man bedenkt, dass Wissensgraphen, die dieselben Ontologien verwenden, problemlos miteinander kombiniert werden können. Für Ontologien gibt es sehr ausgereifte Standards des World Wide Web Consortium (W3C). Es verfolgt seit 2001 die Idee des **Semantic Web**, in dem zu den Webseiten auch semantische Informationen maschinenlesbar abgelegt werden.

Im Mittelpunkt des semantischen Web steht **DBpedia**, der Wissensgraph zu Wikipedia. Er ist mit vielen weiteren öffentlichen semantischen Datenbeständen verknüpft, die man unter dem Stichwort **Linked Open Data** findet. Sie alle können mit unternehmenseigenen Wissensgraphen kombiniert werden. Zu den Standards des W3C zählt RDF, ein XML-Format für die Tripel, in die man einen Wissensgraphen zerlegen und in einer **Wissensbasis** speichern kann. Die Abfragesprache heißt **SPARQL (SPARQL Protocol And RFD Query Language)**.

Zur Interpretation eines Satzes zählt die Erkennung von Objektbezeichnungen, Ereignissen, temporalen und räumlichen Ausdrücken. Objektarten, Relationen und Eigenschaften müssen dem

Vokabular der Ontologie zugeordnet werden und die Objekte müssen eindeutig identifiziert werden. Als **Wikifikation**, oder allgemeiner **Entity Linking**, bezeichnet man die Identifikation eines Objekts durch eine Seite in der Wikipedia bzw. DBpedia. Hierfür wird zunächst durch eine Suchmaschine eine Liste von Kandidatenseiten erzeugt. Anschließend wird für die Kandidaten anhand der thematischen Übereinstimmung zwischen dem Text, der die Objektbezeichnung enthält, und der Beschreibung der einzelnen Objekte in der Wikipedia, der beste Treffer ausgewählt. Schließlich kann man die gewonnenen Informationen als Tripel in den Wissensgraphen einfügen oder für eine SPARQL-Abfrage verwenden.

Oberhalb der Satzebene geht es darum, ganze Dokumente einzuschätzen. Man kann sie nach dem Inhalt kategorisieren, etwa um E-Mails richtig weiterzuleiten oder abzulegen. Oder man kann feststellen, ob ein Beitrag oder Kommentar toxic, also ungemessen formuliert ist, um seine Publikation zu unterdrücken. Oder man kann feststellen, ob ein Text positiv oder negativ ist (**Sentiment Analysis**), welche Emotionen ausgedrückt werden, ja sogar auf die Persönlichkeit der Verfasserin oder des Verfassers schließen, um passend zu reagieren. Auch für die Marktforschung lassen sich diese Informationen nutzen. Wie viele Benchmarktests belegen, liefern für diese Aufgaben trainierte tiefe Sprachmodelle die besten Resultate. Der Aufwand

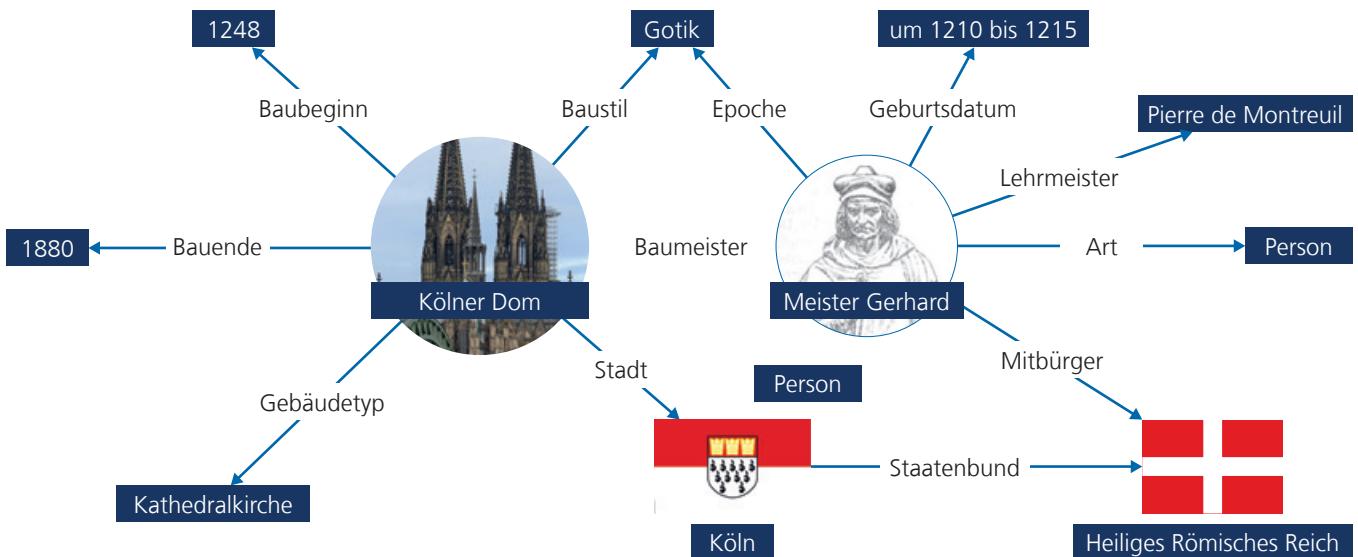


Abb. 8: Wissengraph Kölner Dom.  
Quelle: Eigene Darstellung

für die Vorverarbeitung ist gering, man muss nur Formatierungen entfernen und die Worte bzw. Wortfragmente erkennen. Eine morphologische, syntaktische oder lexikalische Analyse ist nicht nötig.

Die Informationen aus einem Text müssen nicht immer in einen Wissensgraph gelangen. Oft

werden Daten nur zur Weitergabe an andere IT-Systeme extrahiert, zum Beispiel eine Tabelle mit den Posten aus einer Rechnung. Informationsextraktion ist auch eine Teilaufgabe in Frage-Antwort-Systemen und Sprachassistenten. Erstere müssen die Frage verstehen und letztere müssen verstehen, was sie tun sollen.

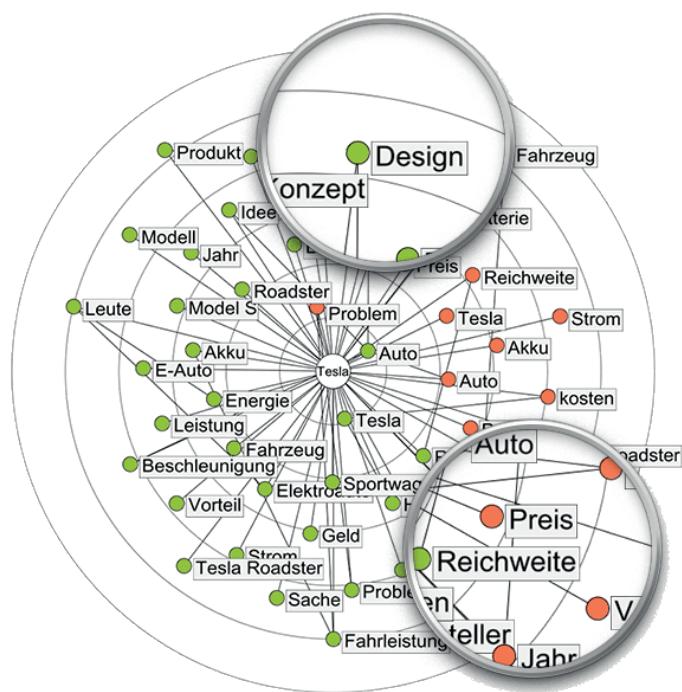
## InMap von Insius

[insi.us/de/produkte/inmap](http://insi.us/de/produkte/inmap)

InMap ist eine Textanalyse-Software des Start-ups Insius. Damit lassen sich aus großen Textmengen, z. B. Interviews, Social-Media-Postings oder CRM-Daten, die Einstellungen von Verbraucherinnen und Verbrauchern gegenüber Marken, Produkten, Services oder anderen relevanten Meinungsgegenständen aufdecken.

Die Algorithmen ermitteln mittels Informationsextraktionsverfahren Konzepte und Aussagen aus Beiträgen und ordnen ihnen eine Tonalität

(positiv, negativ, neutral) zu. Sie unterscheiden zwischen Themenbereichen, da Begriffe unterschiedlich positive oder negative Bedeutung haben können. Und sie extrahieren auch Begründungen. Mit InMap lässt sich die Stärke der Konzepte direkt aus der Nähe zum Zentrum ablesen. InMap kann außerdem für jedes Konzept anzeigen, welche Produkte im Kontext der Diskussionen eine Rolle spielen und welche Webseiten für bestimmte Konzeptwahrnehmungen der Nutzer verantwortlich sind.



Quelle: [insi.us](http://insi.us)

## Expertenmeinung André Lang, Insius



**INSIUS**

»Mit Sprachtechnologien können wir große Textmengen strukturiert, umfassend und nachvollziehbar auswerten. Verlässlichkeit entsteht dabei aus Nachvollziehbarkeit.«

André Lang ist Dipl.-Wirtschaftsinformatiker und geschäftsführender Gesellschafter der Insius UG. Er erforscht und entwickelt seit über zehn Jahren Textanalytics- und NLP-Lösungen im Marktforschungsbereich. Seine Schwerpunkte liegen auf der Textklassifizierung, der Informationsextraktion und der Computerlinguistik.

### Welche wirtschaftliche Bedeutung hat der Einsatz von Sprachtechnologien für Ihr Unternehmen?

*Sprachtechnologien sind der Kern unserer Analysetätigkeit. Nur durch sie haben wir die Möglichkeit, große Textmengen strukturiert, umfassend und nachvollziehbar auszuwerten.*

### Was sind dabei aus Ihrer Sicht die größten Herausforderungen und welche Use Cases besitzen besondere Potenziale (z.B. in den Kriterien Niedrigschwelligkeit, Akzeptanz und Skalierbarkeit)?

*Die größte Herausforderung sehen wir in Verlässlichkeit, die durch Nachvollziehbarkeit entsteht. Nur einem System, das reproduzierbar erwartbares Verhalten zeigt, sind wir bereit, intuitiv Vertrauen zu schenken.*

### Welche ersten Schritte sollten Unternehmen gehen, um konkret Sprachtechnologien zu etablieren?

*Unternehmen sollten prüfen, welche sprachbasierten Informationen ihnen entweder bereits vorliegen oder leicht zugänglich sind. Freitextfelder in Fragebögen, Feedback-E-Mails, aber auch Kommentare auf Webseiten und in Social Media lassen sich mit Sprachtechnologien inzwischen leicht, systematisch und nachvollziehbar auswerten. So kann auch in kleineren Pilotprojekten die nötige Erfahrung gesammelt und ins Unternehmen gebracht werden.*

### Was bedeuten Sprachmodelle wie GPT-3 für Europa? Und was bedeuten sie für die Weiterentwicklung und Forschung in Ihrem Unternehmen?

*Sprachgenerierende Modelle wie GPT-3 haben vor allem im Games-Bereich Potenzial. Bei uns im Unternehmen kommt es nicht zum Einsatz, da wir uns eher auf die Textanalyse fokussieren. Da Sie aber explizit nach Europa fragen: Wir müssen im Blick haben, dass diese Technologien immer auch missbraucht werden können. Ähnlich wie Spam-Mailer ermöglichen sie, mit wenig Aufwand gefärbte Texte zu generieren, mit denen Kanäle geflutet werden können, um eine politische Agenda zu erreichen. Das ist nicht weniger als eine Gefahr für die Demokratie in Europa, weil die bislang bestehenden Skalierungsschranken der schon heute aktiven Troll-Armeen damit wegfallen.*

Wo sehen Sie die Stärken und Schwächen von Deep Learning in der praktischen Anwendung von Sprachtechnologien und NLP in der Praxis – und gibt es Gebiete, in denen Deep Learning nicht eingesetzt werden sollte?

*Da Deep Learning keine Kausalitäten kennt, sondern nur Korrelationen reproduziert, ist das Wissen über den Lern-Input entscheidend. So wie ich heute bei Personen oder Quellen die fachliche Qualifikation und ideologische Richtung überprüfen und reflektieren kann, muss ich das auch bei Deep-Learning-Modellen anhand der Trainingsdaten können. Kann ich das nicht, und liefert das Modell auf unbekannten Trainingsdaten transparente Entscheidungen, sind die Schäden umso größer, je größer die Folgen der einzelnen Entscheidungen sind und je länger systematische Fehler aufgrund von Intransparenz und mangelnder menschlicher Kontrolle unentdeckt bleiben.*

Zukunft der Sprachtechnologien – was können wir in zehn Jahren erreichen?

*Sprachtechnologien ermöglichen uns, den gesammelten und in vielen verschiedenen Formen geäußerten Wissens- und Erfahrungsschatz aller Menschen zugänglich zu machen. Anders als früher, wo stets gezielte Experimente oder Befragungen notwendig waren, bekommen wir in Zukunft direkten Zugang zum gesammelten Wissen aller. Auf dem Weg dahin liegen noch viele einzelne Schritte zur praktischen Umsetzung, aber an dessen Ende werden bessere Produkte und Leistungen stehen.*

Textverstehen ist für Sprachassistenten essenziell. Was ein Sprachassistent tun kann und welche Kommandos ein sprachgesteuertes Gerät ausführen kann, bestimmen seine **Aktionen** – so heißt es beim Google Assistant –, bei Alexa sind es die **Skills**. Damit eine Aktion ausgeführt werden kann, muss intern ein **Rahmen (frame)** mit einer **Absicht (intent)** und **Parametern (slots)** gefüllt werden. Zum Beispiel gehören zum Intent *Finde-Restaurant* die Slots *Anzahl Personen*, *Gegend* und *Art der Küche*, sodass der Frame

*(Finde-Restaurant, Köln, 3, Chinesisch)*

einen ausführbaren Auftrag darstellt. Es geht aber noch weiter. Ein Sprachassistent muss auch verstehen, ob sein Dialogpartner ihm überhaupt gerade einen Auftrag erteilen möchte oder ob er ihm etwas mitteilen oder bestätigen möchte oder eine Rückfrage stellt. Das sind verschiedene **Sprechakte**, auf die ein Sprachassistent unterschiedlich reagieren muss. Für das Erkennen von Sprachhandlungen, Absichten und Parametern verwendet man Methoden der Informationsextraktion. Man kann Stichwörter und Textmuster festlegen, um diese Informationen aus dem Text zu extrahieren. Moderner ist die Nutzung von tiefen Sprachmodellen, die mit großen Textmengen trainiert wurden.

### 3.2.2 Textsynthese

Die Textsynthese ist das Gegenstück zum Textverstehen. Hier wird also automatisiert Text generiert, der anschließend in Sprachsignale verwandelt werden kann. Auch dieser Arbeitsschritt ist abhängig von dem Zweck der Anwendung. Abfragen eines Wissensgraphen liefern Tripel, aus denen verständliche Texte erzeugt werden müssen, und angedockte Systeme können kodierte Rückmeldungen liefern, die man ebenfalls vertexten möchte.

Tiefe Sprachmodelle heben die Textsynthese auf eine sehr viel anspruchsvollere Ebene. Ein Beispiel ist das Zusammenfassen von umfangreicheren Dokumenten, um die wichtigsten Aussagen herauszuziehen. In der letzten Zeit wurden Transformer

entwickelt, die Dokumente mit über 4000 Wortfragmenten verarbeiten können. Eines dieser Modelle, genannt *BigBird*, geht Zusammenfassungen als Übersetzungsaufgabe an, bei der ein langer Text in einer Reihe frei formulierter Sätze übersetzt wird. Es wird auf einem großen Datenbestand vorgenutzt. Ein Feintuning kann stattfinden auf Informatiktexten, medizinischen Texten oder Patenten sowie deren Zusammenfassungen. In allen drei Anwendungsbereichen konnte BigBird die Qualität der Zusammenfassungen entscheidend verbessern. Auch hinter manchen Chatbots verbergen sich tiefe Sprachmodelle, die von Text-zu-Text ohne Zwischenschritte Small-Talk produzieren. Es ist abzusehen, dass tiefe Sprachmodelle für weitere Aufgaben eingesetzt werden, zum Beispiel für Textzusammenfassungen, Textverbesserungen oder als Textroboter für Standardberichte. Kreativschaffende könnten Sprachmodelle zur Inspiration als Textvorschlagssysteme einsetzen.

Oft möchte man eine stärkere Kontrolle über die Textsynthese ausüben, also beispielsweise das Thema, die Emotion, und die Sprecherpersönlichkeit vorgeben. Das Sprachmodell *Megatron-Cntrl* greift während der Texterzeugung auf eine externe Wissensdatenbank mit Relationstripeln zurück und kann damit die erzeugten Texte inhaltlich unterfüttern. Das gleiche Modell kann Texte in bestimmten Textstilen (wie Nachrichten, Märchen, Roman, Werbung) erzeugen oder gemäß unterschiedlicher Sprecherpersönlichkeiten generieren. Schließlich kann man ein tiefes Sprachmodell darauf trainieren, eine lange Geschichte auf zwei Ebenen zu erzeugen. Zunächst schreibt das Modell ein Storyboard, das in einzelnen Sätzen die grobe Abfolge der Ereignisse beschreibt. Anschließend formuliert das Modell die Einzelheiten aus. Hier sind dann auch Eingriffe möglich, sodass man das Storyboard nach seinen Wünschen ändern kann.

### 3.2.3 Maschinelle Übersetzung

Für maschinelle Übersetzungen werden tiefe Sprachmodelle verwendet, die speziell mit Texten in zwei Sprachen trainiert wurden. Hierfür gibt es drei wichtige Anwendungsfelder: Zunächst ist da die reguläre Textübersetzung, die für den

Alltagsgebrauch oder auch für professionelle Übersetzerinnen und Übersetzer maschinell produzierte Entwürfe generiert. Dann gibt es die Möglichkeit, Text auf Fotos zu erkennen und in andere Sprachen zu übertragen. Das eignet sich besonders für Auslandsaufenthalte: Mit dem Smartphone lassen sich Hinweisschilder oder Speisekarten scannen und dann sofort in der eigenen Sprache lesen. Der dritte Anwendungsfall bezieht sich auf die Übersetzung von Gesprächen, also Audiosignalen, in Echtzeit. Besonders anspruchsvoll ist hier die Simultanübersetzung, die mit der Übersetzung bereits beginnt, noch bevor ein Satz zu Ende gesprochen ist.

Für Übersetzungen werden heutzutage Transformer-Netze Ende-zu-Ende trainiert. Das heißt, zur Übersetzung ist keine linguistische oder gar semantische Analyse nötig. Benötigt werden nur große Mengen von bilingualen Texten, also Texte mit Übersetzungen. Für Übersetzungen aus selteneren Sprachen kann das eine Herausforderung sein. Allerdings beherrschen die Übersetzungsdiene der großen Tech-Konzerne inzwischen Hunderte von Sprachen. Die Übersetzungsqualität kann man beim Training der Netze verbessern, indem man Hin- und Rückübersetzer kombiniert und die Unterschiede zwischen den Übersetzungen zu minimieren versucht.

Facebook hat sein M2M-100 Übersetzungsmodell frei zur Verfügung gestellt. Mithilfe eines einzelnen multilingualen Modells und einem gemeinsamen Wortfragmentvokabular für alle Sprachen kann es Übersetzungen zwischen mehr als 100 verschiedenen Sprachen erzeugen. Das Modell ist ein Transformer mit insgesamt bis zu 15 Milliarden Parametern. Die Übersetzungsqualität ist überzeugend und übertrifft die von Übersetzern, die Englisch als Zwischensprache verwenden. Denn es wurden Trainingsdaten von sehr vielen Sprachpaaren verwendet, und durch die gemeinsamen multilingualen Einbettungen werden Informationen von einer Sprache in ähnliche Sprachen übertragen. Es sind sowohl die Skripte zum Sammeln von Trainingsdaten, der Code der Modelle, als auch die Parameter fertig trainierter Modelle verfügbar.

## ella von ella

ella.ag

Die KI-basierte Software ella des gleichnamigen Start-ups kann automatisiert hochwertigen Content und kreative Geschichten erzeugen, egal ob Produktbeschreibung, Infotext oder Kurzgeschichte. Jeder Text wird auf Knopfdruck individuell und innerhalb von nur wenigen Minuten generiert. Die Einsatzmöglichkeiten sind groß, denn die automatische Texterzeugung kann in nahezu allen Branchen genutzt werden. Besonders relevant ist die Medienbranche: ella kann Mitarbeitende in Verlagen und Redaktionen unterstützen und bietet neben Wort-für-Wort-Schreibhilfen auch Recherchehilfe an. Auf Wunsch generiert das System auch Vorschläge für ganze Absätze, erstellt Artikelaufsätze oder liefert Zusammenfassungen von Texten. Da ella bereits 16 Sprachen beherrscht, kann das Unternehmen auch international mitmischen. Der Firmensitz wurde in die Schweiz verlegt, ein Großteil des Teams sitzt noch in Köln.

Bei der akustischen **Simultanübersetzung** können die Worte erst ausgegeben werden, wenn das Modell genügend Kontext gesammelt hat, da Korrekturen nicht mehr möglich sind. Dabei kann man die Spracherkennung und die Übersetzung zur Beschleunigung verzahnen. Wird Sprache auf dem Umweg über Text übersetzt, so gehen Informationen über die Stimme des Sprechers, dessen Geschlecht und die Sprachmelodie verloren. Daher versuchen einige Systeme, wie der Skype-Übersetzer, diese Informationen in einem Audio-Audio-Übersetzungsmodell zu erhalten.

## Linguee von DeepL

linguee.de

Das Kölner Start-up ist durch seinen kostenfreien Online-Übersetzer und sein Wörterbuch Linguee bekannt. Das Übersetzungssystem operiert auf der Basis von Faltungsnetzen. Viel Aufmerksamkeit hat DeepL bekommen, nachdem seine Übersetzungen in Blindtests besser bewertet wurden als die von Google und anderen Konzernen. DeepL profitiert von den Daten, die es mit Linguee seit über zehn Jahren gesammelt hat. Geschäftskunden bezahlen dafür, dass ihre Spracheingaben nicht gespeichert werden.

### 3.3 Interne Verarbeitung

Im mittleren Teil der Sprachprozesskette finden sich Technologien, die aus Text extrahierten Informationen weiterverarbeiten und dazu sprachliche Kompetenzen benötigen. Ein wichtiger Anwendungsfall sind semantische Suchmaschinen und Frage-Antwort-Systeme. Weitere Beispiele sind die Sprachsteuerung von extern angedockten Systemen und Geräten und die Dialogsteuerung in Chatbots.

#### 3.3.1 Semantische Suche und Frage-Antwort-Systeme

Watson von IBM war das erste Frage-Antwort-System, das weltweit Aufsehen erregte. 2011 schlug es in der Frageshow Jeopardy die beiden bisher besten Teilnehmer. Möglich wurde dies durch eine 100 GB große Datenbank von Fakten und Zusammenhängen, die Watson durchsuchen konnte.

Aktuelle **Frage-Antwort-Systeme** stützen sich auf eine externe Informationsquelle. Das kann eine Dokumentensuchmaschine oder eine Wissensbasis mit einem Wissensgraph sein. Für eine Dokumentensuchmaschine muss man die Suchwörter aus dem Eingabetext extrahieren. Die Suchmaschine liefert dann Dokumente, in denen die Suchwörter auftauchen. Die besten Dokumente werden passagenweise untersucht, um Textstellen

zu identifizieren, die am besten zur Frage passen (**reading comprehension**). Die beste davon wird als Antwort ausgegeben. Für eine Wissensbasis muss man aus dem Fragetext eine SPARQL-Abfrage erzeugen. Das ist für Menschen mühsam, denn man muss nicht nur die SPARQL-Syntax kennen, sondern auch das Vokabular und die Relationen der verwendeten Ontologien. Außerdem muss man das oder die Tripel, die man zurückbekommt, verstehen. Diese Mühe nimmt einem das Frage-Antwort-System als Front-End ab. Es kann die SPARQL-Abfrage mit Methoden der Informationsextraktion aus der Frage extrahieren und die Ergebnistripel mit Methoden der Textgenerierung vertexten.

Bei Dokumentensuchmaschinen findet gerade eine Revolution statt. Bisher waren sie **wortorientiert**: Sie finden Dokumente, wenn in ihnen die Suchwörter auftauchen. Sucht man etwa nach »Erfinder des Autos«, so werden Dokumente mit »Entwickler des Automobils« nicht gefunden. BERT ermöglicht es nun, Textpassagen durch Einbettungen sinngemäß zu repräsentieren. So kann man sämtliche Textpassagen der Wikipedia durch Einbettungen darstellen. Anschließend kann man eine Suchanfrage ebenfalls durch ein BERT-Modell in einen Einbettungsvektor transformieren. Das Verfahren liefert nun die Textpassagen, deren Einbettungen eine kurze Vektor-distanz zu der Frage-Einbettung besitzen.

Die semantische Suche kann auch in Anwendungen integriert werden, sodass sie den Nutzerinnen und Nutzern sozusagen über die Schulter schauen und vorausschauend hilfreiche Suchergebnisse oder Texte vorschlagen. Man muss also keine explizite Frage formulieren, ja noch nicht einmal wissen, was man sucht oder dass man etwas suchen könnte. Eingaben für die semantische Suche sind die Texte, die gerade am Rechner bearbeitet werden.

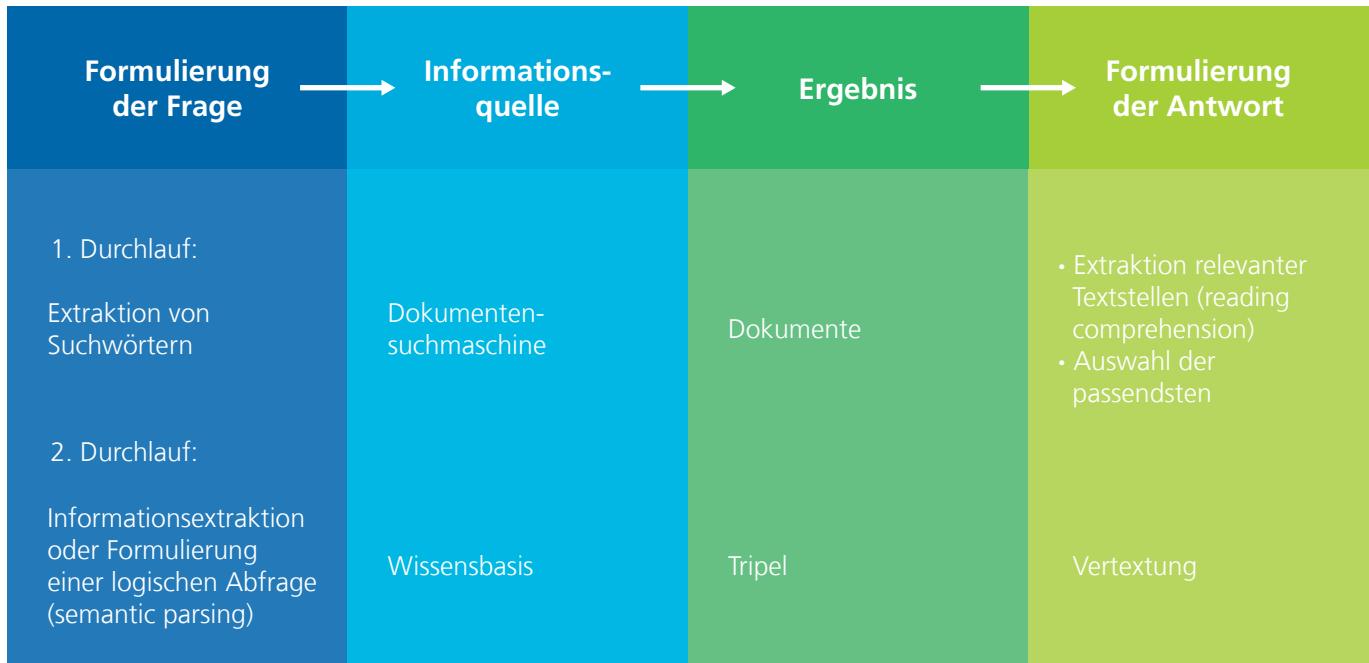


Abb. 9: Frage-Antwort-Systeme. Quelle: Eigene Darstellung

## SEMPRIA-Search von SEMPRIA

[sempria.de](http://sempria.de)

Archivbesitzer (Firmen, Websites, Verlage, TV, Radio ...) haben oft Probleme damit, Archive voll und effizient zu erschließen. Die lokale Suchmaschine SEMPRIA-Search des Start-ups SEMPRIA hilft hier durch höhere Vollständigkeit und Genauigkeit der Suchergebnisse und die Möglichkeit gezielter Fragen. Genutzt wird dazu tiefes, wissensbasiertes Sprachverständnis mit Wissensgraphen und moderner Sprachtechnologie: Jedes Textdokument und jede Suchanfrage wird umfassend inhaltlich analysiert. Die bedeutungsorientierte Suchmaschine verknüpft gleichbedeutende Formulierungen zwischen

Dokumenten und Suchanfragen intelligent miteinander. Neben Texten können auch Scans, Audios und Videos indexiert werden. Suchanfragen können geschrieben oder gesprochen werden.

Durch die Verknüpfungen können Suchanfragen und Dokumente auf der inhaltlichen Ebene abgeglichen werden. SEMPRIA-Search kann abweichendem Wortlaut mit Wissen zu Synonymen, unterschiedlichen Formulierungen, Paraphrasierungen, variierender Terminologie, Fachsprachen und Begriffshierarchien begegnen.

## Überprüfung von Jahresabschlussberichten mit ALI

Das Automated List Inspection (ALI) Tool kombiniert maschinelles Lernen mit Expertenwissen, um den Prozess der Überprüfung von Jahresabschlussberichten zu vereinfachen. ALI ist ein Vorschlagssystem, das Textpassagen aus dem Anhang eines Jahresabschlusses den relevanten gesetzlichen Vorschriften zuordnet, und wurde vom Fraunhofer IAIS in den letzten Jahren in Zusammenarbeit mit der Wirtschaftsprüfungsgesellschaft PwC entwickelt. Seit 2019 wird ALI von PwC im Alltagsgeschäft erfolgreich eingesetzt.

Bisher mussten Wirtschaftsprüfer den gesamten Bericht händisch durchgehen, um für jeden Abschnitt die jeweils relevanten Vorschriften zu identifizieren. Diese Prozesse sind meist hoch repetitiv und sehr zeitaufwändig, da es je nach Anwendungsbereich mehrere Hunderte bis Tausende

potenzieller Anforderungen gibt. Durch ALI lässt sich der Prozess deutlich beschleunigen. Das Vorschlagssystem funktioniert dabei in beide Richtungen: Man kann sich sowohl die relevanten Passagen pro Anforderung als auch die relevanten Anforderungen pro Textpassage anzeigen lassen.

Es gibt weiterhin einen Konsistenzcheck, der den Inhalt von Tabellen auf Korrektheit und Übereinstimmung überprüft. Die integrierte Anonymisierungsfunktion identifiziert automatisch die sensiblen Daten während des Hochladens und macht diese unkenntlich. Weitere geplante Features sind eine automatisierte, qualitative Analyse der Berichte in Hinblick auf Vollständigkeit und Widersprüche, sowie ein Abgleich zwischen Tabelleninhalten und Text.

Semantische Suchverfahren lassen sich auch für Frage-Antwort-Systeme mit einem Wissensgraphen verwenden. Ein BERT-Modell kodiert die Anfrage und erhält eine Reihe von Textpassagen aus einer mit BERT kodierten Wissensdatenbank. Anschließend analysiert ein Transformer die Anfrage und die gefundenen Textpassagen und erzeugt eine Antwort in natürlicher Sprache. Für diese Aufgabe erhielt der Transformer eine Feintuning. Ein Modell dieser Art übertraf bei einem Vergleich sogar GPT-3, obwohl es nur einen kleinen Bruchteil der Parameter hatte. Es kann Fragen auch besser beantworten, als es derzeit unter Verwendung von aufwendig und manuell erstellten Wissensbasen automatisch möglich ist. Wichtig ist, dass das Modell ohne Neutrainings an geändertes Hintergrundwissen angepasst werden kann. Derartige Systeme sind noch im Bereich der Forschung, werden aber von Microsoft, Facebook, Google etc. vorangetrieben. Derzeit konzentriert man sich auf Fragen mit kurzen Antworten von etwa fünf Wörtern.

### 3.3.2 Dialogmanagement

Sprachassistenten und Chatbots sind **Dialogsysteme**. Sie beherrschen Gespräche, die nicht nur aus der Erwiderung auf eine einzelne Frage oder Aufforderung bestehen. Sie sollten sich an Dinge erinnern, die schon erwähnt wurden und erkennen, wann sie etwas nicht verstanden haben. Wenn ihnen Informationen fehlen, sollten sie nachfragen. Manche können auch Small Talk.

Die zentrale Komponente eines Dialogsystems ist der **Dialogmanager**. Er verarbeitet die interne Darstellung, die durch Informationsextraktion aus dem Eingabetext gewonnen wurde, also Absicht (intent), Parameter (slots) und Sprechakte (siehe Kapitel 3.2.1.). Dazu verwaltet er den aktuellen **Zustand**: Auf welche Absicht bezog sich das Gespräch zuletzt, welche Parameter sind schon bekannt, welche müssen noch erfragt werden und was trägt

die aktuelle Eingabe dazu bei? Wurde das Thema gewechselt oder eine Nachfrage gestellt? Wurde etwas missverstanden und sollte man besser nachfragen? Oder sind alle Informationen eingesammelt und kann die Aktion jetzt ausgelöst werden?

Wenn man eine Informationsfrage gestellt hat – »Wie ist das Wetter?«, »Was kostet aktuell der Sprit?«, »Wie lange würde ich jetzt zur Arbeit brauchen?« – verhält sich der Dialogmanager wie ein Frage-Antwort-System. Er kann für seine Antwort auch ein Informationssystem des Unternehmens konsultieren, das ihn einsetzt. Soll er eine Transaktion durchführen, zum Beispiel eine Buchung, einen Kauf oder eine Terminvereinbarung, dann muss er diese ebenfalls in einem externen IT-System auslösen.

Das Ergebnis der ausgelösten Aktion muss der Dialogmanager interpretieren und seinem Gegenüber rückmelden. War die Aktion überhaupt erfolgreich? Woran ist sie gescheitert? Was ist das Ergebnis? Was soll geantwortet werden? Schließlich muss er sich noch das Wichtigste merken und seinen Zustand aktualisieren.

Dialogmanager, die auf eine hinterlegte Wissensbasis oder Datenbank zurückgreifen, werden oft programmiert und Antworten über Schablonen generiert. Aber für die Strategie, die die Antwort auswählt, kann auch ein Netz trainiert werden.

## Expertenmeinung Oliver Trabert, Learnship



**/learnship/**

»Signifikante Potenziale in der Zukunft sehen wir bei der KI-basierten Interaktion mit den Studierenden für einfache Lernaufgaben mit einem Dialog-Bot. Für EdTech wird KI in den kommenden zehn Jahren zum Game-Changer werden.«

Oliver Trabert ist CTO bei Learnship aus Köln. Das Unternehmen bietet digitale Sprach- und interkulturelle Trainings für Unternehmenskunden an. Oliver verantwortet die Produktentwicklung und hat sich in den letzten drei Jahren intensiv mit den Themen NLP, Convolutional Neural Networks und ML-basierten Dialogsystemen beschäftigt. Seine Vision ist die grundlegende Veränderung von Sprachenlernen durch den Einsatz von Künstlicher Intelligenz.

**Was waren für Sie die wichtigsten Durchbrüche in Sprachtechnologien und NLP in den letzten fünf Jahren insbesondere in Bezug auf die Auswirkungen für die wirtschaftliche Wertschöpfung?**

Ein wesentlicher Durchbruch war die Verfügbarkeit von NLP als Product-as-a-Service, wie z. B. von Google. Dort kommen bei uns u. a. Speech-2-Text, Text-2-Speech sowie Syntaxanalysen zum Einsatz. In unserem Anwendungskontext war auch die Grundlagenforschung zu SRI an der Stanford University basierend auf Phonemanalyse sehr wichtig. Indirekt spielten sicherlich auch die Entwicklung von kontextualisierten Sprachmodellen wie ELMo und BERT eine große Rolle für die Weiterentwicklung von Sprachtechnologien.

**Wo setzen Sie Sprachtechnologien / NLP bei sich im Unternehmen ein?**

Wir nutzen Sprachtechnologien u. a. bei der Einstufung unserer Studierenden nach dem gemeinsamen europäischen Referenzrahmen für Sprachen (CEFR) durch die computerbasierte Analyse von Sonogrammen. Hier greifen wir auf ein klassisches Klassifizierungsmodell zurück mit vielen Hunderttausenden klassifizierten Beispieldaten (bestehend aus Sonogramm und Einstufung). Dabei erreichen wir ein über 95 Prozent gutes Ergebnis. Eine weitere Anwendung ist die automatische Erzeugung von Lernaufgaben aus Google News Feeds. Eine Aufgabe wäre z. B., aus einem Newsfeed bestimmte Worte wegzulassen und als Multiple-Choice-Antwort Synonyme, aber auch Antonyme anzubieten.

**Welche wirtschaftliche Bedeutung hat der Einsatz von Sprachtechnologien für Ihr Unternehmen?**

Die automatische Spracheinstufung nach CEFR erspart die manuelle Einstufung durch eine Lehrerin oder einen Lehrer. Pro Einstufungstest spart dies fünf Euro. Es finden pro Jahr ca. 20 000 Einstufungen statt (=100 000 Euro). Die automatische Erzeugung der Lernübungen spart zwei Vollzeitkräfte pro Jahr (auch ca. 100 000 EUR). Weitere signifikante Potenziale in der Zukunft sehen wir bei der KI-basierten Interaktion mit den Studierenden für einfache Lernaufgaben mit einem Dialog-Bot.

**Was sind dabei aus Ihrer Sicht die größten Herausforderungen und welche Use Cases besitzen besondere Potenziale?**

Großes Potenzial beim Sprachlernen hat der Einsatz von Dialogsystemen. Mit diesen könnte die Aktivierung der gelernten Vokabeln, Grammatik und funktionalen Sprache erfolgen. In diesem Fall kann dies die Lehrkräfte ersetzen. Weiterhin kann der Kundensupport über Bot-Systeme signifikant reduziert werden. Hierbei ist allerdings zu

*beachten, dass sich Kundinnen und Kunden nicht immer erwartungskonform verhalten und die Dialoge dies berücksichtigen müssen.*

**Welche ersten Schritte sollte ein Unternehmen gehen, um konkret Sprachtechnologien zu etablieren?**

*Grundsätzlich empfehlen wir, erstmal mit einfachen Anwendungsfällen zu starten und bei den Lösungen in verschiedenen Ausbaustufen zu denken. Das könnte bei einem Chatbot zunächst mit dem Beantworten von standardisierten Anfragen starten und in weiteren Ausbaustufen mit komplexeren Anfragen ergänzt werden. Weiterhin ist die Nutzung von Product-as-a-Service-Modulen hilfreich, bzw. die Nutzung von Modellen und Modulen, die schon andere Entwickler gebaut haben und die dann maximal durch Transferlernen auf den eigenen Anwendungsfall gelernt werden müssen. Eigene Modelle sollten erst entwickelt werden, wenn der Anwendungsfall relevant genug ist und schon genug eigene Erfahrung mit KI gesammelt worden ist.*

**Was bedeuten Sprachmodelle wie GPT-3 für Europa? Und was bedeuten sie für die Weiterentwicklung und Forschung in Ihrem Unternehmen?**

*GPT-3 ist ein interessanter Ansatz. Die Anwendung ist aktuell leider nur im Rahmen der kommerziellen Nutzung von OpenAI möglich. Interessante Alternativen bietet z. B. GPT-Neo als Open Source. Frei verfügbare europäische bzw. deutsche Sprachmodelle wären sicherlich sehr wünschenswert.*

**Wo sehen Sie die Stärken und Schwächen von Deep Learning in der praktischen Anwendung von Sprachtechnologien und NLP in der Praxis – und gibt es Gebiete, in denen Deep Learning nicht eingesetzt werden sollte?**

*Schwächen sehen wir bei der mangelnden Robustheit von Sprachmodellen, aber auch dem Bias auf Basis der beim Einlernen verwendeten Texte. Weiterhin sehen wir in der praktischen Anwendung auch eine Fehleranfälligkeit, speziell nach Updates von Komponenten. Grundsätzlich stellt Deep Learning aktuell nur gute »Filter« bereit, um z. B. Bilder zu klassifizieren oder Sprache zu erkennen. In Show-Effekten, z. B. beim Schreiben von Texten durch GPT-3, entsteht dann der Eindruck, dass alles dahinter auch schon sehr intelligent wäre, das ist aber in Wirklichkeit nicht so. Und da sind wir erst ganz, ganz am Anfang.*

**Was denken Sie, was können wir in zehn Jahren mit Sprachtechnologien erreichen und welche wirtschaftlichen Potenziale kann der Einsatz von Sprachtechnologien noch entfalten?**

*Für EdTech wird KI in den kommenden zehn Jahren zum Game-Changer werden. Lernen könnte dann in Zukunft mit weniger Lehrkräften auskommen. Das kann insbesondere für jene Regionen der Welt eine Chance sein, in denen es ohnehin an Lehrpersonal mangelt und auch in absehbarer Zukunft weiterhin mangeln wird.*

## Expertenmeinung

### Dr. Sven Hartrumpf, Sempria



»Beim wirtschaftlichen Potenzial von Sprachtechnologie stehen wir erst ganz am Anfang. Viele Anwendungen setzen eine digitale Orientierung von Firmen und Institutionen voraus, die mitwachsen muss.«

Dr. Sven Hartrumpf ist Informatiker, Computerlinguist und Geschäftsführer bei der SEMPRIA-GmbH in Düsseldorf.

**Was waren für Sie die wichtigsten Durchbrüche in Sprachtechnologien und NLP in den letzten fünf Jahren insbesondere in Bezug auf die Auswirkungen für die wirtschaftliche Wertschöpfung?**

*Die Verfügbarkeit freier, marktfähiger Lösungen zur Spracherkennung (also: Text-to-Speech) gerade auch für kleinere Sprachen. Dies ist wichtig für die Sprachvielfalt in Europa, die ich als Bereicherung und Chance und nicht als Hindernis sehe. Weiterhin hat das Thema Linked Open Data (LOD; z. B. WikiData) so viel Fahrt aufgenommen, dass es für wissensbasierte Sprachsysteme große neue Möglichkeiten eröffnet.*

**Welche ersten Schritte sollten Unternehmen gehen, um konkret Sprachtechnologien zu etablieren?**

*Firmen sollten in offenen Kommunikationsformaten ermitteln, wo bei einem »erstmal dummen« Computer Datenverluste entstehen können, z. B. verlorene Informationen, unbalancierte Entscheidungsgrundlagen durch Recherche-Lücken, lange Suchprozesse etc. So können die größten Schmerzpunkte identifiziert werden. Für diese kann man dann Lösungen einkaufen und nach einem ersten Erfahrungszyklus*

*den nächsten Schmerzpunkt zum Einsatz von NLP identifizieren. Das Vorgehen ist also vom Schmerzpunkt zur Lösung. Das umgekehrte Vorgehen – von der Lösung zum Schmerzpunkt – ist dagegen mit vielen Risiken verbunden.*

**Wo sehen Sie die Stärken und Schwächen von Deep Learning in der praktischen Anwendung von Sprachtechnologien und NLP in der Praxis – und gibt es Gebiete, in denen Deep Learning nicht eingesetzt werden sollte?**

*Die größte Stärke ist, dass erstmal viel manuelle Arbeit erspart werden kann, indem man die Maschine auf massiven Trainingsdaten trainiert und oft schnell ein gewisses Leistungsniveau erreicht. Als Schwächen möchte ich nennen: Mängel in der Erklärfähigkeit, Nachvollziehbarkeit und Erweiterbarkeit. Für Sprachen mit weniger verfügbaren Trainingsdaten ist es sehr fraglich, wie man zu annähernd vergleichbaren Ergebnissen wie für Englisch kommen soll. Deep Learning (oder KI im Allgemeinen) sollte nicht eingesetzt werden, wenn der Ansatz keine ausreichenden Erklärungen für den nutzenden Menschen ermöglicht und auf dem Deep-Learning-System beruhend kritische Entscheidungen getroffen werden, die Freiheitsrechte von Menschen verletzen können. Beispiel: Ein KI-System, das eine medizinische Behandlung aufgrund der NLP-Analysen von Journal-Artikeln ohne ausreichende Nachvollziehbarkeit empfiehlt, ist moralisch (und hoffentlich weiterhin juristisch) nicht akzeptabel.*

## Leseempfehlungen

### Speech and language processing

Dan Jurafsky, James Martin (2020)

»The first of its kind to thoroughly cover language technology – at all levels and with all modern technologies – this text takes an empirical approach to the subject, based on applying statistical and other machine learning algorithms to large corporations.«

Die dritte Auflage dieses Buches mit vielen Aktualisierungen in Bezug auf tiefes Lernen und Sprachmodelle ist in Arbeit und wird für Ende 2021 angepeilt. Der Stand von Dezember 2020 ist frei herunterladbar.

Verfügbar unter: <https://web.stanford.edu/~jurafsky/slp3/>

### Text to Knowledge Graph - Knowledge Extraction Pipeline with Transformers

Y Prasad (2020)

»This article discusses an automated pipeline based on neural language models that extracts knowledge from Text and populates a Semantic Knowledge Graph.«

Verfügbar unter: <https://medium.com/swlh/text-to-knowledge-graph-683002cde6e0>

### Survey on Challenges of Question Answering in the Semantic Web

Konrad Höffner, Sebastian Walter, Edgard Marx, Ricardo Usbeck, Jens Lehmann, Axel-Cyrille Ngonga Ngomo (2016)

»This survey analyzes 62 different Semantic Question Answering systems, which are systematically and manually selected using predefined inclusion and exclusion criteria. ... We identify common challenges, structure solutions, and provide recommendations for future systems.«

Verfügbar unter: [https://www.researchgate.net/publication/303826023\\_Survey\\_on\\_Challenges\\_of\\_Question\\_Answering\\_in\\_the\\_Semantic\\_Web](https://www.researchgate.net/publication/303826023_Survey_on_Challenges_of_Question_Answering_in_the_Semantic_Web)

# 4 Konversations-KI

Das aktuelle Schlagwort für das Gebiet der sprachbasierten Dialogsysteme ist Konversations-KI (**Conversational AI**). KI-Anwendungen, mit denen man sich unterhalten kann, sind besonders anspruchsvoll, da sie die gesamte Prozesskette der Sprachtechnologien abdecken. Heute können wir über Sprachassistenten mit immer mehr Geräten sprechen. Sprachschnittstellen verdrängen mühsamere Hardware: Wir brauchen uns immer weniger Knöpfe und Tastenkombination zu merken, um Geräte zu bedienen, immer seltener eine Maus, um Menübaum aufzublättern und auch immer seltener eine Tastatur, um Formulare auszufüllen. Statt Texte einzutippen, können wir sie diktieren. Dies ist besonders wichtig für Menschen mit geringer Schreib- und Lesekompetenz, mit Sehschwächen oder motorischen Einschränkungen. Mit Chatbots vergeuden wir außerdem viel weniger Zeit in Hotlines, um Routineanliegen zu lösen. Die Entwicklungen der letzten Jahre haben zudem gezeigt, dass Unterhaltungen mit Dialogsystemen immer anregender werden, erst recht, wenn sie in humanoiden Robotern, Spielzeug oder Avataren verbaut sind.

## 4.1 Chatbots

Seit seiner Gründung 1956 hat das Gebiet der Künstlichen Intelligenz Turings berühmten Test als Herausforderung verstanden. Schon 1964 entstand mit ELIZA ein Chatbot, der erstaunlich geschickt programmiert war und als täuschend echter Gesprächspartner herhalten konnte.

Waren Chatbots ursprünglich rein textbasiert, so beherrschen heutzutage viele auch die gesprochene Sprache. Der große Durchbruch für den massenhaften Einsatz von sprachbasierten Dialogsystemen gelang 2011 mit Siri. Und auch Alexa und ihre Kollegen bei Google oder Microsoft werden mit zunehmendem Einsatz von Deep Learning immer besser: Sie verstehen Dialekte und Denglisch, können interpretieren, was wir sagen,

und sprechen mit einem natürlichen Tonfall. Bislang jedoch sind Gespräche mit Maschinen noch nicht so unbeschwerlich wie mit Menschen. Man muss genau zuhören, zügig antworten, laut und deutlich sprechen, beim Thema bleiben und Geduld bewahren. Aber viele Anliegen sind einfacher Natur und eignen sich prinzipiell für diese Technologie.

Und je besser die Chatbots werden, umso leichter können wir auch getäuscht werden und glauben, mit einem menschlichen Gegenüber zu sprechen. Möglicherweise entstehen dabei sogar emotionale Bindungen. Bei Xiaoice, einem Chatbot, der in China über 600 Millionen Nutzerinnen und Nutzer hat, ist das sogar durchaus gewünscht. Xiaoice kann über beliebige Bereiche plaudern und gibt vor, eine eigene Persönlichkeit zu haben: eine sympathische, intelligente, zugewandte Frau von 18 Jahren. Xiaoice speichert die Äußerungen des Gegenübers und entwickelt eine persönliche Beziehung, sodass in den Gesprächen auch Berufs- und Beziehungsprobleme besprochen werden. Insgesamt ist in asiatischen Ländern die Bereitschaft der Menschen höher, sich auf einen solchen Chatbot einzulassen.

## 4.2 Fähigkeiten und Komponenten

Für Chatbots begegnen einem viele weitere Bezeichnungen, die verschiedene Facetten ihrer Einsatzmöglichkeiten und Fähigkeiten betonen.

- **Funktion:** Mit Chatterbots kann man sich unterhalten, transaktionalen Bots kann man Anweisungen geben, mit Sprachsteuerung Geräte bedienen
- **Gerät:** Digitale oder virtuelle Assistenten können uns auf dem Smartphone, daheim, im Auto, am Arbeitsplatz oder im Kundenservice unterstützen

- **Kanal:** Messenger-Bots findet man in Messaging-Systemen und Social Bots in sozialen Medien, Telefonroboter erreicht man am Telefon
  - **Intelligenz:** Intelligente Bots sind meist lernfähig und sprachlich gewandter, Dialogsysteme können längere Gespräche führen
  - **Körper:** Ein conversational agent kann in einem Avatar oder Roboter stecken
  - **Autonomie:** Gesprächsautomaten, Kommunikationsautomaten oder Kommunikationsroboter automatisieren Dialoge zwischen Unternehmen und ihren Kunden oder Mitarbeitenden
  - **Branche:** In digitalen Spielen gibt es Game Bots und Legal Bots unterstützen in Kanzleien
  - **Medium:** Mit Voicebots kann man sprechen, mit Textbots nur texten
- Regel fertige Software, während die Module an der untersten Schicht auf die Anwendung spezialisiert werden müssen. Dabei hat man die Wahl zwischen Konstruieren und Trainieren. Konstruieren ist pflegeintensiv, auch wenn bei graphischen Editoren keine Programmierkenntnisse erforderlich sind (Stichwort: **low code** oder **no code**). Zwar hat man dabei alle Erwiderungen unter Kontrolle und kann unpassende Äußerungen verhindern, aber die Sprachbeherrschung ist insgesamt eher gering. Dagegen erfordert das weniger aufwendige Trainieren genügend und geeignete Textbeispiele als Basis. Die Sprachkompetenz ist dann sehr hoch, aber man hat die Erwiderungen nicht völlig unter Kontrolle. Heute sind hybride Lösungen gängig, die beide Ansätze verbinden. Deshalb sind Schnittstellen wichtig, damit man schrittweise alternative Verfahren implementieren kann. So lässt sich beispielsweise das Erkennen und Ausdrücken von Stimmungen und Emotionen leichter realisieren.

Ein Textbot muss geschriebenen Text verstehen, also Information zu Sprechakten, Absichten und Slots oder Fragen extrahieren, den Dialog managen und dabei eventuell angedockte Systeme und Geräte ansprechen, und schließlich einen Text für die Antwort formulieren.

Für gesprochene Sprache benötigt man einen Sprachassistenten, also einen Voicebot. Dabei ist der Unterschied zwischen Voicebots, mit denen man sprechen kann, und Textbots, mit denen man in einem Textfenster kommuniziert, gar nicht so groß: Ein Voicebot oder Sprachassistent ist im Kern ein Textbot, der zusätzlich hören und sprechen kann. Zum Hören benötigt er ein Mikrofon und automatische Spracherkennung, die das Audiosignal in Text überführt, zum Sprechen benötigt er eine akustische Sprachsynthese, um Text in ein Audio-signal umzuwandeln, sowie einen Lautsprecher.

Heutzutage bestehen Voicebots aus einer Reihe von Modulen entsprechend den Blöcken der Sprachprozesskette. An der obersten Schicht, für Spracherkennung und -synthese, verwendet man in der

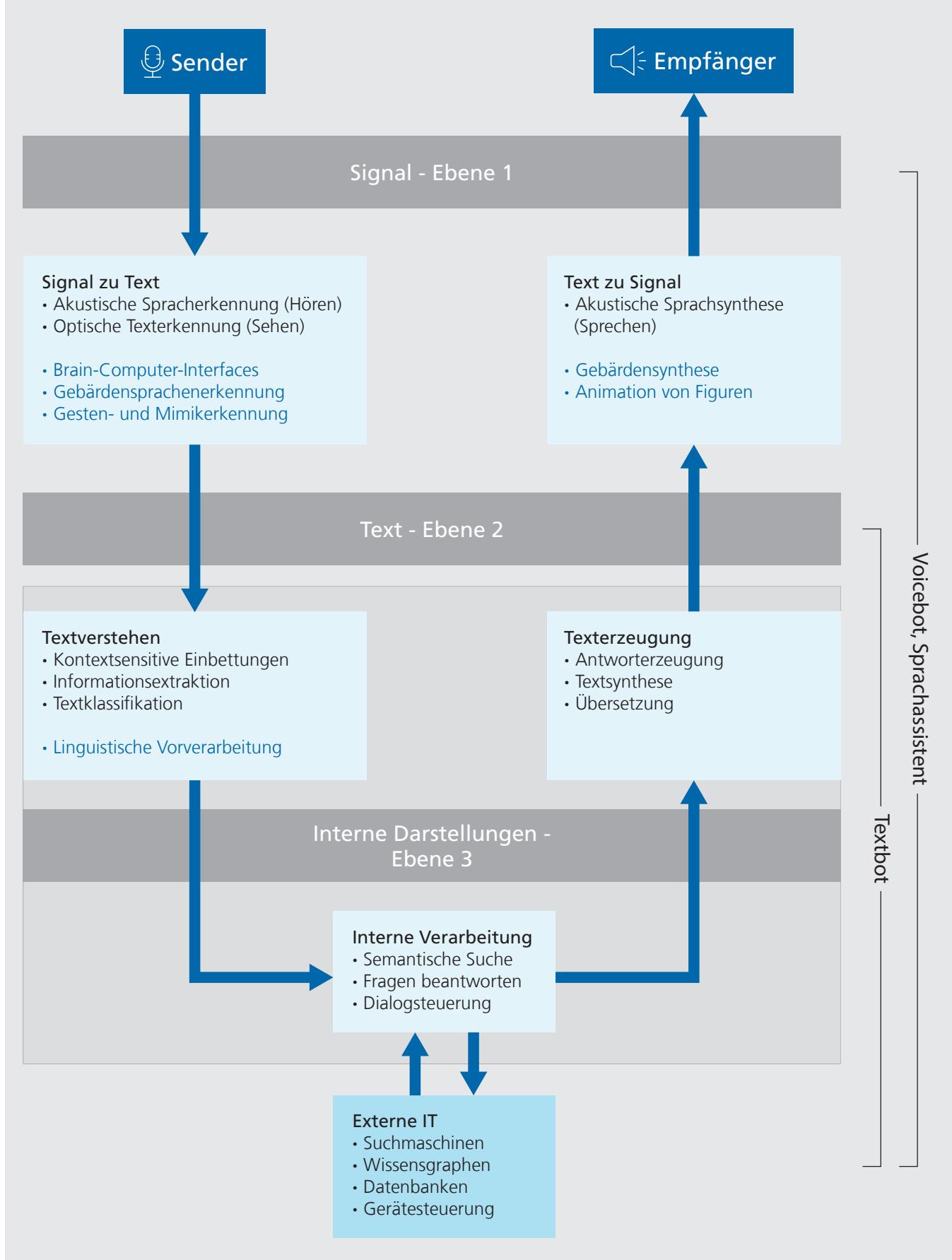


Abb. 10: Prozesskette der Voicebots und Textbots. Quelle: Eigene Darstellung

## 4.3 Beispiel Blender

Ein Beispiel ist der Chatbot Blender, mit dem man sich über beliebige Themen schriftlich unterhalten kann. Er basiert auf einem Transformer mit 9,4 Milliarden Parametern und wurde auf Social-Media-Konversationen trainiert, wobei die Eingabe aus den bisherigen Dialogbeiträgen besteht und die Ausgabe aus der nächsten Antwort. Der Name Blender deutet an, dass der Chatbot mehrere Ziele miteinander verfolgt: Entwicklung einer eigenen »Persönlichkeit«, Rückgriff auf große Wissensbasen und Zeigen von Empathie. Seine Antworten auf Testfragen wurden von menschlichen Dialogpartnern nach Sinnhaftigkeit, Informationsgehalt und der Berücksichtigung der persönlichen Beziehung bewertet. Es zeigte sich, dass sich die menschlichen Dialogpartner genauso gerne mit dem Chatbot wie mit einer echten Person unterhielten. In Zukunft soll Blender auch zu Fragen des Dialogpartners Hintergrundinformationen einholen und in seine Antworten einflechten. Aufgaben kann er noch nicht ausführen.

Auch wenn Blender engagierte Unterhaltungen führen kann, ist er immer noch kein perfekter Dialogpartner. Hin und wieder macht er widersprüchliche Äußerungen und kann die gleiche Phrase innerhalb einer Unterhaltung wiederholen. Manchmal kann er auch Fakten »halluzinieren«, welche in Wahrheit nicht richtig sind. An der Verringerung dieser Probleme wird aktiv geforscht. Das Programm steht als Open-Source-Produkt zur Verfügung.

Es bleibt abzuwarten, ob persönliche Assistenten wie Xiaoice oder Blender wirklich alle Nuancen der menschlichen Interaktion in Matrizen, Vektoren und tiefen Sprachmodellen als statistische Assoziationen und Gewichte kodieren können. Ein nicht zu unterschätzender Vorteil ist sicher die ständige Verfügbarkeit: Hat man so eine geduldige und einfühlsame Gesprächspartnerin wie Xiaoice, dann können Unterhaltungen nicht nur Stunden dauern, sondern auch zu jeder beliebigen Tages- und Nachtzeit in Gang gebracht werden. Ein realer Mensch hat nicht immer die Ausdauer, sich unentwegt die Interessen sowie die kleinen und großen Nöte seiner Mitmenschen anzuhören und ist zudem auch nicht immer erreichbar.

Blender

Nutzer

Hi how are you?

I'm good. How are you? Do you have a favorite food? Mine is lobster!

Doing well. My favorite food is cake. I just bought one because I got promoted at work!

Congratulations on the promotion. What kind of work do you do? I work in software.

Thanks so much, I just want to make my parents proud! I am an engineer.

I'm sure they will be very proud of you. What type of engineering do you work in?

I am an environmental engineer. What is it like to work in software?

It can be a lot of work but it is very rewarding. Do you enjoy your job?

I love it! It's very important to protect the environment.

That's great! Do you have hobbies? I like to play tennis in my spare time.

I'm not very coordinated so I don't play sports... I do like to play the guitar, though.

Abb. 11: Dialog mit Blenderbot (Facebook)  
Quelle: Roller et al. 2020

## 4.4 Anforderungen an die Entwicklung

Für die Entwicklung und den Betrieb von Chatbots gibt es viele Plattformen und noch mehr Anbieter. Bei der Auswahl ist eine ganze Reihe von Kriterien zu bedenken: Skalierbarkeit, Kosten, Kanalwechsel während eines Gesprächs, Übergabe an menschliche Agentinnen und Agenten, Analysedienste, Vorlagen, graphische Editoren, Autokonfiguration, Entwicklungsunterstützung, NLP- und KI-Komponenten. Besonders bedenkenswert sind folgende Kriterien:

- **Sprachen:** Chatbots für Deutsche müssen gut Deutsch können. Oft wird auch eine Reihe anderer Sprachen unterstützt.
- **Schnittstellen:** Schnittstellen in der Sprachverarbeitung ermöglichen es, NLP-Verfahren auszutauschen. Schnittstellen zur Unternehmens-IT sind wichtig, damit ein Chatbot überhaupt benötigte Informationen beschaffen oder die Aufträge des Kunden anstoßen kann.
- **Personalisierung:** Je mehr der Chatbot über die Person weiß, desto effizienter ist das Gespräch. Nachdem man sich identifiziert hat, sollte ein Chatbot auf Daten aus dem CRM-System zugreifen können, statt sie in dem Gespräch zu erfragen. Außerdem sollte er sich Informationen, die im Gespräch gegeben wurden, für die Dauer des Gesprächs merken und soweit nützlich, auch im CRM ablegen. Dies sollte mit dem Einverständnis der Kundinnen und Kunden erfolgen.
- **Dateneigentum:** In einem Gespräch kann man sehr viel über einen Kunden, seine Vorlieben, Ansichten und Gefühle erfahren. Diese Informationen kann man in Wert setzen. Nicht nur, um den Chatbot oder seine Angebote zu personalisieren, sondern auch, um über die Kundinnen und Kunden zu lernen oder Daten

für das Nachtrainieren des Bots zu gewinnen. Darum ist es wichtig, wer die Rechte an diesen Daten hat: man selbst, der Plattformbetreiber oder sogar die Nutzerinnen und Nutzer.

- **Datenschutz:** Gespräche enthalten viele persönliche Daten, die der Datenschutzgrundverordnung unterliegen. Sie dürfen nur zweckgebunden gespeichert werden und müssen sicher sein. Unter Umständen darf der Chatbot nicht bei einem Cloud-Anbieter laufen, sondern muss im hauseigenen Rechenzentrum betrieben werden.
- **Kontrolle der Weiterentwicklung:** Möchte man, dass Chatbots im Betrieb aus ihren Gesprächen weiterlernen, muss man Sicherungsmaßnahmen treffen. Der Chatbot Tay ist bekannt geworden, weil Microsoft ihn nach kurzer Zeit vom Netz nehmen musste. Tay hatte von seinen Gesprächspartnerinnen und -partnern eine rassistische und beleidigende Ausdrucksweise übernommen.
- **Portierung:** Ein Chatbot sollte über verschiedene Kanäle erreichbar sein. Der Trend geht dahin, eine angefangene Unterhaltung auf einem anderen Kanal fortsetzen zu können (Persistenz). Auf einer Entwicklungsplattform sollte man einen einmal gebauten Chatbot auf andere Kanäle portieren können, ohne ihn neu programmieren zu müssen.

Chatbots sollten effektiv, effizient und ansprechend sein, und diese Eigenschaften lassen sich messen. »Effektiv« bedeutet, dass man mit ihrer Hilfe erreicht, was man möchte (task completion). »Effizient« bedeutet, dass dies mit wenig Aufwand gelingt. Mit »ansprechend« ist die Qualität des Gesprächs gemeint.

## Chatbot-Technologie von KAUZ

kauz.net/technologie/chatbot-nlu

Mit der Sprachtechnologie von Kauz lassen sich Chatbots für den Kundenservice, den Verkauf oder die Beratung einfach und unkompliziert erstellen.

Die Chatbots funktionieren über das Internet und können mit Voice-Funktionen verbunden werden, z. B. mit dem Sprachassistenten von Amazon Echo. Kauz kombiniert eine NLU-Engine mit einem eigenen Lexikon und Deep-Learning, um eine interne Repräsentation zu erstellen, die anschließend maschinell weiterverarbeitet werden kann. Das Sprachverständnis ist eingebaut und muss nicht extra trainiert werden. Kauz kümmert sich zudem um neue Terminologien und Ausdrücke. Die NLU-Engine versteht auch bei komplexen Formulierungen den Kontext

und löst Rückbezüge richtig auf. Die Chatbots können ihr Sprachverständnis und ihre Antwortqualität außerdem permanent verbessern. Mit dem Chatbot-Editor lassen sich die Antworten des Chatbots auch einfach verwalten. Der Kunde entscheidet also selbst, wie der Chatbot auf bestimmte Problemstellungen reagiert. Weitere Eigenschaften sind Mehrsprachigkeit, Mehrkanal-fähigkeit und die Multimandantenpflege.

Kauz ist Partner des Beta-Programms von OpenAI. Um herauszufinden, inwiefern GPT-3 für den Einbau in Chatbots geeignet ist, hat Kauz den Horoskop-Bot entwickelt. Es ist ein hybrides Modell und vereint die Funktionalitäten eines Chatbots mit der Software GPT-3.

## Cognigy.AI von Cognigy

cognigy.com/products/cognigy-ai

Cognigy.AI ist eine Plattform des Start-ups Cognigy zum Bau von Textbots und Voicebots für Kunden und Mitarbeitende. 2020 gewann Cognigy.AI einen Vergleich zwischen 40 Dialogplattformen von Gartner, weil es als einzige alle Minimalanforderungen erfüllte. Darunter: Mehrsprachigkeit, NLP, graphische Entwicklungs-umgebung für Nicht-Programmierer, Datenbank-Zugriff und API. Besonders punkteten die Vielzahl der Kommunikationskanäle und die Möglichkeit, Module auszutauschen sowie die flexiblen Reaktionsmöglichkeiten, wenn kein eindeutiger Intent identifiziert werden kann. Hervorgehoben wurde außerdem der Support durch das Entwicklerteam.

Cognigy verfügt über ein Netzwerk an gut ausgebildeten lokalen und globalen Implementierungs- und IT-Systempartnerfirmen

mit den richtigen Branchen- und Themen-spezialisierungen. Darunter die Telekom, die Cognigy.AI in ihre Conversational AI Suite für automatisierten Kundenservice aufnimmt. 2021 wurde Cognigy gemeinsam mit seinem Technologiepartner Deutsche Telekom als einer der ersten Anbieter AIC4 geprüft. Der AIC4-Kriterienkatalog (Artificial Intelligence Cloud Services Compliance Criteria Catalogue) des Bundesamtes für Sicherheit in der Informations-technik (BSI) ist der weltweit einzige anerkannte, relevante Sicherheitsstandard für KI-basierte Anwendungen.

PricewaterhouseCoopers (PwC) hat die AIC4-Prüfung von Cognigy.AI entsprechend den BSI-Vorgaben durchgeführt – und Vorurteilsfreiheit und Transparenz, Sicherheit und Belastbarkeit, Zuverlässigkeit sowie Datenqualität getestet.

## Expertenmeinung

### Dr. Benjamin Geißler, Gothaer Finanzholding AG



**Gothaer**

»Das Potenzial von Audio liegt nicht so sehr im direkten wirtschaftlichen Nutzen, sondern in dem Umstand, dass es die natürlichste Art der Kommunikation für uns Menschen ist.«

Dr. Benjamin Geißler ist Senior Projektleiter für Künstliche Intelligenz bei der Gothaer Finanzholding AG und dort u.a. für die Einführung einer konzernweiten Sprach- und Chatbot-Plattform verantwortlich.

**Was waren für Sie die wichtigsten Durchbrüche in Sprachtechnologien und NLP in den letzten fünf Jahren insbesondere in Bezug auf die Auswirkungen für die wirtschaftliche Wertschöpfung?**

Grundsätzlich Deep-Learning-Modelle, bei Textverarbeitung natürlich BERT und Co. Bei Sprachtechnologien eigentlich primär das großflächige Aufkommen der digitalen Assistenten (Alexa, Siri, Google Home), die das in guter Qualität können.

**Wo setzen Sie Sprachtechnologien/NLP bei sich im Unternehmen ein?**

NLP setzen wir klassisch im Input-Management ein, um zu verstehen, was uns der Kunde, Makler etc. schickt, um entsprechend darauf reagieren zu können und um in Teilen auch schon Daten daraus zu extrahieren. Außerdem nutzen wir NLP bei der Bearbeitung von Texten, z.B. beim Vergleich von Verträgen und Formulierungen. Im Sprachbereich automatisieren wir einfache Kundengespräche und wollen perspektivisch auch die Kundenanfrage richtig routen. Ein weiteres Beispiel ist das Voice-Bot-gestützte Abmelden eines Autos.

**Welche wirtschaftliche Bedeutung hat der Einsatz von Sprachtechnologien für Ihr Unternehmen?**

Eine sehr hohe. Da wir als Versicherer keinerlei »greifbare« Produkte haben, erfolgt alles über Sprache, sowohl schriftlich als auch telefonisch. Im schriftlichen Bereich wird schon länger viel automatisiert verarbeitet, im Audiobereich kommt das aktuell erst und bietet somit noch einen deutlich größeren Effizienzhebel, da hier bisher so gut wie nichts automatisiert wurde. Neben den klassischen Aspekten wie Qualität und Effizienz spielt zum Beispiel beim Input-Management das Thema Schnelligkeit eine Rolle, weil schriftliche Anfragen von Kundinnen und Kunden deutlich schneller in der Sachbearbeitung ankommen.

**Was sind dabei aus Ihrer Sicht die größten Herausforderungen und welche Use Cases besitzen besondere Potenziale?**

Bei Text ist es die Klassifikation der Dokumente, die am einfachsten und am besten skaliert werden kann. Datenextraktion und Verständnis sind deutlich aufwendiger, da hier noch in stärkerem Maße die eigenen Spezifika berücksichtigt werden müssen. Primäre Herausforderung ist die Beschaffung von guten, passenden und sauber annotierten Trainingsdokumenten, ohne die die KI nicht lernen kann. Bei Audio ist es genau umgekehrt. Hier sind sehr spitze und klar abgegrenzte Use Cases am vielversprechendsten, da man mit diesen recht schnell und einfach lernen kann, wie der Kunde spricht. Neben der eigentlichen Technik, d.h. die Audiostreams korrekt zu transkribieren, liegt eine große Herausforderung darin, den Anrufer am Telefon so zu steuern, dass er die Antworten, die die Maschine braucht, möglichst in der Art und Weise liefert, wie sie die Maschine auch verwenden muss, ohne dass der Anrufer das Gefühl hat mit einem völlig »starren und dummen« System zu sprechen. Nur wenn das gut funktioniert, kann man den Anrufer davon überzeugen, so etwas zu nutzen. Das größte Potenzial bieten klassische Operations Use Cases. Der Einsatz im Vertrieb ist deutlich

*komplexer, zumal man hier auch nicht immer automatisieren möchte.*

### **Welche ersten Schritte sollte ein Unternehmen gehen, um konkret Sprachtechnologien zu etablieren?**

*Aus unserer Erfahrung sollte man Schritt für Schritt vorgehen. Wichtige Punkte sind: Technologie, Potenziale und Business Value verstehen, Datenschutz und Co. klären, die richtigen Systeme und Partner auswählen bzw. aufbauen, und dann vor allem schnell erproben und bei Erfolg skalieren.*

*Außerdem sollte man beachten, dass jeder KI-Use-Case eingebettet ist in einen größeren Geschäftskontext aber auch den Kundenkontext. So steht z.B. die Telefonnummer für ein spezifisches Thema auf einem Kundenbrief. Unter dieser Nummer rufen die Kunden aber nicht nur an, wenn sie genau zu dem Thema eine Frage haben, sondern auch, wenn sie einfach irgendeine Anfrage haben. Hier gilt dann für uns das Prinzip: »Der Kunde ist immer richtig bei uns, wir müssen schauen, dass wir ihn richtig routen«. Und das ist dann eine Herausforderung, die es bei der Entwicklung der KI zu berücksichtigen gilt.*

### **Was bedeuten Sprachmodelle wie GPT-3 für Europa? Und was bedeuten sie für die Weiterentwicklung und Forschung in Ihrem Unternehmen?**

*Solange GPT-3 (bzw. andere Sprachmodelle) noch nicht auf Deutsch in ausreichender Qualität verfügbar ist, ist es für uns nicht interessant. Zumal die Texte, die bei uns erstellt werden, meistens inhaltlich komplex sind (Verträge etc.). Da fällt es einer KI schwer, das in guter Qualität selbst zu erstellen und dann auch noch so, dass es jemand anderes versteht und es auch rechtlich korrekt ist.*

*Als Nicht-Technologie-Anbieter nutzen wir nur Services im Bereich der Sprachtechnologien und partizipieren entsprechend an den Weiterentwicklungen der Technologiekonzerne. Man*

*sieht aber schon, dass die Ergebnisse der neusten Forschung immer erst einmal in Englisch und Chinesisch verfügbar sind. Für andere Sprachen, insbesondere Deutsch, braucht man immer ein wenig Geduld.*

### **Wo sehen Sie die Stärken und Schwächen von Deep Learning in der praktischen Anwendung von Sprachtechnologien und NLP in der Praxis – und gibt es Gebiete, in denen Deep Learning nicht eingesetzt werden sollte?**

*Stärken sind klar die aktuell sehr guten Ergebnisse. Die Schwächen sind, dass die eigentlichen Ideen und Konzepte, die dahinter stecken, sehr alt sind (1950-60) und nur durch die aktuell verfügbaren Datenmengen und Rechenleistungen realisiert werden können. Gerade Deep Learning benötigt für gute Ergebnisse ja beides in großen Mengen, was aus verschiedenen Blickwinkeln langfristig nicht gut ist (z.B. hoher Energieverbrauch, Daten sammeln, teuer). Hier braucht es, glaube ich, grundlegend neue Ansätze, um effizienter zu werden. Deep Learning ist eines von vielen Werkzeugen, das man in vielen Bereichen einsetzen kann. Man muss nur wie bei allem darauf achten, wie man es einsetzt und sich gerade bei KI-Systemen der zentralen Stolperfallen bewusst sein (Bias, Wahrscheinlichkeiten, »shit in shit out«).*

### **Was denken Sie, was können wir in zehn Jahren mit Sprachtechnologien erreichen und welche wirtschaftlichen Potenziale kann der Einsatz von Sprachtechnologien noch entfalten?**

*Das Potenzial von Audio liegt nicht so sehr im direkten wirtschaftlichen Nutzen, sondern in dem Umstand, dass es die natürlichste Art der Kommunikation für uns Menschen ist und es uns durch Audiotechnologien noch leichter fallen wird, verschiedene Dinge und Services zu konsumieren. Das hilft, die Kluft zwischen Kunden und Dienstleistern auf einfache Art und Weise zu überbrücken und das in einer Vielzahl von Situationen, gerade bei sehr einfachen Dingen.*

## Leseempfehlungen

### **Chatbots: The Definitive Guide**

Artificial Solutions (Hrsg.) (2020)

»In this in-depth guide you'll learn: What chatbots are & how they work; the must have chatbot features, the value of chatbots for business; and lots more. So, if you want to become a chatbot pro, this guide is for you«

Ein illustrativer Einstieg in Chatbots. Nur die Passagen über KI sind kritisch zu lesen.

Verfügbar unter: <http://marketing.artificial-solutions.com/rs/177-TDV-970/images/Chatbots-the-definitive-guide-2020.pdf>

### **Marketing mit Sprachassistenten**

Thomas Hörner (2019), Springer Gabler

»Mit diesem Buch lernen Sie, wie Sie Sprachassistenten im Marketing Ihres Unternehmens nutzen. Es liefert viele praktische Hilfestellungen, mit denen Sie Ihren Geschäftsbetrieb effizienter gestalten.«

Verfügbar unter: <https://sprachassistenten.marketing/literatur/buch-marketing-mit-sprachassistenten/>



# 5 Bedeutung für die Wirtschaft

Bereits heute nutzen Unternehmen Sprachtechnologien, um interne Prozesse zu optimieren, für die Ansprache ihrer Kundinnen und Kunden und als Schnittstellen zu Geräten. Die Boxen in dieser Studie haben Beispiele dafür aus Nordrhein-Westfalen zusammengetragen. Sie zeigen zudem, dass die Software oftmals als cloudbasierter Dienst angeboten wird. Einige Dienste können ohne weitere Anpassung zum Einsatz kommen, andere erbringen Teilaufgaben, aus denen komplexere Anwendungen zusammengesetzt werden. Bei NLP-Anwendungen ist in der Regel eine Anpassung für den Einsatz in Unternehmen erforderlich. Chatbots bilden eine besondere Klasse von NLP-Anwendungen, da sie auf bestimmten Kanälen kommunizieren und dafür auf Plattformen betrieben werden müssen. NLP-Werkzeuge und Entwicklungsumgebungen stellen Software und Services zur Verfügung, um Chatbots und andere NLP-Anwendungen zu bauen. In diesem Kapitel wird aufgezeigt, welche Einsatzbereiche es in Unternehmen gibt und welche Chancen und Potenziale sich dadurch für die Wirtschaft ergeben. Ein besonderer Fokus liegt abschließend auch darauf, wie der Einsatz von Sprachtechnologien vertrauenswürdig gestaltet werden kann.

## 5.1 Wirtschaftliche Potenziale

Den Technologien rund um Sprache werden in mehreren Studien große finanzielle Potenziale zugesprochen. Allein das Marktvolumen von Anwendungen zur Spracherkennung wurde 2019 noch auf 10,7 Mrd US Dollar beziffert, soll aber laut Prognose für 2025 auf fast 27,2 Mrd US Dollar ansteigen (Mordor Intelligence 2020).

Auch die Anzahl der Nutzerinnen und Nutzer von Sprachassistenten wird stetig steigen. Die Märkte für Anwendungen werden daher weiterwachsen, auch dank technologischer Weiterentwicklungen und der zunehmenden Vernetzung im Internet der Dinge (IoT). Allein 2020 nutzten Menschen

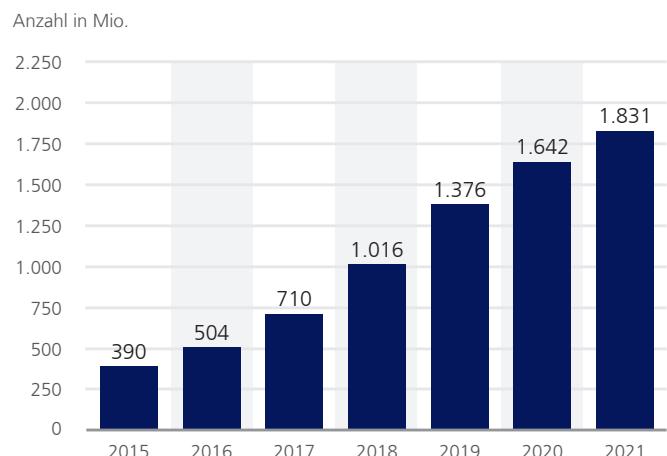


Abb. 12: Prognose weltweite Nutzung von digitalen Sprachassistenten bis 2024. Quelle: Vailshery 2021

weltweit 4,2 Mrd digitale Sprachassistenten (Hochrechnung aus 2019). Bis 2024 soll sich die Zahl auf 8,4 Mrd Nutzungen verdoppeln, d. h. viele Menschen werden mehrere digitale Sprachassistenten parallel nutzen (Vailshery 2021).

Andererseits scheint sich die Akzeptanz von Sprachassistenten zurzeit noch stark auf den privaten Lebensbereich zu beschränken. Dabei ist die Sprachinteraktion mit mobilen Endgeräten führend (GfK SE & ZVEI 2017).

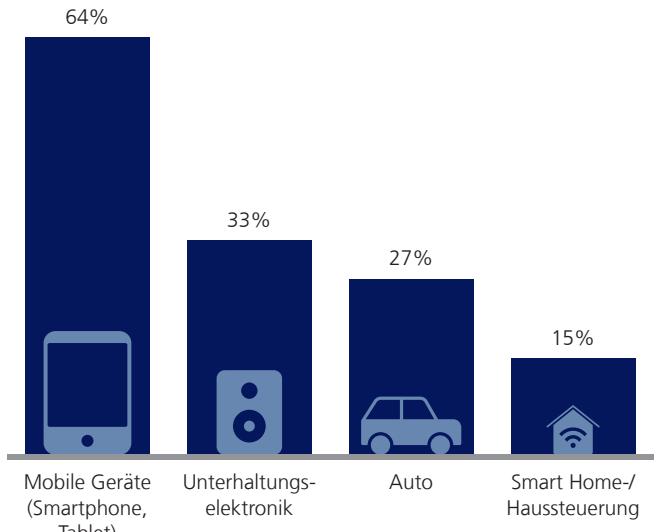
Diese hohen Nutzungszahlen im privaten Lebensbereich können mehrere Ursachen haben. Zum einen ist hier die Verfügbarkeit von sprachverarbeitenden Applikationen deutlich größer. Gleichzeitig ist anzunehmen, dass die Nutzungsanforderungen im beruflichen und B2B-Segment komplexer sind, man denke an Fachbegriffe, Hintergrundgeräusche im Shopfloor, im Produktionsbereich oder an Datenschutzvorgaben.

Dies bestätigt eine weitere Studie, die Genauigkeit und fehlende Dialekterkennung als überwiegende Barrieren für die Einführung von Sprachtechnologien in Unternehmen anführt. Hohe Kosten, fehlende Anwendungssprachen und Komplexität der Einführung werden als weitere Barrieren benannt (Speechmatics 2021).

Auch bei der Kundenansprache geben noch gut 38 Prozent an, lieber 15 Minuten auf einen menschlichen Berater oder eine Beraterin zu warten, als mit einem Bot zu interagieren (KPMG 2019). Das bedeutet, dass zwar eine gewisse Bereitschaft zur Nutzung von Chatbots gegeben ist, die Akzeptanz bei Kundinnen und Kunden aber durchaus noch ausbaufähig ist.

Einer Befragung in Deutschland zufolge können sich viele Menschen vorstellen, Assistenzsysteme (Spracheingabe) im Kontext von einfachen Bankgeschäften zu verwenden, beispielsweise um einen Termin zu vereinbaren oder einen Geldautomaten zu finden (Diesel 2020). Dabei sind genannte Hemmnisse Datensicherheit, Zuverlässigkeit und Einfachheit der Nutzung, die zu berücksichtigen sind.

So auch in den USA: In einer amerikanischen Umfrage unter 400 Entscheidungsträgern wurde



Frage: Bei welchen Geräten nutzen Sie die Sprachsteuerung?  
Befragung von 396 Nutzern von Sprachsteuerung ab 14 Jahren in Deutschland im Oktober 2017

Abb. 13: Gerätespezifische Nutzung von Sprachsteuerungen.  
Quelle: GfK SE & ZVEI 2017

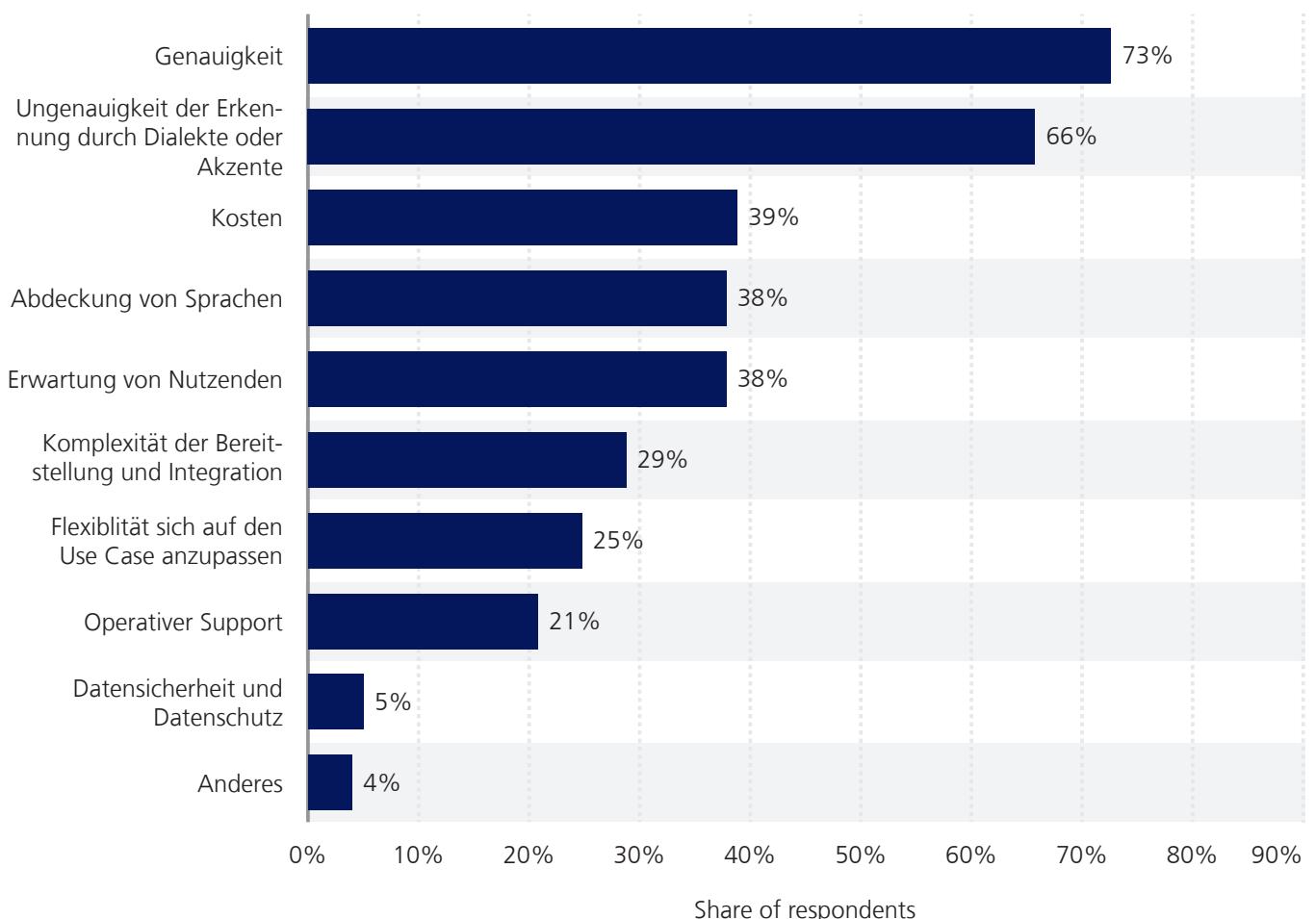


Abb. 14: Hindernisse für die Einführung von Sprachtechnologie weltweit ab 2020. Quelle: Speechmatics 2021

## Kundenpräferenzen zum KI-Einsatz im Alltag

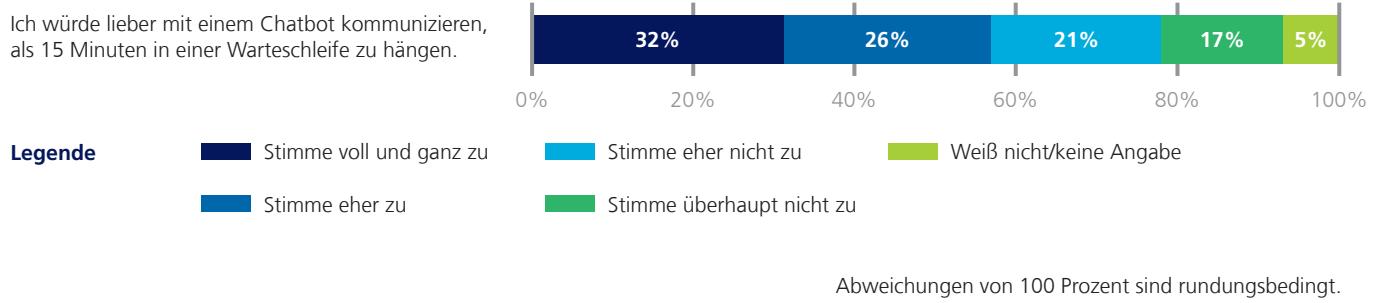


Abb. 15: Interaktion Chatbot vs Mensch. Quelle: KPMG 2019

herausgefunden, dass sich die präferierte Anwendung von sprachgesteuerten Interaktionen auf einfache Tätigkeiten wie Käufe, Order-Nachverfolgung oder Suchfunktionen beschränkt (Adobe & Advanis 2019a). Gleichzeitig planen über 90 Prozent der befragten Personen, signifikant in Sprachtechnologien zu investieren oder die bestehenden Investitionen weiter zu erhöhen (Adobe & Advanis 2019b).

Die wirtschaftlichen Zukunftsaussichten von Sprachtechnologien sind bereits heute nicht von der Hand zu weisen. Sie besitzen noch ein großes

Durchdringungspotenzial, werden zunehmend stärker auch im B2B-Umfeld eingesetzt und sind bei weitem noch nicht an ihrem Einsatz-Zenit angekommen. Auch unter Berücksichtigung von der teilweise noch eingeschränkten Nutzungsbereitschaft, wird davon ausgegangen, dass die Nutzung exponentiell zunehmen wird. Die Herausforderung dabei besteht insbesondere in der Schaffung von Sprachtechnologielösungen, die die Anforderungen, Bedürfnisse und Wünsche der Nutzerinnen und Nutzer berücksichtigen.

## Sprachassistenten könnten bald den Finanzbereich erobern

Anteil der Befragten, die sich folgende Finanzdienste über einen Sprachassistenten vorstellen können (in %)

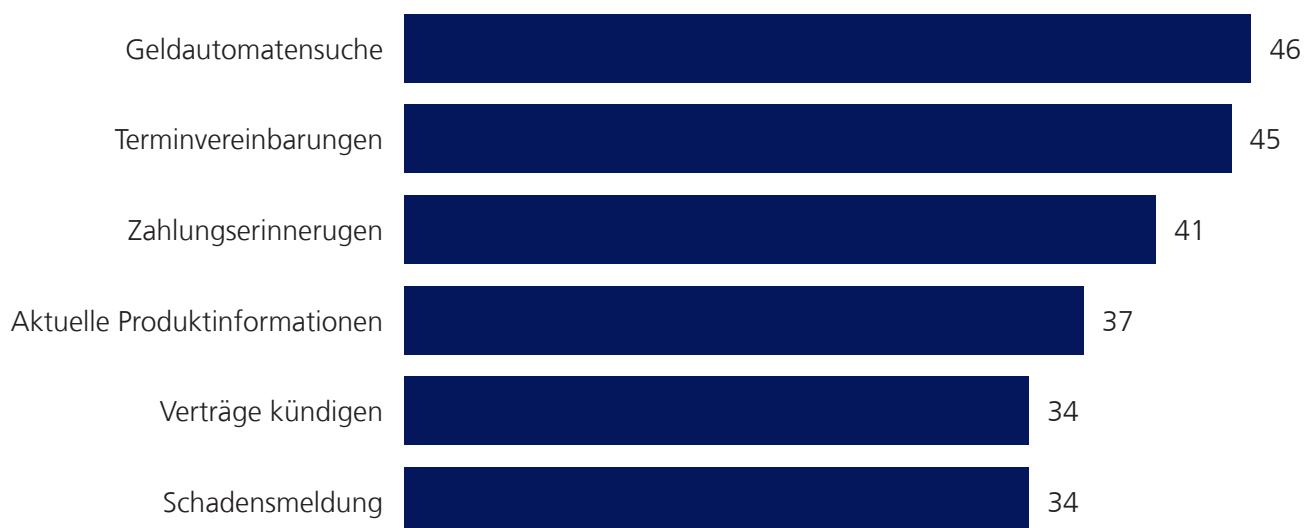


Abb. 16: Finanzdienstleistungen als Beispiel für den Einsatz von Sprachassistenten. Quelle: Diesel 2020



## 5.2 Einsatzbereiche im Unternehmen

Neben den branchenunabhängigen Einsatzmöglichkeiten lassen sich auch branchenspezifische Anwendungen mit hohem Potenzial für

Sprachtechnologien identifizieren. Dabei ist zu überlegen, zu welchem Zweck und an welcher Schnittstelle sie kurz-, mittel- und langfristig eingesetzt werden sollen. Wie bereits in den Expertenmeinungen angeführt, bieten die Bereiche der Organisations- und Verwaltungsprozesse große Potenziale zur Vereinfachung. Durch den Wegfall von repetitiven

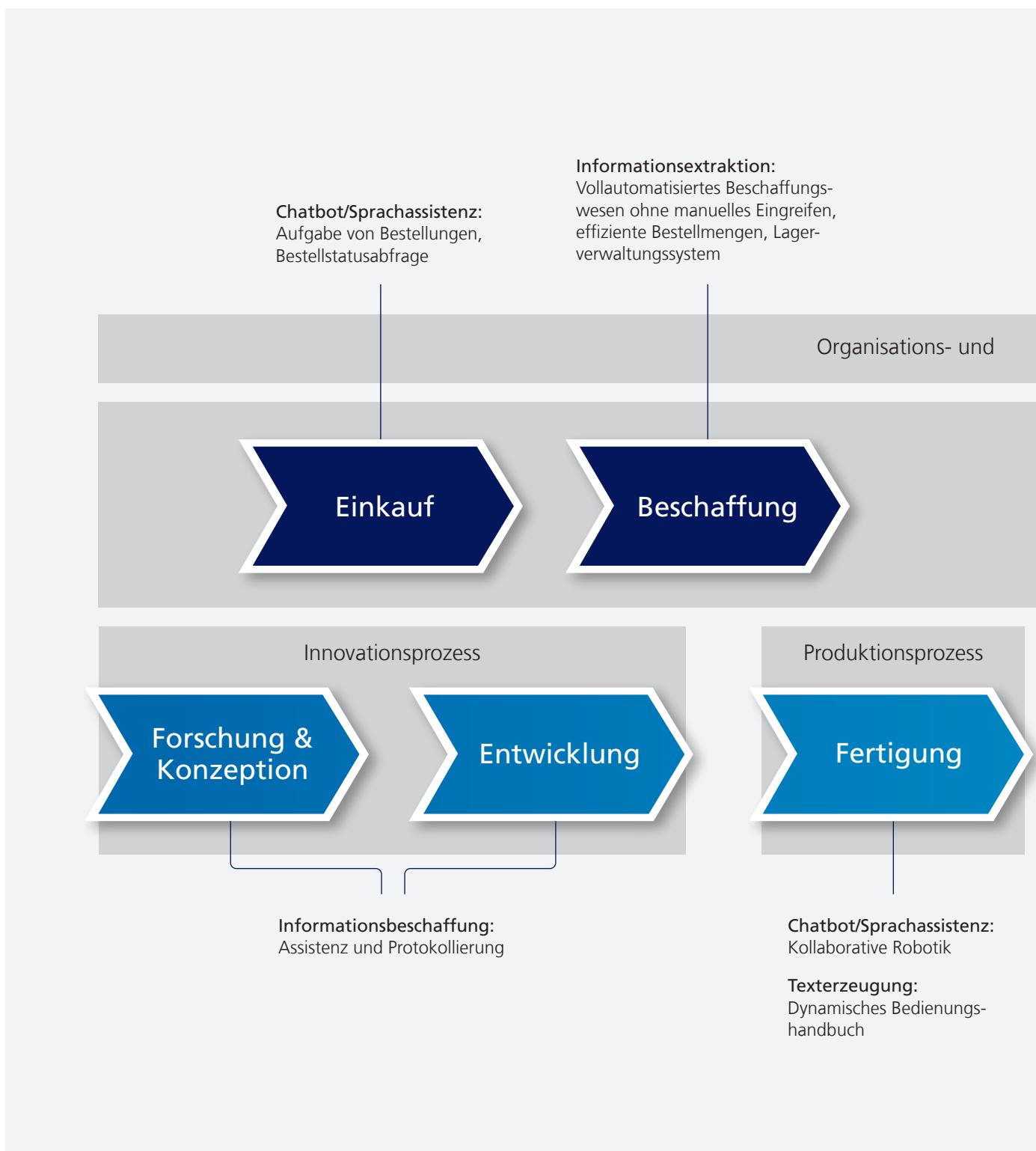


Abb. 17: Organisations- und Unterstützungsprozesse. Quelle: Eigene Darstellung

Tätigkeiten bei gleichzeitiger Übernahme höherwertiger Aufgaben, kann sich die Arbeitsqualität und -zufriedenheit insbesondere aus Sicht der Arbeitnehmerinnen und Arbeitnehmer verbessern. Beispiele sind Sprachassistenten und Chatbots, die einen Teil der Anfragen automatisch beantworten können. Neben dem Einsatz in Verwaltungs- und

kundennahen Prozessen können Sprachtechnologien auch große Potenziale in den primären, direkt wertschöpfungsbezogenen Aktivitäten entfalten. Einen Denkanstoß mit ersten Einsatzmöglichkeiten für das eigene Unternehmen bietet die nächste Grafik.

**Informationsextraktion:**  
Dokumentenmanagement,  
Rechnungsautomatisierung,  
Vertragsmanagement, KPI-Extractor

**Informationsbeschaffung:**  
Automatisierte Listenprüfung

**Texterzeugung:**  
Diktierfunktion mit Fachbegriffen  
und konkreten Formulierungen

**Informationsextraktion:**  
Belegverteilung, Jobmatching

**Chatbot/Sprachassistenz:**  
Bewerbungsberatung,  
Gesundheitscoach,  
Lernassistenz

**Chatbot/Sprachassistenz:**  
Empfangsroboter oder  
EDV-Support

Unterstützungsprozesse

Legal

Personal

IT & EDV

Verkaufs- und Distributionsprozesse

Marketing &  
Vertrieb

Logistik

Kundendienstprozess

Service &  
Kunden-  
dienst

**Informationsextraktion:**  
Meinungs- und Themen-  
monitoring, Verkaufsberatung

**Chatbot/Sprachassistenz:**  
Verkaufsberatung

**Informationsextraktion:**  
automatisierte Fallbearbeitung

**Informationsextraktion:**  
Beschwerdemanagement, automatisierte  
Fallbearbeitung, Kundenanfragen  
kategorisieren

**Informationsbeschaffung:**  
Assistenz im Callcenter, Auftragsklärung

**Chatbot/Sprachassistenz:**  
Aftersales-Beratung

## Expertenmeinung

### Sebastian Glock, Cognigy



**COGNIGY**

»Das Aufsetzen und Betreiben von NLP und KI-Anwendungen ist nicht mehr allein IT-Fachleuten und Data Scientists vorbehalten, sondern steht breiten Anwenderschichten offen. Das öffnet KI-Applikationen für völlig neue Anwendungsbereiche und wird in den nächsten Jahren dafür sorgen, dass NLP in immer mehr Bereichen alltäglich genutzt wird.«

Sebastian Glock ist Senior Technology Evangelist bei Cognigy, einem führenden Anbieter für Conversational AI. Sebastian arbeitet an der Schnittstelle zwischen Technologie und Markt und ist bei Cognigy verantwortlich für den Bereich Product Marketing.

**Was waren für Sie die wichtigsten Durchbrüche in Sprachtechnologien und NLP in den letzten fünf Jahren insbesondere in Bezug auf die Auswirkungen für die wirtschaftliche Wertschöpfung?**

*NLP hat in den letzten Jahren dramatische Fortschritte erzielt, insbesondere in den Bereichen Speech-to-Text, also dem Verstehen gesprochener Sprache sowie in der zuverlässigen Erkennung von Nutzerabsichten in frei formulierten Äußerungen. Den größten wirtschaftlichen Impact hat allerdings die Demokratisierung von KI-Technologie durch Low-Code.*

Was sind dabei aus Ihrer Sicht die größten Herausforderungen und welche Use Cases besitzen besondere Potenziale (z.B. in den Kriterien Niedrigschwelligkeit, Akzeptanz und Skalierbarkeit)?

*Die Akzeptanz von Sprachtechnologie ist besonders hoch, wenn sie einen echten Mehrwert liefert. Ein gutes Beispiel ist Sprachbedienung im Auto: Sie funktioniert zuverlässig, bequem und macht die Fahrt sicherer. Mit modernen Tools können vergleichbar starke Use Cases auch im B2B- und B2C-Umfeld so umgesetzt werden, dass Nutzer begeistert sind und die Einführung solcher Anwendungen wirtschaftlich schnell rentabel ist. Das größte Potenzial hat der Einsatz von Voice-Bots: Gesprochene Sprache ist die natürlichste Form menschlicher Kommunikation und ist als Kanal zur Kunden- und Mitarbeiter-Kommunikation längst nicht ausgeschöpft.*

**Welche ersten Schritte sollten Unternehmen gehen, um konkret Sprachtechnologien zu etablieren?**

*Unternehmen können heute aus einer Vielzahl von schlüsselfertigen Anwendungen wählen. Entscheidend für eine erfolgreiche Etablierung von Sprachtechnologie sind vor allem drei Aspekte:*

- 1. Unternehmen dürfen Sprachtechnologie nicht als »Chatbot-Projekt« oder Insellösung begreifen. Use Cases für Sprachanwendungen finden sich in allen Unternehmensbereichen. Nur ein Plattform-Ansatz hebt abteilungsübergreifende Synergien und vermeidet Silos.*
- 2. Der Einsatz von Sprachtechnologie sollte von Anfang an cross-funktional begleitet werden: Wenn Fachabteilungen, Marketer und IT-Experten in agilen Teams gemeinsam arbeiten, entstehen die besten Lösungen.*
- 3. Durch kontinuierliche Messung der Nutzung, des Erfolgs und des Nutzerverhaltens wird Sprachtechnologie dauerhaft erfolgreich und zum festen Bestandteil von Prozessen und Services.*

Wo sehen Sie die Stärken und Schwächen von Deep Learning in der praktischen Anwendung von Sprachtechnologien und NLP in der Praxis – und gibt es Gebiete, in denen Deep Learning nicht eingesetzt werden sollte?

*Deep Learning zeigt beeindruckende Leistungen bei der Verarbeitung großer Datenmengen und der Analyse von Mustern. Im Kontext von Conversational AI hilft Deep Learning bei der Erkennung von Nutzerabsichten, also den Zielen, die Nutzer mit frei formulierten Anfragen erreichen wollen. Limitierungen ergeben sich z. B. aus der Frage der Erklärbarkeit: Algorithmische Entscheidungen, die auf Deep Learning basieren, sind nicht mehr wie bei einem trivialen Programmablauf nachvollziehbar. Das ist insbesondere dann problematisch, wenn z. B. menschliche Vorurteile aus Trainingsdaten via Deep Learning ihren Weg in KI-Systeme finden. KI-Entwickler und Anwender müssen sich diese Risiken bewusst machen und beim Einsatz von Deep Learning berücksichtigen.*

Was denken Sie, was können wir in zehn Jahren mit Sprachtechnologien erreichen und welche wirtschaftlichen Potenziale kann der Einsatz von Sprachtechnologien noch entfalten?

*Conversational AI – und Sprachtechnologie als ein zentraler Bestandteil – wird in zehn Jahren einen Reifegrad erreicht haben, der völlig natürliche Dialoge zwischen Mensch und Maschine alltäglich macht. Das Verstehen gesprochener und geschriebener Sprache wird menschenähnliches Niveau erreichen und es in einigen Kontexten sogar übertreffen. Virtuelle Agenten, z. B. im Kundenservice, werden nicht nur präzise, individuelle Antworten auf fast alle Fragen geben, sie werden auch den Großteil aller Anliegen schnell und kompetent lösen. Immer mehr positive Erfahrung mit exzellenten digitalen Services per Sprache wird einen Image-Wandel beschleunigen, bei dem Bots von »Hilfsarbeitern« zu voll akzeptierten Experten werden. Das »Reden mit Maschinen« wird in allen Lebensbereichen zum Alltag gehören: Von virtuellem Service, über KI-Unterstützung in Standardsoftware bis zu sprachgesteuerten IoT-Devices.*

## 5.3 Branchen mit hohem Potenzial

Neben diesen branchenunabhängigen Einsatzmöglichkeiten lassen sich auch branchenspezifische Anwendungen für Sprachtechnologien identifizieren. Besonders hoch ist das Potenzial in der Medien-, Gesundheits-, Automobil- und Versicherungsbranche, in Handel und Marketing, im E-Commerce und Verwaltungssektor sowie in Rechts- und Steuergesellschaften.

### Handel und Marketing

Im Online-Handel (E-Commerce) werden Sprachassistenten vor allem im Kaufprozess eingesetzt. Dabei ist zu unterscheiden, ob es sich um einen Versorgungskauf oder um einen Erlebniskauf handelt. Versorgungskäufe müssen so effizient und einfach wie möglich sein. Die Kaufentscheidung und -abwicklung soll möglichst schnell und ohne große Wahlmöglichkeiten abgewickelt werden. Da bei wiederholten Bestellungen Produkte und Anbieter bekannt sind, sind die Ansprüche an die

Sprachassistenten vergleichsweise beherrschbar. Bei einem Erlebniskauf ist das anders. Die Kundinnen und Kunden erwarten oftmals mehr Informationen und Wahlmöglichkeiten als bei einem Versorgungskauf. Windowshopping, das Stöbern, das Entwickeln von Kaufkriterien und das Vergleichen von Produkten, welche zu einem Erlebniskauf aus Kundenperspektive gehören, sind technologisch abzubilden. Um Produkte zu empfehlen, können Sprachassistenten Vorzüge aus Produktbeschreibungen extrahieren, auf Basis von vorangegangenem Verhalten Empfehlungen aussprechen oder Bedarfe im Dialog identifizieren. Chatbots können die komplette Kundeninteraktionskette (Produktsuche, Bestellung, Versandinformationen und Kontakt sowie Aftersales) abdecken. Hierfür gibt es etablierte Plattformen, die auf das jeweilige Unternehmen anpassbar sind. Mehrsprachigkeit ist ebenfalls möglich.

### Medien

Sprachmodelle machen rasante Fortschritte und finden gerade im kreativen Bereich immer neue Einsatzmöglichkeiten. Für die Medienbranche wird

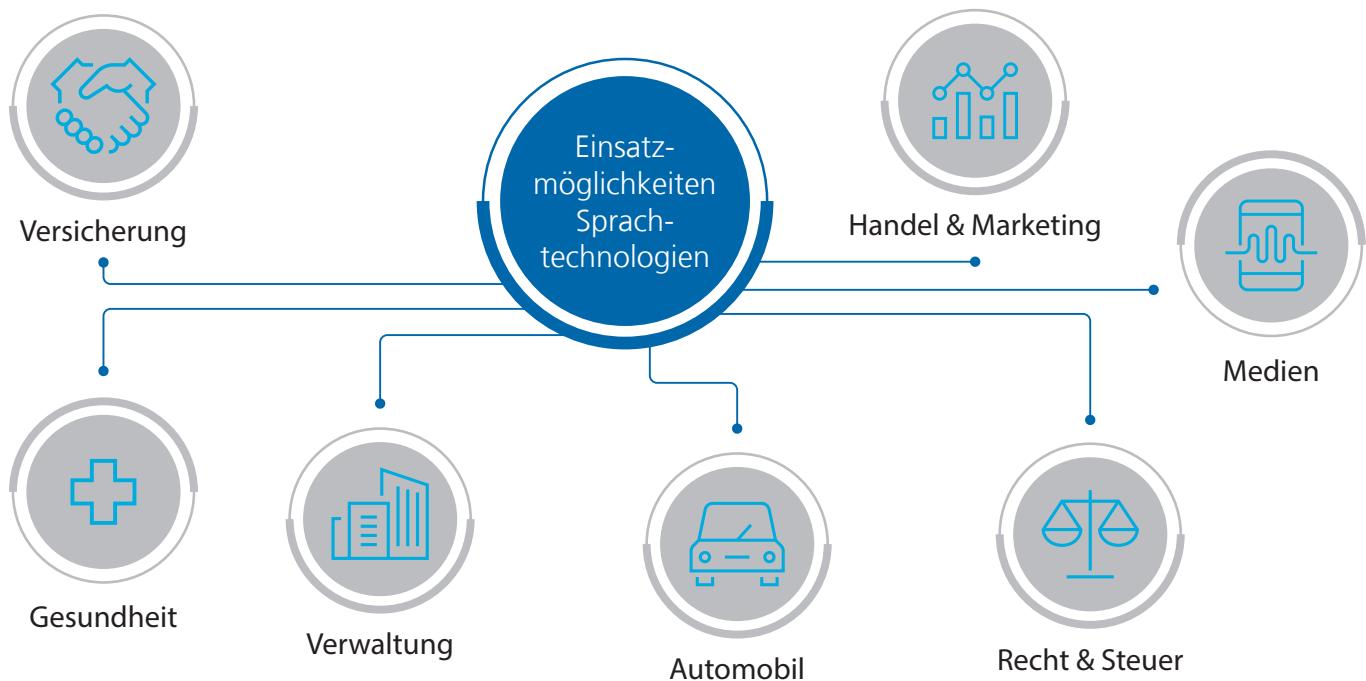


Abb. 18: Ausgewählte Anwendungsbereiche für Sprachtechnologien gruppiert nach wirtschaftlichen Branchen.  
Quelle: Eigene Darstellung

diese Entwicklung von besonderer Bedeutung sein und viele Arbeitsprozesse grundlegend ändern. Schon jetzt können Texte für Aktien- und Sportnachrichten vollautomatisch generiert werden. Für die Leserinnen und Leser ist dabei nicht mehr zu erkennen, ob der Text von einem Menschen oder einer Maschine verfasst wurde (»Robot Journalism«). Hauptsächlich wird die KI-basierte Texterzeugung aber zur Unterstützung von Journalistinnen und Journalisten zum Einsatz kommen: Beispielsweise können sie sich auf die Recherche konzentrieren und ihre Ergebnisse im Anschluss von einer KI zusammenfassen lassen. Bei der Auswahl von O-Tönen oder Textpassagen kann die semantische Suche helfen, indem sie ganze Medienarchive automatisiert durchsucht und passende Inhalte ausspielt.

Da sich Sprache nicht nur in Texten manifestiert, sondern auch in Audio- und Videodateien, betrifft die automatisierte Sprach- und Textverarbeitung letztlich die gesamte Medienproduktion. Davon werden also auch Synchron-, Musik-, Video- und Spielestudios profitieren. Mit Sprachtechnologien wird es auch einfacher, personalisierte Inhalte zu generieren, passgenaue Konsumvorschläge zu unterbreiten und sogar Medieninhalte dynamisch zu erzeugen, um das Nutzererlebnis, das Engagement, die Zufriedenheit und auch das Kaufverhalten zu beeinflussen. In digitalen Welten und Spielen könnten künstliche Charaktere sich natürlicher ausdrücken, sich an frühere Begegnungen erinnern und sogar emotionale Beziehungen zu den Spielerinnen und Spielern aufbauen. Intelligente Medientechnologie wird also zunehmend selbst neue Medieninhalte produzieren.

## Recht & Steuern

In Anwaltskanzleien und Steuergesellschaften helfen Sprachtechnologien in der Mandantenkommunikation oder in der Navigation auf der Kanzleiwebseite. Durch den Einsatz von Chatbots bei Standardfragen kann man versuchen, menschliche Intermediäre in der Kommunikation zwischen Mandanten und Anwälten auszuschalten, was das Vertrauen erhöhen könnte. Spracherkennung

mit speziell nachtrainierten Modellen ermöglicht es, Dokumente, Urteile oder Rechnungen zu diktieren und automatisch zu erstellen. Dadurch werden Schreibarbeiten reduziert und es kann ein effizienterer Dokumentenworkflow etabliert werden. Über Informationsextraktion und semantische Suche können automatisch Referenzurteile und Vergleichsfälle herausgesucht werden, auf die eine eigene Formulierung gestützt werden kann.

## Verwaltung

Das Onlinezugangsgesetz (OZG) verpflichtet Bund, Länder und Kommunen dazu, ihre Verwaltungsleistungen bis Ende 2022 über Verwaltungsportale auch digital anzubieten. Dies führt kurzfristig zu Veränderungen in der Verwaltung, in deren Zuge große Mengen an Text digitalisiert werden. Mit Informationsextraktionverfahren können zahlreiche Prozesse verschlankt und vor allem beschleunigt werden. Gegebenenfalls können ganze Aktenvorgänge automatisiert geprüft, analysiert und bearbeitet werden. Durch Sprachassistenten und Chatbots kann die Schnittstelle zu Bürgerdiensten freundlicher, intuitiver und barrierefreier gestaltet werden. Der Vorteil: Bürger-Services wären rund um die Uhr erreichbar, auch am Wochenende. Sprachtechnologien könnten in diesem Zusammenhang auch Texte vereinfachen oder übersetzen und vorlesen, wodurch »Amtssprache« automatisch in einfachere Sprache transformiert würde und auch Personen mit geringen Deutschkenntnissen besser erreicht würden.

## Versicherung

Auch in der Versicherungsbranche können Sprachassistenten massive Prozessvereinfachungen mit sich bringen. Textbots könnten durch die Webseite oder durch Datenerfassungsprozesse führen, Voicebots könnten im telefonischen Kundenservice 24/7 Fälle aufnehmen, Informationen bereitstellen, über Produkte beraten oder das Beschwerdemanagement vollständig übernehmen. Auch der komplette Vertragserstellungsprozess könnte durch Sprachassistenten abgedeckt werden, sodass die Anträge lediglich nachstehend menschlich geprüft werden müssten. Die Mühsal der Vertragsausfüllung auf

einem ausgedruckten Formular kann in Zukunft in einer Art Unterhaltung abgewickelt werden, wobei Fehler direkt bemerkt und korrigiert werden können. Der zusätzliche Benefit: eine höhere Kundenzufriedenheit. Wichtig ist aber, dass bei Bedarf ein menschlicher Berater oder eine Beraterin unmittelbar eingeschaltet werden müssen.

## Automobilindustrie

Sprachassistenten sind in der Automobilbranche nicht neu, Mercedes Benz stellte beispielsweise schon 2010 mit LINGUATRONIC eine Sprachsteuerung im Autocockpit vor. Die Funktionen, die durch Sprache gesteuert werden können, nehmen laufend zu. Man kann sich Nachrichten vorlesen lassen, die Sitzheizung anstellen oder die Route aktualisieren. Die Spracherkennung wird immer besser und kann sogar anhand der Stimme erkennen, wer spricht oder wer fährt, um zum Beispiel die jeweilige Temperatur in den entsprechenden Klimazonen der Fahrer-, Beifahrerseite oder der Rückbank anzupassen.

## Gesundheitswesen

Sprachtechnologien haben im Gesundheitssektor das Potenzial, Mitarbeitende zu entlasten, indem die Arbeitsbelastung durch Technologie reduziert wird, Verwaltungsaufgaben abgebaut werden und schlussendlich mehr Zeit für Erholung und patientennahe Tätigkeiten bleiben. Dies ist beispielsweise durch Spracherkennungslösungen denkbar, die von Ärztinnen und Ärzten sowie medizinischen Fachkräften genutzt werden, um Anamnese-gespräche oder Behandlungen zu dokumentieren. Ferner können in diesem Zusammenhang auch mit Hilfe von Klassifizierungslösungen, Kodierungen vergleichbar und Krankheitsmuster bzw. Zusammenhänge sichtbarer gemacht werden. Bei komplexen, langwierigen und neuartigen Operationen, bei denen kontinuierlich Daten aus dem Computer abgefragt werden müssen, kann eine Sprachsteuerung viel Zeit sparen, weil das medizinische Personal nicht mehr zwischen OP-Tisch und Tastatur wechseln muss. Im Kontext des Patientenzimmers der Zukunft besteht das Potenzial, den Patientinnen und

Patienten eine höhere Aufenthaltsqualität zu ermöglichen. Zum Beispiel durch Sprachsteuerung zur Bedienung von Licht, Rollläden, Fernseher oder zur Kontaktaufnahme mit dem Pflegepersonal. Weiterhin können Chatbots von Versorgungseinrichtungen und in Praxen zur Interaktion mit dem Team, Kundinnen und Kunden sowie zu behandelnden Personen genutzt werden, etwa zur Terminplanung, Abfrage von Informationen oder in der assistierten Bearbeitung von Anträgen. Der Datenschutz und die Vertraulichkeit von personen-, behandlungs-, und patientenbezogenen Daten muss in diesem Zusammenhang auch beim Einsatz von Sprachassistenten beachtet werden.

## 5.4 Vertrauenswürdiger Einsatz von Sprachtechnologien

Wie KI-Anwendungen insgesamt sind Sprachtechnologien geeignet, Routinetätigkeiten zu erleichtern oder zu ersetzen und werden dadurch auch Jobprofile verändern. Sprachbasierte Anwendungen bergen aber auch besondere Gefahren. Mit gefälschten Stimmen – und sogar passend dazu generierten Lippenbewegungen – können Personen falsche Worte in den Mund gelegt und so Fakenews verbreitet werden, die man auf Anhieb nicht mehr erkennen kann. Das wird vor politischen Wahlen ein zunehmendes Problem. Mit Social Bots lassen sich Meinungen auch in großem Stil manipulieren. Chatbots können reale Menschen vortäuschen, auf Dating-Plattformen sind bereits Chatbots gefunden worden, die sich als Frauen ausgeben.

Über die Möglichkeiten des Missbrauchs muss die Gesellschaft aufgeklärt werden, Betrugsvorwürfe und Straftaten müssen verfolgt werden. Vielerorts wurden KI-Ethikkommissionen eingesetzt, die sich darüber hinaus mit der Frage beschäftigen, wie Unternehmen KI-Anwendungen so bauen und einsetzen können, dass sie sicher und akzeptabel sind. Normierungsgremien erarbeiten entsprechende Richtlinien und es gibt erste Schulungen für die Entwicklung vertrauenswürdiger und abgesicherter KI.



Abb. 19: Dimensionen für den vertrauenswürdigen Einsatz von Künstlicher Intelligenz & Sprachtechnologien.

Quelle: Eigene Darstellung

Nordrhein-Westfalen treibt mit einem Flagship-Projekt die Entwicklung der wichtigsten Standards voran, und zwar in Abstimmung mit dem Bund und der High Level Expert Group (AI HLEG) der EU. Aus den Standards wurde ein Prüfkatalog als operativer Rahmen für die Zertifizierung von KI abgeleitet. Er orientiert sich am Risiko der KI-Anwendungen. Je höher die möglichen Schäden, umso mehr ist in vorbeugende oder Gegenmaßnahmen zu investieren. Zur Ermittlung der Gefahren und Maßnahmen werden verschiedene Handlungsfelder unterschieden: Ethik und Recht, Transparenz, Fairness und Diskriminierungsfreiheit, Autonomie und Kontrolle, Verlässlichkeit, Sicherheit, Datenschutz.

## Ethik und Recht

Hier geht es um die Frage, ob die KI-Anwendung gesellschaftliche Werte und Gesetze respektiert. Sie muss noch vor dem Start der Entwicklung geklärt werden und sollte alle betroffenen Akteure berücksichtigen. Wie auch bei den übrigen Handlungsfeldern muss im Lauf der Entwicklung und bis in den Betrieb hinein kontrolliert werden, ob die geforderten Maßnahmen tatsächlich umgesetzt wurden. Bei Sprachassistenten und Chatbots könnten zum Beispiel folgende Überlegungen relevant sein:

- Als Maschine kann ein Bot nicht haften. Kommt es durch falsche Beratung oder Täuschung zu einem Haftungsfall, muss geprüft werden, ob ein individuelles Fehlverhalten des Unternehmens vorliegt.
- Smarte Lautsprecher wurden nicht speziell für Kinder entworfen und können nicht gezielt gesperrt werden. Kinder könnten Privates erzählen, Antworten erhalten, die sie nicht verstehen, Musik mit Texten hören, die nichts für ihre Ohren sind, oder Waren bestellen, obwohl sie noch nicht geschäftsfähig sind.

## Autonomie und Kontrolle

Hier ist zu fragen, ob die Aufgabenverteilung zwischen KI-Anwendung und Mensch klar ist und der Mensch die KI-Anwendung selbstbestimmt und effektiv nutzen kann. Dies kann man mit Methoden des Usability-Engineering testen. Hier sind einige Überlegungen zu Sprachassistenten:

- Smarte Lautsprecher sollten nicht versehentlich aktiviert werden können.
- Ein Chatbot wird den Kunden nicht immer verstehen. Aus der Beratung durch einen Chatbot

sollte man deshalb jederzeit aussteigen und möglichst nahtlos zu einem menschlichen Berater wechseln können.

- Sprachassistenten sollten sich respektvoll ausdrücken. Umgekehrt muss der Sprachassistent mit Beleidigungen umgehen können. Kinder könnten sich den Befehlston im Umgang mit dem smarten Lautsprecher auch bei Menschen angewöhnen. Alexa lobt Kinder deshalb inzwischen für einen höflichen Umgangston.
- Chatbots wie Xiaoice können auf das Entstehen einer persönlichen Beziehung zu ihren Nutzern angelegt sein. Allerdings muss man darauf achten, dass keine Abhängigkeiten entstehen.

## Fairness und Diskriminierungsfreiheit

Behandelt die KI-Anwendung alle Betroffenen fair und werden Personen, die zu einer schützenswerten Gruppe gehören, nicht diskriminiert? Ein mit Daten trainiertes Modell kann grundsätzlich nicht besser sein als diese Daten. Wenn die Daten zu umfangreich sind, um sich bereinigen zu lassen, muss man unfaire oder diskriminierende Ergebnisse im Nachhinein erkennen und unterdrücken oder korrigieren. Das trifft besonders bei großen Sprachmodellen zu, deren Trainingsbasis praktisch das ganze Web ist, oder bei Sprachassistenten, die im Betrieb weiterlernen:

- Chatbots und auch Sprachübersetzer können Geschlechterstereotype aus ihren Trainingsdaten einlernen. Ein Beispiel aus dem Englischen: »The doctor« wird als der Arzt, also als männlich aus der geschlechtsneutralen Form übersetzt, und aus »the nurse« wird die Pflegerin, also in der weiblichen Form übersetzt.
- Spätestens seit Microsofts Chatbot Tay weiß man, dass tiefe Sprachmodelle diskriminierende oder rassistische Äußerungen lernen können.

## Transparenz

Sind Funktionsweise und Entscheidungen der KI-Anwendung nachvollziehbar? Der Mensch muss immer wissen, dass sein Gegenüber eine Software ist und er muss wissen, wozu sie dient, was sie kann und was nicht. Antworten sollten begründet werden können. Einige Transparenzprobleme kann man mit Usability-Experimenten aufdecken und durch die Gestaltung der Interaktion reduzieren. Erklärbarkeit von Ergebnissen ist bei tiefen neuronalen Netzen und damit auch bei tiefen Sprachmodellen ein inhärentes Problem. Deshalb ist der im ersten Kapitel erwähnte Ansatz von Forschern bei Google interessant, die tiefe Sprachmodelle so trainieren möchten, dass sie ihre Äußerungen begründen können.

- Ein Sprachassistent oder Frage-Antwort-System sollte seine Antwort auf Nachfrage belegen können. Geht das nicht mit einem tiefen Sprachmodell, so sollte man einen Wissensgraph verwenden bzw. den Dialogmanager programmieren.
- Kunden sollten immer wissen, wann sie mit einem Bot sprechen und wann mit einem Menschen.
- Man sollte immer wissen, wann ein Mikrofon aufnimmt. Das gilt für smarte Lautsprecher wie für Smartphones, sprechendes Spielzeug oder für Chatbots in der Kundeninteraktion via Text.
- Durch eine gefälschte Stimme kann eine falsche Identität vorgespielt werden. Deep Fakes müssen deshalb kenntlich gemacht werden.

## Verlässlichkeit

Funktioniert die KI-Anwendung zuverlässig und ist sie robust? Ein mit Daten trainiertes Modell sollte möglichst richtige Antworten geben, bzw. anzeigen, wie unsicher seine Antwort ist. Robust bedeutet, dass kleine Änderungen in der Eingabe nur kleine Änderungen in der Ausgabe haben. Und schließlich sollte eine KI-Anwendung prüfen, ob ihre Eingabe überhaupt zulässig ist. Die Unsicherheit einer Antwort kann man technisch abschätzen und Robustheit kann

man statistisch messen. Aber ob eine Eingabe zulässig ist, muss unter Umständen in der Software geprüft werden, in die das KI-Modul eingebaut ist.

- Sprachassistenten und Sprachschnittstellen sollten Dialekte und Sozialekte verstehen. Sie sollten auch mit Personen kommunizieren können, die eine andere Muttersprache haben.
- Sprachanwendungen sollten den Sprecher oder die Sprecherin identifizieren, Gäste und Kinder erkennen, damit sie passend reagieren.
- Auf inhaltlicher Ebene sollten Chatbots Zweifelsfälle erkennen und nachfragen.
- Frage-Antwort-Systeme und Sprachgeneratoren sollte merken, wann irreale Fragen gestellt werden, die man nicht beantworten kann, wie nach dem Sieger eines Spiels, das noch gar nicht stattgefunden hat.

## Sicherheit

Ist die KI sicher gegenüber Angriffen, Unfällen und Fehlern? Diese Gefährdungen können sich als Funktionsausfall oder starke Funktionsänderung der KI-Komponente, sowie durch unautorisierten Informationsabfluss äußern. Wie schon bei der Erkennung unzulässiger Eingaben wären hier vor allem Maßnahmen außerhalb des KI-Moduls zu treffen.

- Mikrofone können eventuell durch Fenster und Türen aktiviert werden. Man hat sogar nachgewiesen, dass sie Lichtsignale als Audiosignale interpretieren und sich durch Laserstrahlen kapern lassen. Abgesehen davon, dass der Sprachassistent durch einen Angriff gesteuert werden kann, kann man auch die Lautsprecher dazu bringen, laute oder hochfrequente und schädliche Töne zu produzieren.
- Sprach-Apps auf dem Smartphone und Skills externer Anbieter können betrügerischen Zwecken wie Phishing dienen. Sie können den Stopp-Befehl uminterpretieren, im Hintergrund weiterlaufen und Unterhaltungen übertragen.

Möglicherweise beschränken sich die Zulassungstests mancher Anbieter nur auf die erste Version, sodass spätere Versionen mit Schadfunktionen unentdeckt bleiben können.

- Man weiß, dass tiefe Sprachmodelle personenbeziehbare Daten aus ihrem Trainingsmaterial enthalten können. Solche Daten kann man einem Modell durch geschickte Anfragen eventuell entlocken.

## Datenschutz

Schützt die KI die Privatsphäre und sonstige sensible Informationen? KI-Anwendungen verarbeiten oftmals sensible Informationen, wie Geschäftsgeheimnisse, personenbezogene oder persönliche Daten. Daher ist sicherzustellen, dass die einschlägigen datenschutzrechtlichen Bestimmungen wie die Datenschutz-Grundverordnung (DSGVO) und das Bundesdatenschutzgesetz (BDSG) eingehalten werden.

- Durch Pannen sind Sprachaufzeichnungen mehrfach an ganz unbeteiligte Personen gelangt.
- In Ausgabe 11/2020 der Stiftung Warentest wurden aktuelle smarte Lautsprecher lediglich mit befriedigend oder ausreichend bewertet. Der Grund ist die mangelhafte Durchsetzung des Datenschutzes.

Insbesondere müssen personenbeziehbare Daten, die man aus Texten extrahiert oder die einem Chatbot mitgeteilt werden, gesetzeskonform verarbeitet und gespeichert werden. Das ist umso wichtiger bei smarten Lautsprechern, die im Laufe der Zeit so viele Daten ansammeln können, dass man nicht nur einzelne Profile erstellen, sondern ganze soziale Strukturen im Haushalt herauslesen kann.

## Leseempfehlungen

### **Beyond Touch – Voice Commerce 2030: Wie Voice-assisted Interfaces den Handel in Europa revolutionieren werden**

Monitor Deloitte (Hrsg.) (2018)

»Sind Voice-assisted Interfaces die nächste Revolution nach Touch, die den Handel grundlegend verändern wird? Dieser Frage sind die Vertreter führender Handelsunternehmen aus Versand, Supermarkt und Discount, Google und wir bei Monitor Deloitte im Rahmen dieser Studie gemeinsam nachgegangen.«

Verfügbar unter: [https://www.thinkwithgoogle.com/\\_qs/documents/8031/Beyond\\_Touch\\_\\_Voice\\_Commerce\\_2030.pdf](https://www.thinkwithgoogle.com/_qs/documents/8031/Beyond_Touch__Voice_Commerce_2030.pdf)

### **Studien & Statistiken zu Sprachassistenten – Nutzung, Marktanteile, Marktführer, Aussichten**

Thomas Hörner (2020)

»Wir können hier nicht garantieren, immer die aktuellsten (und auch nicht alle wirklich existierenden) Studien und Marktzahlen zu Alexa, Google Assistant und Co. haben. Aber wir wollen versuchen, Ihnen im Folgenden eine erste hilfreiche Übersicht zu bieten. ... wir nehmen hier bevorzugt Studien auf, die entweder repräsentativ sind oder zumindest hohe Ansprüche an Studiendesign und -durchführung einhalten bzw. die Grundlagen empirischer Forschung einhalten.«

Verfügbar unter: <https://sprachassistenten.marketing/allgemein/studien-zu-sprachassistenten-nutzung-marktanteile-marktfuehrer-aussichten/>

### **Vertrauenswürdiger Einsatz von Künstlicher Intelligenz**

Dr. Armin B. Cremers, Dr. Stefan Wrobel, et al., Fraunhofer IAIS (Hrsg.) (2019)

»Mit der von uns ins Leben gerufenen Zertifizierung von Künstlicher Intelligenz wollen wir aus Nordrhein-Westfalen heraus die Qualitätsmarke >KI Made in Germany< weiter etablieren, indem sie zuverlässige, sichere Technologien erkennbar macht und nachhaltig schützt. Die Zertifizierung fördert den freien Wettbewerb unterschiedlicher Anbieter und leistet einen Beitrag zur Akzeptanz von Künstlicher Intelligenz in der Gesellschaft.«

Verfügbar unter: [https://www.iais.fraunhofer.de/content/dam/iais/KINRW/Whitepaper\\_KI-Zertifizierung.pdf](https://www.iais.fraunhofer.de/content/dam/iais/KINRW/Whitepaper_KI-Zertifizierung.pdf)

### **KI-Prüfkatalog. Leitfaden zur Gestaltung vertrauenswürdiger Künstlicher Intelligenz**

Maximilian Poretschkin et al. (2021) Fraunhofer IAIS

»Der KI-Prüfkatalog ist ein praxistaugliches Dokument, das von unabhängigen Prüforganisationen als Grundlage für zukünftige Produktprüfungen genutzt werden kann. Gleichzeitig stellt er Unternehmen das Handwerkszeug zur Verfügung, mit dem sie bereits im Entwicklungsprozess ihre Systeme selbst evaluieren und verbessern können und sich so auf zukünftige regulatorische Anforderungen vorbereiten können.«

Verfügbar unter: <https://www.ki.nrw/publikationen/>



# 6 Perspektiven

Wie sehr sich die Entwicklung der Sprachtechnologien in den letzten Jahren beschleunigt hat, verdeutlicht die unten stehende Zeitleiste. Wurden Sprachtechnologien zunächst in Forschungseinrichtungen vorangetrieben, so sind seit 2016 große IT-Konzerne führend. Zu den wichtigsten Durchbrüchen und den aktuellen Trends rund um Sprachtechnologien sind schon einige Experten in Interviews zu Wort gekommen. Auch unsere Autoren haben über diese Fragen diskutiert und möchten ihre Einschätzung zum Abschluss der Studie teilen.

## Was waren wichtigsten Durchbrüche in den letzten zehn Jahren?

Der Durchbruch kam durch Deep Learning ab 2014. Auf allen Stufen der Sprachverarbeitung erreichten tiefe Netze eine Genauigkeit auf menschlichem Niveau, sodass der praktische Einsatz möglich wurde (Fehlerrate wurde reduziert von 30 Prozent auf 8 Prozent).

Ein wichtiger Fortschritt war die Transformer-Architektur, die 2017 von Google erstmals publiziert wurde. Sie ist heute die Basis vieler moderner NLP-Ansätze, welche in Benchmarks sehr gut abschneiden.

Damit eng verbunden ist ein weiterer Durchbruch im NLP-Bereich: die Nutzung vortrainierter Modelle. Große vortrainierte Modelle wie BERT und GPT-3 nutzen oft die Transformer-Architektur und bieten den Vorteil, dass sie für die jeweiligen NLP-Aufgaben unter Nutzung von, relativ gesehen, wenig Trainingsdaten angepasst werden können. Damit wurde die in der KI-Forschung schon länger untersuchte Idee des sogenannten Transferlernens – dem Anwenden von gelerntem Wissen auf einer neuen Domäne bzw. einer neuen Aufgabe – sehr populär im NLP-Bereich.

Weitere wichtige Kerntechnologien im NLP-Bereich sind Wissensgraphen und Graph Neural Networks.

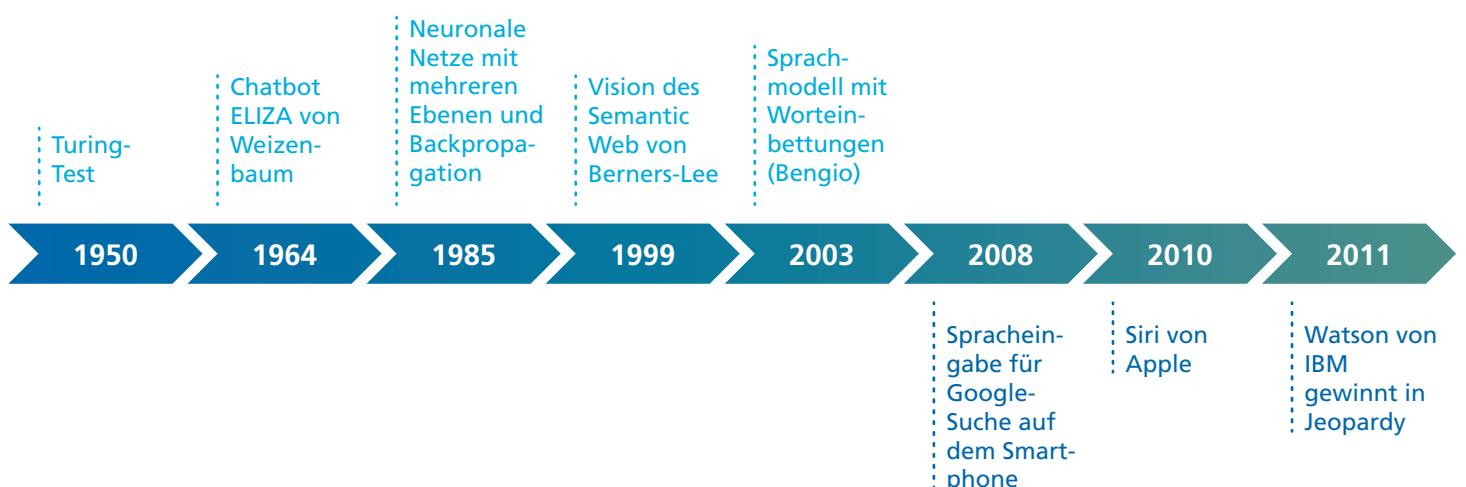


Abb. 21: Zeitleiste der wichtigsten Entwicklungssprünge aus Forschung (oben) und Unternehmen (unten).  
Quelle: Eigene Darstellung

Deren Basistechnologien wurden vor mehr als fünf Jahren entwickelt, aber haben in der letzten Zeit enorm an Bedeutung gewonnen. Sie erlauben es, NLP-Aufgaben auf Graphstrukturen auszuführen und somit menschliches Wissen besser mit KI-Algorithmen zu verknüpfen. Das wird speziell bei der wirtschaftlichen Wertschöpfung hohe Relevanz haben, da es weiterhin wichtig sein wird, das Wissen der Fachleute in den jeweiligen Bereichen mit den Vorteilen moderner KI-Systeme zu kombinieren.

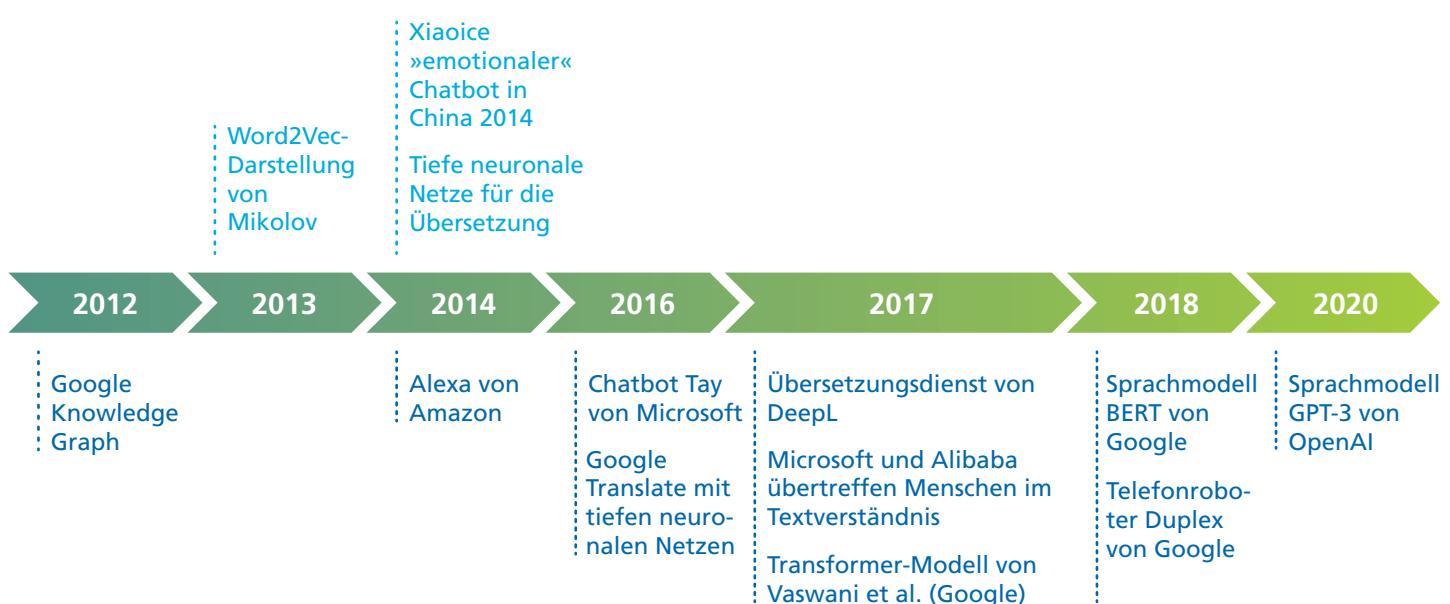
Das Potenzial von Einbettungen ist noch lange nicht erschöpft. Die von BERT erzeugten Einbettungen von Wortfragmenten haben das Textverständnis entscheidend verbessert und sind aus modernen Sprachmodellen nicht mehr wegzudenken. In den Graph Neural Networks werden Tripel aus Wissensgraphen eingebettet, um hybride Frage-Antwort-Systeme zu bauen. Multilinguale Einbettungen ermöglichen Übersetzungen zwischen beliebigen Sprachen, die auch ohne große Trainingsmengen leicht auf neue Sprachen zu erweitern sind. Zunehmend werden Einbettungen auch auf Text und Bilder gleichzeitig angewendet. Das führt zu künstlich intelligenten Systemen, die sowohl hören

und lesen als auch sehen können. Erste Einsatzmöglichkeiten sind multimediale Anwendungen, wie die textuelle Beschreibung von Bildern und Videos oder sogar die Erzeugung von Bildern aus Textbeschreibungen.



Abb. 20: Künstlich generierte Bilder, die DALL-E, als Reaktion auf »Ein Sessel in Form einer Avocado« produziert hat. DALL-E nutzt die gleiche Architektur wie GPT-3, wird aber mit Bildern und Bildunterschriften trainiert.  
Quelle: Chen et al. 2015

Dagegen steckt die Generierung von Videos aus textuellen Beschreibungen noch in den Kinderschuhen. Forscher haben die Sprachprozesskette schon so erweitert, dass sie mit einer Videoaufnahme von einem Sprecher beginnen, die Sprache erkennen, den Text übersetzen, in gesprochene Sprache umwandeln und anschließend zum Ton



passend die Lippenbewegungen in dem Video verändern. Man erhält dann beispielsweise ein Video, in dem man Elon Musk fließend Hindi sprechen sieht.

Die automatische Beschreibung von Bildern ist eine anspruchsvolle Aufgabe. BERT-Modelle, die auf großen Mengen von Teilbildern mit zugehörigen Objektnamen vortrainiert werden und anschließend ein Feintuning auf Bildern mit Bildunterschriften durchlaufen, übertreffen inzwischen die menschliche Genauigkeit.

### Wo sehen Sie die Schwächen von Deep Learning und tiefen Sprachmodellen?

Für tiefe Sprachmodelle benötigt man umfangreiche Trainingsdaten. Wenn man sie selbst nicht hat, muss man vortrainierte Modelle nehmen und kann sie mit deutlich weniger Daten auf die Zielaufgabe feintonen. Wenn man den vortrainierten Modellen nicht vertraut und aus Zeit- oder Kostengründen keine Trainingsdaten für die konkrete Aufgabe bereitstellen kann oder möchte, dann fallen Deep-Learning-Technologien aus. In diesem Zusammenhang gilt allerdings das Motto: »Nicht mit Kanonen auf Spatzen schießen«. Das bedeutet, tiefe Sprachmodelle sollte man nicht bei leichten Aufgaben einsetzen, die man mit einfachen Verfahren bereits sehr gut lösen kann.

Für eine wirtschaftliche Verwertung ist die Kernfrage, wie man die Qualität der Ergebnisse sicherstellen kann, wenn die Menge der Trainingsdaten unüberschaubar groß ist. Setzt man tiefe Sprachmodelle als Sprachassistenten ein, dann sind sie noch weit weg davon, perfekt zu sein. Sie werden in vielen Fällen Antworten geben, die der gesunde Menschenverstand ausschließt. Damit sie noch sinnvoller eingesetzt werden können, müsste man den Modellen mehr Wissen über die Welt beibringen. Die Integration von Wissensgraphen ist ein vielversprechender Weg.

### Worin liegen die größten Potenziale in der Praxis und was sind die aktuellen Herausforderungen?

Sprachtechnologien bringen verschiedene Vorteile: Zeitersparnis durch Beantwortung sich wiederholender oder monotoner Fragen und Fälle. Das erlaubt Experten, sich auf interessante Fälle zu fokussieren. Man kann Computer nutzen und hat dabei die Hände frei. Das ist wichtig im Labor, im OP, im Auto, in Lager- und Fabrikhallen. Die größte Barriere in der Nutzung von Computern waren schon immer die Bedienschnittstellen. Bei jeder neuen Version musste man sich umstellen. Sprache bzw. Text, wenn die Umgebung keine Sprache zulässt, ist die natürlichste Art zu kommunizieren. Statt vom Menschen zu verlangen zu wissen, wie man bestimmte Systeme bedient, kommen Sprachschnittstellen dem Menschen ein großes Stück entgegen.

Der letztgenannte Vorteil ist natürlich auch die größte Herausforderung: eine möglichst intuitive und natürliche Kommunikation mit dem Sprachassistenten. Und die Ansprüche werden steigen. Sprachassistenten werden immer kompliziertere Zusammenhänge verstehen müssen und die Brücke zu immer komplexeren Anwendungssystemen schlagen müssen. Es ist klar, dass man sich nach der Größe der Zielgruppe richtet. Darum findet man die besten Sprachassistenten heutzutage im B2C-Bereich: Smartphones, Wohnungen und Autos. In Unternehmen ist es wichtig, dass Fachpersonal nahtlos einzubinden und gezielt zu unterstützen. Nur so kann der Einsatzbereich einer Anwendung vergrößert und damit ein hoher hohen Mehrwert geboten werden.

## Welche ersten Schritte sollten Unternehmen gehen, um konkret Sprachtechnologien zu etablieren?

Wie gelingt der Einstieg? In wenigen Jahren werden Sprachassistenten aus der Kundenkommunikation, also Marketing und Service, nicht mehr wegzudenken sein. Aber für den Einstieg, um mit der Technologie vertraut zu werden, macht die Optimierung von internen Prozessen vielleicht mehr Sinn. Um ein Beispiel zu nennen: Agenten und Vertreterinnen, die im Auto zu Kunden unterwegs sind, brauchen bestimmte Informationen, um sich auf den nächsten Besuch vorzubereiten. Die kann ihnen ein Sprachassistent im Auto liefern und zurechtlegen. Intern, in so einer Anwendung, muss ein Sprachassistent nicht gleich perfekt sein. Er muss nur etwas Zeit einsparen. Wenn er sich hier aber bewährt, könnte man als nächstes Sprachassistenten für den Kundenkontakt entwickeln. Oft findet man aber auch weitere Stellen zur Optimierung interner Prozesse, zum Beispiel im Umgang mit Texten und Dokumenten. Moderne Anwendungen setzen tiefe Sprachmodelle ein. Dazu benötigt man Daten. Deswegen sind die ersten Schritte: Anwendungsfälle identifizieren

- Datenquellen identifizieren, auf die das System zugreifen soll
- Prüfen, welche Trainingsdaten es gibt bzw. wie man sie bekommen kann
- Für alle weitere Schritte sollte ein Lösungsanbieter zu Rate gezogen werden

## Was bedeuten Sprachmodelle wie GPT-3 für Europa?

Große Sprachmodelle sind von sehr hoher Bedeutung. Sie sind der aktuelle Stand der Technik bei einer Reihe von KI-Aufgaben und liefern eine teilweise beeindruckende Performance. Das wirft die Frage nach der Kontrolle über solche Modelle auf. GPT-3 hat zum Beispiel 175 Milliarden Parameter und das Training eines solchen Modells kostet viele Millionen Euro. Sehr große Unternehmen wie Microsoft können Exklusivlizenzen für GPT-3 und andere große Sprachmodelle kaufen. Für diese Unternehmen werden sich die Ausgaben vermutlich um ein Vielfaches bezahlt machen, weil sie sie in unzähligen Anwendungsfällen einsetzen können. Andere Unternehmen, die ein konkretes Problem lösen möchten, werden kaum das Personal, Geld und die Rechenressourcen aufbringen können, um so ein Modell selbst zu trainieren. Wo stehen in diesem Szenario europäische Unternehmen? Sollen sie abhängig sein von den wenigen Großunternehmen, die große Sprachmodelle trainieren können? Wer sichert die Qualität der Modelle? Sollen Forscher außerhalb bestimmter Großunternehmen in der Lage sein, an solchen Fragestellungen arbeiten zu können? Angesichts der Tatsache, dass KI inzwischen Teil unseres Alltags ist und in fast allen Unternehmensbereichen eine Rolle spielt, müssten Deutschland und Europa investieren, um konkurrenzfähig zu bleiben – das ist keine Aufgabe für ein einzelnes Unternehmen und es sollte auch keine Option für Deutschland und Europa sein, die Souveränität über KI-Anwendungen abzugeben.

## Leseempfehlungen

### SYNTHETISCHE MEDIEN

WDR Innovation Hub (2021)

»Doch was sind Synthetische Medien eigentlich? Wie wirken sie? Und wohin entwickelt sich die Zukunft mit Synthetischen Medien? Das beantworten wir – der WDR Innovation Hub – in unserem Zukunftsreport >Synthetische Medien<. Gemeinsam mit Zukunftsforchern haben wir vier Szenarien erarbeitet, wie die Welt in zwei bis fünf Jahren mit Synthetischen Medien aussehen kann. Warum vier? Weil es nicht die EINE Zukunft gibt. Zukunft kann immer verschiedene Ausprägungen haben. Welche genau, das wissen wir nicht. Aber wir wollen auf alle Möglichkeiten vorbereitet sein.«

Verfügbar unter [https://zukunft.wdr.de/assets/pdf/WDR-Zukunftsreport\\_SynthetischeMedien.pdf](https://zukunft.wdr.de/assets/pdf/WDR-Zukunftsreport_SynthetischeMedien.pdf)

### Natural Language Understanding: Instructions for (Present and Future) Use

Roberto Navigli, ICJAI (2018)

»An intelligent system should be able to explain the process followed for understanding text, especially if a decision has to be taken. ...A fundamental question is whether we need explicit lexical semantics at all in order to perform NLU: cannot we just understand text by comparing and transforming its latent representations to those of other texts, without making the effort of identifying and associating explicit semantics? Potentially yes, but would it be better? Posterity will judge.«

Eine Überlegung zu den Potenzialen von Einbettungen und eine Gegenüberstellung mit symbolischen Darstellungen wie in Wissensgraphen.

Verfügbar unter: <https://www.ijcai.org/Proceedings/2018/0812.pdf>



## 7 Primärquellen

Adobe & Advanis 2019 (a,b): *State of Voice Technology for Brands* [verfügbar unter: [https://www.slideshare.net/slideshow/embed\\_code/key/v76vHDbO0WLw](https://www.slideshare.net/slideshow/embed_code/key/v76vHDbO0WLw), via <https://blog.adobe.com/en/2019/05/14/91-of-brands-are-investing-in-voice-how-to-make-it-work.html#gs.1929z8>, zuletzt aufgerufen: 20.05.2021].

Diesel, A. & Psyma (Hrsg.) (2020): *Sprachassistenten – Zukunft der Kundeninteraktion auch im Finance Bereich?* [verfügbar unter: <https://newsroom.psyma.com/de/themen/finanzmarktforschung/sprachassistenten-zukunft-der-kundeninteraktion-auch-im-finance-bereich/>, zuletzt aufgerufen: 20.05.2021].

GfK SE & ZVEI (Hrsg.) (2017): *Jeder fünfte Befragte nutzt bereits Sprachsteuerung* [verfügbar unter: [https://www.zvei.org/fileadmin/user\\_upload/Presse\\_und\\_Medien/Pressebereich/2018-023\\_Jeder\\_fuenfte\\_nutzt\\_Smart\\_Home>Anwendungen/Pr\\_2018-023\\_Grafik\\_2\\_Sprachsteuerung\\_01.pdf](https://www.zvei.org/fileadmin/user_upload/Presse_und_Medien/Pressebereich/2018-023_Jeder_fuenfte_nutzt_Smart_Home>Anwendungen/Pr_2018-023_Grafik_2_Sprachsteuerung_01.pdf), zuletzt aufgerufen: 19.05.2021].

KPMG (Hrsg.) (2019): *How may AI assist you?* [verfügbar unter: <https://hub.kpmg.de/hubfs/LandingPages-PDF/ki-studie-how-may-ai-assist-you.pdf>, zuletzt aufgerufen: 19.05.2021].

Mordor Intelligence (Hrsg.) (2020): *Voice Recognition Market - Growth, Trends, COVID-19 Impact, and Forecasts (2021 - 2026)* [verfügbar unter: <https://www.mordorintelligence.com/industry-reports/voice-recognition-market>, zuletzt aufgerufen: 20.05.2021].

Roller, S. et al. (2020): *Recipes for building an open-domain chatbot* [verfügbar unter: <https://www.aclweb.org/anthology/2021.eacl-main.24.pdf>, zuletzt aufgerufen: 19.05.2021].

Speechmatics (Hrsg.) (2021): *Trends and Predictions for Voice Technology in 2021* [verfügbar unter: <https://www.speechmatics.com/wp-content/uploads/2021/01/Speechmatics-Report-Trends-Predictions-Voice-Technology-2021.pdf>, zuletzt aufgerufen: 19.05.2021].

Vailshery, L. S. (2021): *Number of digital voice assistants in use worldwide from 2019 to 2024* [verfügbar unter: <https://www.statista.com/statistics/973815/worldwide-digital-voice-assistant-in-use/>, zuletzt aufgerufen: 19.05.2021].



# **8 Impressum**

## **Herausgeber**

Kompetenzplattform KI.NRW  
Geschäftsführung Dr. Christian Temath  
c/o Fraunhofer-Institut für Intelligente  
Analyse- und Informationssysteme IAIS  
Schloss Birlinghoven  
53757 Sankt Augustin  
<http://www.iais.fraunhofer.de>

## **Kontakt**

Florian T. Jovy-Klein  
KI Manager KI.NRW  
Telefon 02241 14-2186  
[florian.jovy-klein@iais.fraunhofer.de](mailto:florian.jovy-klein@iais.fraunhofer.de)

## **Coverbild**

Maya Schwarzer, Fraunhofer IAIS

## **Layout und Satz**

Maya Schwarzer, Fraunhofer IAIS



