Woran forschen Sie,

Christoph Purschke?

Ich bin derzeit in der schönen Lage, etwas Neues aufbauen zu können. Nachdem ich mich in den letzten Jahren an der Uni Luxemburg vor allem mit Aspekten der hiesigen Mehrsprachigkeit befasst habe, besonders mit der Engführung von Sprachen(politik), Ideologie und Diskurs, aber auch mit partizipativer Forschung *(Citizen Science)*, bin ich momentan dabei, einen neuen Arbeitsschwerpunkt zu entwickeln. Seit Oktober arbeite ich als Professor für Computerlinguistik am Institut für Luxemburgistik, und im Rahmen dieser Stelle eröffnen sich mir eine Reihe spannender Forschungsfährten.

Darunter sind zunächst ganz praktische Vorhaben; so planen wir den Aufbau einer Forschungsgruppe für Kultur und Digitalität. Dort soll zum einen geisteswissenschaftliche Forschung mit sprachtechnologischen Methoden gefördert und sichtbar gemacht werden. Dazu gehört auch deren Einbindung in die Studienprogramme der Fakultät. Darüber hinaus geht es mir aber auch um eine kritische Begleitung des digitalen Wandels aus der Sicht der Geisteswissenschaften, etwa im Hinblick auf die Rolle von sogenannten „künstlichen Intelligenzen“ als automatisierten sozialen Akteuren – mit teilweise reeller Entscheidungsgewalt. Hier können die Geisteswissenschaften Grundlegendes zum Verständnis von Kultur unter digitalen Vorzeichen beitragen, ebenso wie zu einem nachhaltigen Einsatz von Technologie im Dienste der Gesellschaft.

Ein Forschungsfeld, das mich in den nächsten Jahren viel beschäftigen wird, ist die Arbeit mit Methoden und Modellen der automatischen Sprachverarbeitung *(Natural Language Processing)*. Wir alle nutzen solche Modelle täglich, egal ob wir einen Begriff googeln, Siri eine Notiz diktieren oder uns einen Satz ins Englische übersetzen lassen. Hinter diesen Anwendungen stecken komplexe Sprachmodelle, die mit Hilfe großer Datenmengen trainiert werden. So gut diese Anwendungen mittlerweile auch funktionieren, gibt es dennoch große Bereiche der Alltagspraxis, für die sie „blind“ sind und die sie kaum bewältigen können. Man kann sich das sehr gut vor Augen führen, wenn man auf *Google Translate* einen halbwegs komplexen Satz vom Französischen ins Luxemburgische übersetzt. Das geht, ergibt dann aber den sprachlichen Unsinn, wie man ihn aus Spam-E-Mails kennt.

Allein dieses Beispiel zeigt, wie limitiert Sprachtechnologie derzeit eigentlich noch ist – und weshalb Expert.innen diese kaum je als „intelligent“ bezeichnen. Und dabei handelt es sich hier noch um einen unproblematischen Fall. Es hat sich nämlich gezeigt, dass solche Modelle in hohem Maße kulturelle Stereotype sowie rassistische und sexistische Vorurteile reproduzieren. Um das zu testen, muss man nur versuchen, auf *Google Translate* den Satz „Die Ärztin gibt der Patientin eine Spritze“ ins Französische übersetzen zu lassen. Grammatisch ist das Ergebnis tadellos, allerdings werden aus weiblichen Referenzen plötzlich männliche. Für diese „Fehlleistung“ der Modelle gibt es eine Reihe von Gründen; der wichtigste bezieht sich auf die den Modellen zugrundeliegenden Daten. Häufig werden Sprachtechnologien vor allem auf Basis englischer Daten trainiert, hinzu kommen soziale und demografische Einschränkungen (männlich, „weiß“, „westlich“, bessergestellt). Auch gilt, dass die Demografie der Branche einem bestimmten Profil folgt (nämlich demselben). Welche Auswirkungen dieser „Bias“ auf große Teile der Menschheit hat, zeigt das Buch *Invisible Women* von Caroline Criado Perez eindrücklich.

Sprachtechnologien erkennen weder Ironie noch sprachspezifische Kodierungen von Gender zuverlässig.

Mit meiner Arbeit möchte ich dazu beizutragen, sprachtechnologische Anwendungen weiterzuentwickeln, um ihnen z. B. die Vorurteile abzutrainieren. Dazu ist es nötig, ihre Datenbasis zu diversifizieren, also ein breiteres Spektrum an Sprecher.innen einzubeziehen. Darüber hinaus wollen wir Sprachmodelle gezielt mit kulturellen Kontextdaten konfrontieren. Sprache steht ja nicht isoliert da, sondern ist in komplexe, soziokulturell geprägte Routinen eingebettet, die entscheidend für Sinn und Verständnis von Äußerungen sind. Und genau an diesen Kontexten scheitern Sprachtechnologien bislang: Sie erkennen weder Ironie noch eben sprachspezifische Kodierungen von Gender zuverlässig. Mein Hintergrund als Soziolinguist kommt mir für diese Aufgabe sehr zupass.

Überhaupt existieren die meisten Anwendungen – und ein Großteil der Forschung – bislang vor allem für „große“ Sprachen wie Chinesisch, Spanisch oder Deutsch. Je kleiner die Sprachgemeinschaft, und je weniger Ressourcen es also gibt, die für das Training benutzt werden können, um so schlechter fallen meist die Ergebnisse aus. Teil meiner Arbeit soll deshalb auch sein, die Entwicklung von Ressourcen und Anwendungen für kleinere Sprachen wie Luxemburgisch voranzubringen. So arbeite ich an einem Tool für die automatische Korrektur luxemburgischer Texte, und auch die Entwicklung einer Spracherkennung für Luxemburgisch steht auf dem Programm des Instituts. Es könnten also spannende Jahre werden. ◆

Christoph Purschke war schon einiges, Dialektologe zum Beispiel oder Soziolinguist. Dass er jetzt vor allem mit Computern arbeitet, hat auch mit seiner Neigung zum Basteln als Lebensprinzip zu tun, heißt aber nicht, dass er nicht auch zukünftig in den Diskurs hineinstänkern wird, etwa wenn es um die Mehrsprachigkeit im Land und ihre politische Ideologisierung geht. In seiner Freizeit backt er sehr leckeres Brot.