PICTURE ALLIANCE / DPA/ BERND VON JUTRCZENKA

An der **SPRACHE** soll man sie erkennen

Weil ihre Herkunft oft nicht leicht zu klären ist, will das Bundesamt für Migration und Flüchtlinge die Dialekte von Asylsuchenden mit einer Computersoftware analysieren

„Schibboleth“ ist hebräisch und bedeutet „Ähre“. Das Wort ist mehr als 3000 Jahre alt. Schon in der Bibel kommt es vor. Dort entscheidet die Art, wie man es ausspricht, über Leben und Tod. Die Gileaditer verlangten von den Flüchtlingen aus dem Gebiet der verfeindeten Ephraimiter, das Wort auszusprechen. Die Ephraimiter sprachen das *Sch* nämlich wie ein *S* aus: Sibboleth. Das war für sie das Todesurteil.

VON PHILIPP HUMMEL

Auch heute spielt es noch eine entscheidende Rolle für Flüchtlinge, wie sie sprechen, wenn auch die Folgen weit weniger drastisch sind. Bei vielen Asylbewerbern lässt sich die Herkunft nicht leicht überprüfen. Beim Bundesamt für Migration und Flüchtlinge (BAMF) haben etwa 60 Prozent der seit 2015 angekommenen Asylbewerber keine gültigen Ausweispapiere vorgelegt. Um die Herkunft dennoch zu klären, setzt das BAMF seit 1998 auch Sprachanalysen ein.

Am Dialekt sollen Gutachter erkennen, wo ein Asylbewerber aufgewachsen ist. Doch nun plant das BAMF mit einem neuen Verfahren, die Sprechweise von Flüchtlingen auch digital zu untersuchen. „Es geht darum, automatisiert einen Sprecher einem Dialekt zuzuordnen“, erklärt Julian Detzel, Referent im Bereich Grundsatzstrategie Digitalisierung und IT-Programmmanagment. Er leitet die „Stimmbiometrie-Machbarkeitsstudie“, so der Titel des Pilotprojekts.

Als technologische Basis für das System soll eine Software zur Sprecherauthentifizierung dienen, die dann so verändert wird, dass der Computer danach Dialekte erkennen kann. Bei der Sprecherauthentifizierung wird geprüft, ob ein bestimmter Sprecher derjenige ist, der er zu sein vorgibt. Banken und Versicherungen verwenden solche Systeme, um ihren Service sicher übers Telefon anzubieten.

Dahinter stecken Algorithmen zur Mustererkennung, die auf „maschinellem Lernen“ basieren: Das System bekommt sogenannte Trainingsdaten eingespielt, beispielsweise eine Reihe von Sprachaufnahmen des immer gleichen Passwortsatzes von verschiedenen Sprechern. Dazu verraten die Entwickler der Software, bei welchen der Aufnahmen es sich um den gesuchten Sprecher handelt und bei welchen um Betrüger. Aus den Daten bildet das System ein Modell abstrakter Parameter. Diese gleicht es dann ab. Die Software lernt dabei aus den Trainingsdaten. Sie passt sich so an, dass einerseits echte Kunden in möglichst wenigen Fällen für einen Betrüger gehalten werden und andererseits möglichst wenige Betrüger fälschlicherweise als Kunden durchgehen. Völlig perfekt funktioniert dieser Filter nicht. Wie eng er ist, geben die Software-Entwickler vor.

Das BAMF will nun nicht die Stimme eines Kunden von der eines Betrügers unterscheiden, sondern prüfen, ob ein Asylbewerber den Dialekt einer bestimmten Region spricht oder nicht. Dazu will man in einem ersten Schritt eine Software entwickeln, die erkennt, ob ein Sprecher levantinisches Arabisch spricht. Dieser Dialekt ist in Teilen Syriens, aber auch in Jordanien, dem Libanon und weiteren Ländern verbreitet. Verläuft der Test erfolgreich, will das BAMF das Projekt auf weitere Dialekte ausweiten. „Die Idee ist, von Asylantragstellern eine separate Sprachprobe aufzunehmen und einer automatischen Dialektanalyse zu unterziehen“, sagt IT-Referent Detzel. Die Dialektsoftware muss keine gerichtsfesten Analysen liefern. Vielmehr soll sie als Assistenzsystem dem Entscheider im Asylverfahren einen weiteren „Indikator“ an die Hand geben. Unterscheiden sich die Angaben des Asylbewerbers von den Ergebnissen der Software, müsste nach wie vor ein menschlicher Gutachter eine Sprachanalyse durchführen.

Das BAMF betritt mit diesem Vorstoß technologisches Neuland. Die Welt hat mit etwa einem Dutzend internationaler Experten für Sprachtechnologien gesprochen. Keinem war ein vergleichbares System bekannt. Das BAMF wäre damit die erste Institution weltweit, die eine massenhafte Dialektanalyse in Asylverfahren einsetzt.

Doch wie zuverlässig kann das funktionieren? In einem Wettbewerb haben Forscher Anfang des Jahres verschiedene Algorithmen daraufhin getestet, wie gut sie vier arabische Dialekte und Standard-Arabisch zuordnen können. Das beste System schaffte eine Treffsicherheit von über 75 Prozent. Mit schlichtem Raten würde man bei fünf Dialekten nur eine Quote von 20 Prozent erzielen. Marcos Zampieri von der Universität Köln hat den Wettbewerb mit organisiert. Er ist davon überzeugt, dass eine Dialektsoftware auch bei Asylverfahren eingesetzt werden könnte: „Kein System ist perfekt, aber die besten funktionieren vernünftig. Das könnte für eine Vorauswahl nützlich sein, oder um die Einschätzung eines Sprachanalysten zu stützen.“

Allerdings handelte es sich bei den Sprachproben um Mitschnitte aus dem Programm des Fernsehsenders al-Dschasira. Bei Sprachaufnahmen von Asylbewerbern hätte man es mit ganz anderem Datenmaterial zu tun. Georgina Brown von der University of York schreibt in einer Studie, man könne nicht davon ausgehen, dass die Ergebnisse für einen Satz Sprachdaten sich auf einen anderen ohne Weiteres übertragen lassen. In manchen Sprachen ließen sich die Dialekte vielleicht besser unterscheiden. Und bestimmte Systeme könnten sich mit einem Dialekt leichter tun als mit einem anderen.

Eine besondere Schwierigkeit liegt in den Trainingsdaten. Dirk Hovy, ein deutscher Computerlinguist an der Universität von Kopenhagen, erklärt, die Daten müssten möglichst repräsentativ die Gruppe der zu untersuchenden Asylbewerber abbilden, was etwa das Alter oder die echte Herkunft angeht. Sonst würden diejenigen benachteiligt, deren Sprechweise das System nicht gelernt hat. „Einen perfekten Datensatz zu erstellen, ist praktisch unmöglich,“ sagt Hovy. „Schon weil Sprache sich dauernd ändert.“ Wenn man aber über eine breite Datenbasis verfüge, könne man zumindest eine verwendbare Annäherung schaffen. Das sei jedoch sehr aufwendig.

Tomi Kinnunen von der University of Eastern Finland warnt davor, dass die Software die Trainingsdaten einfach auswendig lernen und dann nicht von diesen auf neue Daten abstrahieren könnte. Außerdem müssten die Trainingsdaten idealerweise unter denselben technischen Bedingungen aufgezeichnet werden wie die späteren Sprachproben der Asylbewerber, das betrifft beispielsweise Aufnahmegeräte und Umgebungsakustik. „Sonst besteht die Gefahr, dass das System im Einsatz mehr Fehler macht.“

Wie ein solches System reagiert, wenn jemand versucht, einen Dialekt zu imitieren, dazu gibt es nach dem Stand dieser Recherche bisher noch keine Ergebnisse. Dieses Szenario muss man aber in Asylverfahren durchaus in Erwägung ziehen. Im Moment hat das BAMF vor, seine Trainingsdaten bei externen Anbietern einzukaufen, die Sprachdatenbanken erstellen. Testen will man das System entweder ebenfalls mit diesen Daten, mit Dolmetschern oder anderen Freiwilligen, die Arabisch sprechen. Als Softwarebasis soll ein System der Firma Nuance zum Einsatz kommen, „dem führenden Technologieanbieter im Bereich Stimmbiometrie“.

Die Tests für das Dialekterkennungssystem sollen bald starten, „voraussichtlich innerhalb der nächsten zwei Wochen“, heißt es bei BAMF. Mit einem routinemäßigen Einsatz des Systems sei aber nicht vor 2018 zu rechnen.

http://ttg.uni-saarland.de/vardial2017/