Originalveröffentlichung in: Busch, Anna/Trilcke, Peer (Hrsg.): DHd2023: Open Humanities, Open Culture. 9. Jahrestagung des Verbands Digital Humanities im deutschsprachigen Raum e.V.: Universität Luxemburg & Universität Trier, 13. bis 17. März 2023. Konferenzabstracts. – Genf: Zenodo, 2023. S. 415-417.
DOI: *https://doi.org/10.5281/zenodo.7688632*

Projektvorstellung – Sprachanfragen. Empirisch gestützte Erforschung von Zweifelsfällen

Lang, Christian
lang@ids-mannheim.de
Leibniz-Institut für Deutsche Sprache, Deutschland

Tu, Ngoc Duyen Tanja
tu@ids-mannheim.de
Leibniz-Institut für Deutsche Sprache, Deutschland

Creative Commons - Attribution 4.0 International

Publikationsserver des Leibniz-Instituts für Deutsche Sprache
URN: http://nbn-resolving.de/urn:nbn:de:bsz:mh39-122034

Schneider, Roman
schneider@ids-mannheim.de
Leibniz-Institut für Deutsche Sprache, Deutschland

Volodina, Anna
volodina@ids-mannheim.de
Leibniz-Institut für Deutsche Sprache, Deutschland

Einführung

Das im Januar 2022 gestartete Projekt „Sprachanfragen“ (https://www.ids-mannheim.de/gra/projekte2/sprachanfragen/) verfolgt das Ziel, Sprachanfragedaten – also Daten, die im Rahmen von verschiedenen Sprachberatungsszenarien entstehen, wie beispielsweise (1) – zu erfassen, aufzubereiten und ein wissenschaftsöffentliches Monitorkorpus aus ihnen zu erstellen. Dazukommend wird eine Rechercheschnittstelle entwickelt, mit der die Sprachanfragen systematisch wissenschaftlich analysierbar gemacht werden.

(1) „[Frage:] Heißt es "Dramaform" oder "Dramenform" [...]?

[Antwort:] In allgemeinsprachlichen Wörterbüchern ist diese Zusammensetzung nicht erfasst. Im allgemeinen Schreibgebrauch wird - wie eine Internetrecherche ergab - die Form mit Fugen-en vorgezogen.“

Sprachanfragen bieten einen authentischen Einblick in Probleme und Themen, die Sprecher:innen außerhalb der linguistisch-fachwissenschaftlichen Gemeinschaft beschäftigen. Wie Breindl (2016, 86f.) ausführt, bietet eine systematische Auswertung der Sprachberatungspraxis eine wertvolle Grundlage für die Erforschung einer großen Bandbreite verschiedener Fragestellungen. So können diese Daten u. a. dazu benutzt werden, um (i) Zweifelsfälle zu analysieren, wodurch Normierungslücken aufgedeckt werden können, und um (ii) Sprachwandelphänomene nachzuvollziehen. Ebenfalls können Sprachanfragen herangezogen werden, um (iii) Strategien zu erforschen, wie fachspezifische Inhalte von Nicht-Fachpersonen erfragt werden. Dadurch können bspw. die Zugangswege zu grammatischen und orthographischen Inhalten in einem webbasierten Informationssystem optimiert werden. Eine mögliche Optimierung wäre, Sprachanfragen automatisch in Form eines Chatbots zu beantworten.

Das Poster gibt einen Überblick über das Projekt, zeigt erste Ergebnisse und bietet einen Ausblick auf Überlegungen zur Konzeption eines Chatbots zur automatisierten Beantwortung von Sprachanfragen.

Datengrundlage

Das Monitorkorpus wird zum einen aus ~50.000 Sprachanfragen, die an den Sprachberatungsservice des WAHRIG-Verlags per E-Mail geschickt wurden, aufgebaut. Diese decken einen Zeitraum von 1999 bis 2018 ab. Die zugehörigen Antworten werden ebenfalls in das Korpus aufgenommen. Zum anderen wird das Korpus kontinuierlich mit Sprachanfragen erweitert, die im Leibniz-Institut für Deutsche Sprache eingehen. Um mehr Daten für das Trainieren eines Chatbots zu generieren, werden darüber hinaus Sprachanfragen aus Online-Quellen, wie z.B. gutefrage.net, extrahiert.

In einem ersten Schritt werden die Daten aufwendig vorverarbeitet. Dabei werden sie anonymisiert, um den Datenschutz zu gewährleisten und das Korpus wissenschaftsöffentlich zur Verfügung stellen zu können. Für die Anonymisierung ist die Nutzung eines NamedEntity-Erkenners, wie in anderen Arbeiten geschehen (vgl. u.a. Bleicken et al., 2016; Kleinberg et al., 2017), nicht optimal, da u. a. Namen ebenfalls Teil der Fragestellung sein können (vgl. (2)). Somit müssen automatisierte Lösungswege gefunden werden, um primär tatsächlich personenbezogene Daten zu ersetzen und die anschließende manuelle Nachkorrektur maßgeblich zu erleichtern.

(2) "[...] Der Genitiv des Wortes "Paulus" [...] sollte wie lauten: "Pauli" oder "Paulus'"? [...]"

Darüber hinaus werden die Sprachanfragen nach orthographischen und terminologischen Kriterien strukturiert, indem sie mit grammatischen Termini (z. B. „Dativ“, „Fugen-s“, „Getrenntschreibung“) annotiert werden. Basis dafür ist die terminologische Ressource der Abteilung Grammatik des Leibniz-Instituts für Deutsche Sprache, die sogenannte Wissenschaftliche Terminologie (WT, https://grammis.ids-mannheim.de/terminologie). Diese beinhaltet ~6.000 Termini aus der Domäne Deutsche Grammatik (vgl. u.a. Suchowolec et al., 2019). Berücksichtigt werden Uni- (z.B. „Substantivierung“), Bi- (z.B. „indirekte Rede“) und Trigramme (z.B. „negationsinduzierend additive Konnektoren“). Mit Hilfe eines Pattern Matchings werden vorkommende Termini in den lemmatisierten Sprachanfragen automatisiert detektiert. Über exakte Treffer hinaus werden bei der Annotation von Unigrammen auch Teiltreffer am Anfang oder am Ende eines Lemmas aufgenommen, bspw. „ Genitivbezug", „ Dativform“, „Muss- Komma“. Somit werden auch Ausdrücke erfasst, die einen Terminus als Erst- oder Zweitglied beinhalten, als Ganzes jedoch nicht als Termini in der WT auftreten.

Um zu evaluieren, wie gut die Automatisierung der beiden Vorverarbeitungsschritte funktioniert, wird ein Subkorpus aus 1.000 zufällig extrahierten Sprachanfragen erstellt. Dieses wird manuell anonymisiert sowie terminologisch annotiert und als Goldstandard bei der Auswertung der automatischen Methoden herangezogen.

Ausblick: automatisierte Beantwortung von Sprachanfragen

Eine weiterführende, zukünftige Zielsetzung ist zudem, bei ausreichender Größe des Monitorkorpus, einen Chatbot zur automatischen Beantwortung von Sprachanfragen zu entwickeln. Dafür werden die Sprachanfragen nach den zugeordneten Termini gruppiert und ein Modell je Gruppe trainiert. Als Baseline wird ebenfalls ein regelbasierter Chatbot implementiert. Denkbar wäre auch eine Kombination aus regelbasiertem und trainiertem Chatbot. Das Ziel ist es, mit einem solchen Sys-

tem eine nicht-kommerzielle und offene (im Sinne von Veröffentlichung des Quellcodes) Alternative zu anderen Online-Grammatik- und Rechtschreibhilfetools (z. B. Deepkomma, Duden-Mentor, LanguageTool oder Studi-Kompass) zu schaffen, die durch nahtlose Anknüpfung an die umfassenden sprachwissenschaftlichen Ressourcen des hauseigenen wissenschaftlichen Informationssystem zur deutschen Grammatik grammis (https://grammis.ids-mannheim.de/) umfangreiche Materialien zum weiterführenden Selbststudium auf verschiedenen Komplexitätsstufen bietet.

Im Fokus der automatischen Beantwortung soll also nicht nur die Korrektur, sondern es sollen auch die sprachwissenschaftlichen Hintergründe einer Frage stehen. Zum Beispiel sollen im Fall der folgenden authentischen Sprachanfrage: „Was ist korrekt: „Haushalthilfe" oder „Haushaltshilfe", „Haushaltpflege" oder „Haushaltspflege"?" über die Angabe der korrekten Variante hinaus das zugrundeliegende Phänomen (Fugenelemente) benannt und entsprechende Artikel aus grammis verlinkt werden.

Bibliographie

Bibliographisches Institut GmbH. 2022. *„Duden Mentor."* https://mentor.duden.de/.

Bleicken, Julian, Thomas Hanke, Ute Salden, und Sven Wagner. 2016. "Using a Language Technology Infrastructure for German in order to Anonymize German Sign Language Corpus Data." In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 16)*, 3303-3306. Portorož, Slovenia.

Breindl, Eva. 2016. "Sprachberatung im interaktiven Web". In *Die Kodifizierung der Sprache. Strukturen, Funktionen, Konsequenzen*", herausgegeben von Wolf-Peter Klein und Sven Staffeldt, 85–109. WespA – Würzburger elektronische sprachwissenschaftliche Arbeiten 17. Würzburg.

gutefrage.net GmbH. o.J. *"guteFrage."* https://www.gutefrage.net/

Kleinberg, Bennett, Maximilian Mozes, Yaloe van der Toolen, und Bruno Verschuere. 2017. *"NETANOS - Named entity-based Text Anonymization for Open Science."* Preprint. Open Science Framework. https://doi.org/10.31219/osf.io/w9nhb.

LanguageTooler GmbH. o. J. *"LanguageTool."* https://languagetool.org/de.

Mannheim: Leibniz-Institut für Deutsche Sprache. o. J. *"Grammatisches Informationssystem ‚grammis'."* http://grammis.ids-mannheim.de.

Suchowolec, Karolina, Christian Lang, und Roman Schneider. 2019. "An empirically validated, onomasiologically structured, and linguistically motivated online terminology. Re-designing scientific resources on German grammar." *International Journal on Digital Libraries* 20: 253-268.

Uniprof LLP. 2016-2022. *"Studi-Kompass."* https://studi-kompass.com/generatoren/online-rechtschreibpruefung.

Wefelscheid, Cornelius. o. J. *"DeepKomma."* https://deepkomma.de.