Datensammlung in der Romanistik – Eine Analyse von Normierung und Standardisierung in E-Mails

Laura Bothe, Sybille Große

Romanisches Seminar, Universität Heidelberg

Den E-Mailaustausch gibt es seit nunmehr 50 Jahren (Delfa 2021). Er ist heute ein weltweit anerkanntes Kommunikationsmedium, sowohl in formellen als auch informellen Kommunikationskontexten. Gerade deshalb stellt sich aus linguistischer Sicht die Frage nach der Standardisierung und Normierung von E-Mails (Große 2012). Nach einer Studie von *Statista* wurden im Jahr 2021 ca. 319 Milliarden E-Mails am Tag verschickt (Statista 2023). Tendenz steigend. Trotzdem ist das Interesse der Linguisten an E-Mails in den letzten Jahren zurückgegangen. Das Aufkommen der sozialen Medien (Rentel und Schröder 2018) und der Wechsel vom informellen auf einen formelleren Gebrauch der E-Mail (Souchier u. a. 2019), der sich auf die Kommunikationssituationen und verwandten Versprachlichungsstrategien auswirkt, können hierfür als Erklärung herangezogen werden. Zudem haben sich die Fragen der Datenerhebung und -verarbeitung im Rahmen der Analyse von internetbasierter Kommunikation (IBK) auf der Grundlage von sich in den *Digital Humanities* etablierenden Standards in den letzten Jahren zu einer immer größeren Herausforderung entwickelt (Beißwenger 2017). Gleichzeitig wird die Datenerhebung in sozialen Medien wie *Twitter* oder *Telegram* von den Unternehmen durch die vorgegebenen Privatsphäre-Regelungen und entsprechenden Download-Tools vereinfacht. Ein umfangreiches, zeitlich relevantes, zahlreiche Schreiber:innen umfassendes und aus formellen und informellen E-Mails bestehendes Korpus anzulegen, ist bereits durch den restriktiveren Zugang zu den Daten mit erheblichem Mehraufwand verbunden. Da es keine zentrale Sammelstelle gibt, an der die E-Mails abrufbar sind, ist die Forschung in diesem Bereich im Wesentlichen auf Daten-Spenden angewiesen. Dabei stoßen wir auch im akademischen Kontext auf großes Misstrauen potenzieller Spender:innen im Bereich der Datensicherheit. Tatsächlich beinhalten E-Mails persönliche und personenbezogene, also sogenannte „sensible Daten“, deren Löschung und Anonymisierung gerade bei multilingualen Daten, die in Emails keinesfalls ausgeschlossen werden können, eine Hürde darstellen kann.

Zwar besteht durch die Digitalisierung gleichfalls in der linguistischen Forschung die Möglichkeit immer komplexere Daten in Korpora zusammenzufassen, dieses Unterfangen setzt jedoch eine Infrastruktur und Datenmodelle voraus, die dieses unterstützen. Am Beispiel

Publiziert in: Vincent Heuveline, Nina Bisheh und Philipp Kling (Hg.): E-Science-Tage 2023. Empower Your Research – Preserve Your Data. Heidelberg: heiBOOKS, 2023. DOI: https://doi.org/10.11588/heibooks.1288.c18070 (Freier Zugang – alle Rechte vorbehalten)

des Projekts *Zwischen Briefen und E-Mails: Dynamiken der Normierung und Standardisierung* werden die Herausforderungen thematisiert, denen wir begegnen, wenn es um die Akquise, Verarbeitung und Publikation von mehrsprachigen IBK-Daten geht.

1 Einleitung

50 Jahre E-Mail-Verkehr

Die Kommunikation per E-Mail findet nunmehr seit 50 Jahren statt. Diese computerbasierte, zuerst für professionelle Kontexte entwickelte Methode der Nachrichtenübermittlung hat sich zwischen dem Ende der 1970er und den 1990er Jahren zu einer eher informellen bzw. persönlichen Kommunikationsform entwickelt (Rentel und Schröder 2018). Die in den letzten Jahrzehnten sichtbar gewordene materielle und soziale Metamorphose, die der Brief in Richtung der E-Mail durchlief (ebd.), führte zu einer starken Heterogenität der einzelnen Mails z.B. in ihren Formalitätsgraden und den Kommunikationssituationen. In einer Prognose von 2022 geht Statista (2023) davon aus, dass im Jahr 2023 347 Milliarden E-Mails geschrieben werden – Tendenz steigend. Dieser steigenden Kommunikation zum Trotz ist die linguistische Auseinandersetzung bzw. das Interesse an der linguistisch-kommunikativen Erforschung dieser Form der Kommunikation rückläufig. Nach einer relativ großen Zahl von europäischen Arbeiten um die Jahrtausend-Wende (Baron 1998; López Alonso und Seré 2003; Ziegler und Dürscheid 2007; Anis 1999) ist das Forschungsinteresse an der E-Mail-Kommunikation seither zurückgegangen. Gerade die aktuelle Heterogenität und die Entwicklung der E-Mails hin zu einer oft in distanzierteren, in professionellen Kontexten genutzten Kommunikationsform wirft Fragen rund um die Standardisierung von E-Mails auf: In welchen Schreibsituationen wird eine Standardisierung am deutlichsten manifest? Wie zeigt sie sich im sprachlichen Gebrauch in den einzelnen romanischen Sprachen und hier vor allem im Spanischen, Portugiesischen und Französischen? Gibt es neben den impliziten Normen auch explizite Festschreibungen einer empfohlenen Verwendung? Was wissen wir über die Anwendung spezifischer einzelsprachlicher oder auch diskursiver Empfehlungen?

Derartige Fragen lassen sich nur auf Grundlage einer exhaustiven Menge an Forschungsdaten analysieren. Hier liegt ein Grund, warum die E-Mail-Kommunikation nach einem ersten lebhaften Forschungsinteresse aus dem linguistischen Forschungsfokus rückte. Der aufkommende prominentere Datenschutz und die damit verbundenen Einschränkungen in der Datenakquise erschweren die Erstellung umfassender E-Mail-Korpora. Die Hürden und Herausforderungen auf dem Weg zu einem mehrsprachigen Korpus aus spanischen, französischen und portugiesischen E-Mails, das zur Beantwortung der genannten Forschungsfragen dienen könnte, sind Gegenstand der vorliegenden Abhandlung.

2 Erstellung von Korpora der E-Mail-Kommunikation

2.1 Herausforderungen

Um Normierungs- bzw. Standardisierungsprozesse in E-Mails quantitativ zu erforschen, sollte das zu erstellende Korpus möglichst syn-, aber auch diachrone Daten enthalten, um auch einzelne Perioden der Standardisierung nachzeichnen zu können. Eine hohe Anzahl von Schreiber:innen aus den drei Sprachräumen ist die Voraussetzung, um wechselnde Schreibsituationen sowie die eingangs erwähnte Heterogenität abbilden zu können. Unsere Daten müssen Metadaten wie Sozioprofession, Generationenzugehörigkeit, Formalität des Schreibkontextes etc., beinhalten, um Variation, Innovation, Wandel und Standardisierung an verschiedenen Schreibsituationen nachzuvollziehen und benennen zu können.

Bei der Datenerhebung stellen sich demzufolge bereits zwei Herausforderungen. Eine erste ist die rechtliche Herausforderung des Speicherns und Verarbeitens jener Daten, die durch den Gesetzgeber geschützt sind und personenbeziehbare Informationen beinhalten (siehe §3 BDSG bzw. auf Europäischer Ebene Artikel 5 der Europäischen Datenschutz-Grundverordnung). Die zweite Herausforderung ist ethischer Natur. In E-Mails kommen nicht nur personenbeziehbare, sondern zugleich auch persönliche Informationen, wie Haltungen und Meinungen zum Ausdruck. Dies führt gerade bei der Datenakquise zu Zurückhaltung in der Zustimmung möglicher Spender:innen zur Nutzung der Daten als linguistische Forschungsgrundlage. Die Datensammlung und -verarbeitung ist nur im Einklang mit einer Datenanonymisierung möglich, um einerseits im rechtlichen Rahmen zu forschen und andererseits, den Spender:innen das nötige Vertrauen in unsere Forschung und den Schutz ihrer Daten zu bieten.

2.2 Die Sammlung erster Test-Daten

In einem Pilotprojekt haben wir 2021 zunächst zur Spende von E-Mails bei französischen und spanischen Stiftungen und Vereinen, die im sprachlichen Bereich agieren, aufgerufen. Da ein E-Mailverlauf häufig aus mehr als nur einer Nachricht mit einem Absender und einem Empfänger besteht, stellte sich das Einholen der sogenannten informierten Einwilligung als schwierig heraus. Im Regelfall sind in einer E-Mailkonversation mehrere Personen direkt oder indirekt involviert, deren Einverständnis nur schwer zu erhalten ist.

Trotz dieser Hürde konnten wir ein Sample von ca. 1000 französischen und spanischen E-Mails nutzen, um einen ersten Schritt in die Richtung einer automatisierten Datenverarbeitung zu gehen. Eine solche bietet uns die Möglichkeit, das Risiko für uns, aber zugleich für die Spender:innen. zu minimieren und Daten gesetzeskonform und ethisch sammeln sowie verarbeiten zu können.

3 Ein Weg in Richtung austomatisierten Datenverarbeitung

3.1 Der anonymizer zur Anonymisierung mehrsprachiger Daten

Zur Datenverarbeitung haben wir die universitäre Infrastruktur der Universität Heidelberg bemüht. Hier hat uns das *Scientific Software Center* (SSC) bei der Erstellung eines Algorithmus zur Anonymisierung unterstützt. Das SSC ist dem *Interdisziplinären Zentrum für Wissenschaftliches Rechnen* (IWR) der Universität Heidelberg angegliedert.

Der entwickelte Algorithmus ist als beta version auf github zu finden (Git Hub Repository: `https://github.com/ssciwr/anonymize`). Es handelt sich um einen Prototypen in das .eml Dateien eingelesen und.txt Dateien ausgeworfen werden. Ein erstes Modul säubert den Umgebungstext (An, Von, Datum, Betreff etc.) und extrahiert Sätze mit Satzerkennungstool *SpaCy* (Montani u. a. 2023), auf Grundlage der NLP-Modelle „fr_core_news_sm“ „es_core_news_sm“.

Ein zweites Modul wendet das Named-Entity-Recognition (NER)-Tools *Stanza* auf die Sätze an. Pro Satz werden also mit *Stanza* (Qi u. a. 2020) persönliche Daten (*Named Entities*) extrahiert und durch die jeweiligen Entitäten-Namen (Person, Organisation, Orte) ersetzt. Da *Stanza* auf einsprachigen Modellen basiert, bedarf der Algorithmus demzufolge einer Voreinstellung für die gewünschte Sprache. Da die NER für die einzelnen Sprachen auf verschiedenartig trainierten Modellen basiert, sind die Anonymisierungs-Ergebnisse für unsere französischen und spanischen Test-E-Mails unterschiedlich ausgefallen.

3.2 Technische Hürden der Anonymisierung

Einige Hürden stellten sich uns bei der Anonymisierung mit den tools *SpaCy* und *Stanza* in den Weg. Da die Satztrennung erheblich zu einer korrekten Erkennung von NER beiträgt, beginnt die Schwierigkeit bei der korrekten Erfassung eines Satzes in den E-Mails. Wie bereits eingangs ausgeführt, sind E-Mails allerdings eine heterogene Kommunikationsform und vereinen Merkmale aus verschiedenen Textsorten und Diskurstraditionen, wie Brief, Textnachricht oder administratives Schreiben. Sie können sowohl Spuren von distanzkommunikativen Schreibens als auch Nähe-Markierungen (Koch und Oesterreicher 1985) aufweisen. Beide tools, *SpaCy* als auch *Stanza*, sind auf distanzsprachlichen Modellen trainiert, was zu zahlreichen Fehlern insbesondere in den französischsprachigen E-Mails führte.

Ein weiteres Hindernis ist der Umgebungstext, der in unseren Test-E-Mails uneinheitlich und deshalb schwierig zu säubern war. Da dieser stark von den sprachlichen Voreinstellungen des jeweiligen, die E-Mails generierenden PC's abhängt, müssen an diesem Punkt mehrsprachliche Umgebungstexte noch stärker berücksichtigt werden.

Bei der Benutzung von *Stanza* wird darüber hinaus ein generelles Problem der Behandlung von Zahlen unterschiedlicher Formate deutlich. Sowohl Postleitzahlen als auch Telefonnummern werden in der beta-Version nicht anonymisiert.

Auch Signaturen sind sehr vielgestaltig aufgebaut und können so unterschiedlich gut extrahiert werden. Eine voreingestellte Signatur kann Informationen, z.B. Grußformeln, enthalten, die für die Analyse wichtig sind. Sie kann aber auch gesponsorte Werbungen aufweisen, die statistische Zugriffe verfälschen dürften, da der Wordcount von ihnen betroffen ist. Da die Erkennung personenbezogener Daten in *Stanza* auf einem Datenmodell aus Wikipe-dia-Artikeln (*wikiner*) basiert, kommt es außerdem vor, dass die NER zu sensibel ist. Zu viele Entitäten oder auch zu wenige werden als Personen, Organisationen oder Orte ausgewiesen. Im Gegensatz zu dem französischen Wikipedia-Modell, basiert das spanische NER-Modell in *Stanza* auf Medientexten. Es konnte festgestellt werden, dass die Anonymisierung unserer spanischen Testdaten weniger Fehler in Bezug auf die Sensibilität enthielt.

3.3 Die transformers Datenbank als ein Lösungsansatz

Ein erster Lösungsansatz, den wir gemeinsam mit dem SSC gefunden haben, ist die Benutzung der *transformers* Datenbank. Anders als *Stanza* ist die *transformers*-Datenbank mehrsprachig, weshalb die Sprache nicht mehr ausgewählt werden muss; das Tool basiert auf einem generalisierten multilingualen Korpus und ist so auf verschiedensprachige Datensets anwendbar. Die ersten Tests mit der NER aus der Datenbank zeigten sowohl in den französischen, als auch in den spanischen E-Mails verbesserte Ergebnisse in der *Named Entity Recognition*. Gemeinsam mit einem stärkeren Fokus auf die Säuberung der Rohdaten, erscheint der Wechsel des NER-Tools als für das zukünftige Projekt vielversprechend. Eine zweite Möglichkeit wäre ein eigenes *One shot training*. Hier würde ein von den Ingenieur:innen des SSC entwickeltes Modell basierend auf von uns annotierten Daten trainiert. Diese Lösung bedeutet allerdings einen erhöhten Aufwand im Projekt, da eine umfangreichere Menge Daten manuell annotiert werden müsste.

4 Zur Datenveröffentlichung

Ähnlich wie in Beißwenger u. a. (2017) beschrieben, stellt sich auch für unser Projekt im Anschluss an die Verarbeitung der Daten die ethische und rechtliche Frage nach der Veröffentlichung sprachgebrauchsbezogenen Daten in der linguistischen Forschung. Diese stellt sich im Übrigen zugleich bei Projekten, die mit *Twitter* oder *Telegram*-Nachrichten als Datengrundlage arbeiten. Die Nutzer stimmen bereits bei der Anmeldung auf den einschlägigen Plattformen der Verarbeitung ihrer Daten zu – häufig ohne sich dessen im Detail bewusst zu sein. Es bleiben persönliche und teils auch personenbeziehbare Daten, die wir für die Forschung benutzen und somit auch in Ausschnitten veröffentlichen möchten. Aus dem von Beisswenger et al. vorgestellten Rechtsgutachten zum Dortmunder Chatkorpus geht hervor, dass Daten aus der internet-basierten Kommunikation durch die sehr unterschiedlichen enthaltenen personenbezogenen Informationen rechtlich anders behandelt werden sollten. In diesem Beispiel, das im Bereich des Datenschutzes auch auf unser Projekt übertragbar ist, wurden so Subkorpora zu verschiedenen Kommunikati-

onssituationen gebildet, um die Daten kontextbezogen auf ihre Personenbeziehbarkeit zu prüfen.

Zur rechtlichen Komponente kommt noch die Methodische hinzu: verschiedene Kontexte müssen zwingend klassifiziert werden und würden im Anschluss datenschutzrechtlich nuanciert behandelt werden. Im Chatkorpus von Beisswenger et al. wurden so Daten mit besonders hoher Sensitivität aus z.B. psychosozialen Beratungen vollständig aus dem Korpus gelöscht. Bereits die Erfassung dieser Daten gilt laut des Gutachtens als unzulässig, wenn sie, unter anderem, zum Zwecke der Veröffentlichung dient. Das für das Chatkorpus in Auftrag gegebene Rechtsgutachten weist zudem darauf hin, dass bei der Erhebung bereits ein konkreter Zweck für die Datensammlung angegeben werden müsse (Montani u. a. 2023), hier wird die Forschung bislang nicht unter § 28 Abs. 3 Nr. 4 BDSG (Archivzwecken im öffentlichen Interesse) geführt.

5 Die zukünftige Fusion der Datensammlung und Datenverarbeitung in einer Spendenwebseite

Zukünftig soll aus dem Projekt eine Spendenwebseite im Netz werden, bei der E-Mail-Spender:innen E-Mails direkt hochladen und so keine Interaktion von den Spender:innen mit den Forschenden mehr nötig ist. Während andere E-Mail-Korpora auf Datensets aus online Archiven basieren, wie das in der Germanistik ansässige Projekt *CodE Alltag* (Krieg-Holz u. a. 2016), sind Spendenwebseiten gerade bei internetbasierter Kommunikation bereits an anderer Stelle genutzt worden. So konnte das Projekt *What's up*, (Ueberwasser und Stark 2017) 617 Chatverläufe sammeln. Derzeit läuft ein weiteres Projekt der Universitäten Lothringen, Lüttich und Strasbourg bei dem Audionachrichten gesammelt werden (Glikman und Fauth 2022).

Die von uns angedachte Webseite soll auch die umgehende Anonymisierung der Daten ermöglichen. Dazu planen wir, auf dem Pilotprojekt aufzubauen und den in Zusammenarbeit mit dem SSC entwickelten Algorithmus in die Webseite zu integrieren. Des Weiteren sollen die Nutzer:innen der Webseite während des *uploads* Angaben zu ihrer Person und der Schreibsituation machen können. Die Spender:innen sollen unmittelbar die Möglichkeit erhalten, die Daten zur Verarbeitung an die Universität freizugeben und damit zugleich eine sogenannte *Informierte Einwilligung* für den Verlauf erteilen. Wir stellen im Gegenzug Transparenz in Bezug auf die Datennutzung her, indem wir unser Projekt vorstellen und die Spender:innen über die Art der Weiterverarbeitung informieren. So können wir Metadaten sammeln, die wir zu unserer Auswertung verwenden, aber gleichzeitig technisch absichern, dass es sich um nutzbare Daten handelt. In einem letzten Schritt kann mit Hilfe der Metadaten und Einwilligungen die Möglichkeit einer Veröffentlichung unseres Korpus rechtlich geprüft werden.

Danksagung

Besonderer Dank gilt dem Scientific Software Center der Universität Heidelberg und speziell Frau Inga Ulusoy für die Entwicklung des *anonymizer* Algorithmus.

Literaturverzeichnis

Anis, Jacques. 1999. *Internet communication et langue française.* 191. Paris: Hermes Sciences Publications. ISBN: 2-7462-0063-5.

Baron, Naomi S. 1998. „Letters by phone or speech by other means: the linguistics of email". *Language and Communication* 18 (2): 133–170. DOI: https://doi.org/10.1016/S0271-5309(98)00005-6.

Beißwenger, Michael, Hrsg. 2017. *Empirische Erforschung internetbasierter Kommunikation.* Berlin, Boston: De Gruyter. DOI: https://doi.org/10.1515/9783110567786.

Beißwenger, Michael, Harald Lüngen, Jan Schallaböck, John H. Weitzmann, Axel Herold, Pawel Kamocki, Angelika Storrer und Julia Wildgans. 2017. „Rechtliche Bedingungen für die Bereitstellung eines Chat-Korpus in CLARIN-D: Ergebnisse eines Rechtsgutachtens". In *Empirische Erforschung internetbasierter Kommunikation,* 7–46. Berlin, Boston: De Gruyter. DOI: https://doi.org/10.1515/9783110567786-002.

Delfa, Christina Vela. 2021. *La comunicación por correo electrónico: análisis discursivo de la correspondencia digital.* Madrid,Frankfurt: Iberoamericana;Vervuert.

Glikman, Julie, und Camille Fauth. 2022. „Un nouvel accès à la parole spontanée : les vocaux". In *XXXIVe Journées d'Études sur la Parole – JEP 2022.* ISCA. DOI: https://doi.org/10.21437/JEP.2022-17.

Große, Sybille. 2012. „Sprache und Öffentlichkeit in realen und virtuellen Räumen". Kap. Französische E-Mails: Briefmodelle im Abschwung, herausgegeben von Annette Gerstenberg, Claudia Polzin-Haumann und Dietmar Osthus, 126–139. Romanistischer Verlag. ISBN: 978-3-86143-202-9.

Koch, Peter, und Wulf Oesterreicher. 1985. „Sprache der Nähe — Sprache der Distanz. Mündlichkeit und Schriftlichkeit im Spannungsfeld von Sprachtheorie und Sprachgeschichte". *Romanistisches Jahrbuch* 36 (1): 15–43. DOI: https://doi.org/10.1515/9783110244922.15.

Krieg-Holz, Ulrike, Christian Schuschnig, Franz Matthies, Benjamin Redling und Udo Hahn. 2016. „CodE Alltag: A German-Language E-Mail Corpus". In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16),* 2543–2550. Portorož, Slovenia: European Language Resources Association (ELRA).

López Alonso, Covadonga, und Arlette Seré, Hrsg. 2003. *Nuevos Generos Discursivos: Los Textos Electronicos.* 219. Madrid: Biblioteca nueva. ISBN: 978-8497422017.

Montani, Ines, Matthew Honnibal, Matthew Honnibal, Sofie Van Landeghem, Adriane Boyd, Henning Peters, Paul O'Leary McCann u. a. 2023. *explosion/spaCy: v3.5.2: Pretraining improvements, bug fixes for spans and spancat and more.* DOI: https://doi.org/10.5281/zenodo.7820813.

Qi, Peng, Yuhao Zhang, Yuhui Zhang, Jason Bolton und Christopher D. Manning. 2020. „Stanza: A Python Natural Language Processing Toolkit for Many Human Languages". In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations.* Association for Computational Linguistics. DOI: https://doi.org/10.18653/v1/2020.acl-demos.14.

Rentel, Nadine, und Tilman Schröder. 2018. *Sprache und digitale Medien.* Berlin: Peter Lang Verlag. DOI: https://doi.org/10.3726/b12951.

Souchier, Emmanuel, Étienne Candel, Gustavo Gomez-Mejia und Valérie Jeanne-Perrier. 2019. *Le numérique comme écriture. Théories et méthodes d'analyse.* Paris: Armand Collin. ISBN: 978-2-200-61858-2.

Statista. 2023. „Number of sent and received e-mails per day worldwide from 2017 to 2026". Besucht am 15. Mai 2023. https://www.statista.com/statistics/456500/daily-number-of-e-mails-worldwide/.

Ueberwasser, Simone, und Elisabeth Stark. 2017. „What's up, Switzerland? A corpus-based research project in a multilingual country". *Linguistik Online* 84 (5). DOI: https://doi.org/10.13092/lo.84.3849.

Ziegler, Arne, und Christa Dürscheid, Hrsg. 2007. *Kommunikationsform E-Mail.* Tübingen: Stauffenburg Verlag Brigitte Narr GmbH. ISBN: 978-3-86057-686-1.