Orte in narrativen biographischen Interviews: automatische Methoden und manuelle Analysen

Ruppenhofer, Josef
ruppenhofer@ids-mannheim.de
Leibniz-Institut für Deutsche Sprache, Deutschland

Flinz, Carolina
carolina.flinz@unimi.it
Universität Mailand

Schmidt, Thomas
thomas.schmidt@ids-mannheim.de
Leibniz-Institut für Deutsche Sprache, Deutschland

Das sogenannte Israelkorpus ist ein Korpus gesprochener Sprache, das von Anne Betten und MitarbeiterInnen in den Jahren 1989 bis 2012 erstellt wurde und aus 274 Aufnahmen narrativer autobiographischer Interviews mit Emigranten aus deutschsprachigen Regionen Mitteleuropas besteht, die vorwiegend in den 1930er Jahren zur Auswanderung gezwungen wurden. Es besteht aus drei Subkorpora, die unter anderem in der Datenbank für Gesprochenes Deutsch (DGD), einem Korpusmanagementsystem des Leibniz-Instituts für Deutsche Sprache, abrufbar und recherchierbar sind: IS (Emigrantendeutsch in Israel), ISW (Emigrantendeutsch in Israel: Wiener in Jerusalem) und ISZ (Zweite Generation deutschsprachiger Migranten in Israel).

In unserem Beitrag untersuchen und vergleichen wir automatische und manuelle Zugänge zu Ortsnennungen in den biographischen Interviews des ISW-Korpus. Orte spielen im Israelkorpus eine besondere Rolle. Einerseits dienen sie als geographische Bestimmungen vor, während und nach der Emigration. Andererseits haben sie auch eine Funktion innerhalb der Erinnerungsarbeit während der Interviews, die sehr stark mit der emotionalen Dimension verbunden ist. Im Rahmen des Projekts *Orte und Erinnerung. Eine Kartographie des Israelkorpus* soll die emotionale Funktion von Ortsnennungen auf dem ganzen Israel-Korpus untersucht werden. Unsere Arbeit gehört daher in die Reihe neuerer Arbeiten, die mit korpuslinguistischen Methoden das gesamte Korpus untersuchen (Flinz 2019; Flinz/Brambilla 2019), während bisherige grammatische, syntaktisch-stilistische oder dialoglinguistische Untersuchungen am Israel-Korpus sich auf wenige qualitativ untersuchte Interviews beschränkten.

Um die Untersuchung der Ortsnennungen im Gesamtkorpus zu ermöglichen, müssen wir das relevante Ortskonzept so operationalisieren, dass computerlinguistische und NLP-Werkzeuge Ortsnennungen mit hoher Präzision und einer hohen Trefferquote (EN recall) auffinden und den menschlichen Analystinnen zur Valdierung und Interpretation vorlegen können. In unserer Pilotstudie evaluieren wir hier zunächst, wie gut sich verfügbare Werkzeuge und Ressourcen ‘out of the box’ dafür eignen. Zu diesem Zweck haben wir das relevante Ortskonzept händisch durch Expertenannotation auf alle Transkripte des ISW-Teilkorpus angewendet. Parallel dazu haben wir die Daten einerseits mithilfe eines state-of-the art Named-Entity-Recognizers (Akbik et al 2018) annotiert, der unter anderem auch Orte (‘Locations’) auszeichnet, und andererseits programmatisch alle Wörter, die von GermaNet (Hamp & Feldweg 1997) dem Wortfeld ORT zugeordnet werden, als Ort annotiert.

In der Auswertung der parallelen Annotationen zeigen wir, dass das Ortskonzept mit guter Übereinstimmung händisch annotiert werden kann (kappa > 0.9), es aber weder durch die Annotationen des NER-Systems noch durch die Annotationen auf der Grundlage von GermaNet adäquat erfasst wird. Die NER-Annotationen decken nur Eigennamen ab (z.B. *Wien*, *Israel*), während für die Auswertung auch Bezeichnungen durch Appellativa (z.B. *Lager*, *Ausland*, *Grenze*) sehr wichtig sind. Die Erkennung von Ortsnennungen in Form von Appellativa mithilfe von GermaNet ist für unsere Forschungsfragen ebenfalls nicht ausreichend. Einerseits werden viele Konzepte, deren Bedeutung eine Ortsfacette (im Sinne von Cruse und Croft 2004) besitzen, von GermaNet nicht mit einer eigenen Ortsbedeutung ausgewiesen. Ein zentrales Beispiel hierfür ist *Schule*, welches in GermaNet den Wortfeldern GRUPPE, KOGNITION, GESCHEHEN und ARTEFAKT zugeordnet wird, aber nicht dem Wortfeld ORT. Andererseits besitzen viele relevante Wörter wie im Fall von *Schule* auch andere Bedeutungsfacetten und die räumliche ist im konkreten Kontext nicht unbedingt wichtig (z.B. im Fall der Erwähnung einer ‘Schule’ der Musikgeschichte).

Der Vergleich der händischen und automatischen Annotationen legt nahe, dass wir für die automatische Annotation aller Orte in den Israel-Korpora eine auf das Problem zugeschnittene Lösung brauchen. Die Annotationen eines auf unseren Expertenannotationen trainierten Systems werden wir mit den oben genannten out-of-the-box Annotationen vergleichen. Als weiteren Ansatz werden wir automtisch alle Wörter/Konzepte identifizieren, die in der Wikidata-Ressource (Vrande#i# und Krötzsch 2014) einen Link zu OpenStreetMap besitzen und damit implizit als geographisch lokalisierbar ausgewiesen werden. Beispielsweise wird dadurch das Konzept *Schule* erfasst, das von GermaNet nicht als ORT ausgewiesen wird.

Neben praktischen Erkenntnissen hat uns der Vergleich der manuellen und automatischen Annotationen auch auf theoretische Fragen und Möglichkeiten aufmerksam gemacht, die wir vorher nicht Betracht gezogen hatten. So erwägen wir, die von uns im weiteren Arbeitsverlauf

manuell korrigierten NER-Annotationen als separate Sicht auf die Daten in die korpuslinguistische Analyse der Beziehung zwischen Orten und Emotionen mit einzubeziehen.

Bibliographie

Akbik, Alan / Blythe, Duncan / Vollgraf, Roland (2018): "Contextual string embeddings for sequence labeling." *Proceedings of the 27th International Conference on Computational Linguistics*.

Brambilla, Marina / Flinz, Carolina (2019): Orte und entgegengesetzte Emotionen (Liebe und Hass) im Korpus ISW. In: Studi Germanici. Im Druck.

Croft, William, / Cruse, D. Alan (2004): *Cognitive linguistics*. Cambridge University Press.

Flinz, Carolina (2019): “Multiword units and N-Grams naming FEAR in the Israel-Corpus”. Corpas Pastor, G. / Mitkov, R. Computational and Corpus-Based Phraseology. Springer Verlag. Lecture Notes in Computer Science. 86—98.

Hamp, Birgit / Feldweg, Helmut (1997): "GermaNet - a Lexical-Semantic Net for German." Proceedings of the ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications. Madrid.

Vrande#i#, Denny and Markus Krötzsch (2014). "Wikidata: a free collaborative knowledgebase". *Commun. ACM* 57, 10 (September 2014). 78-85.