Kombinierte Text- und Geo-Suche zum Durchsuchen einer Georeferenzierten Online-Bibliographie

Bastian Entrup[1], Vera Ermakova[2], Ines Schiller[2] und Henning Lobin[2]

[1]Angewandte Sprachwissenschaft und Computerlinguistik
bastian.entrup@germanistik.uni-giessen.de
[2]Zentrum für Medien und Interaktivität
{vera.ermakova|ines.schiller|henning.lobin}@zmi.uni-giessen.de
Justus-Liebig Universität Gießen
Germany

1 Einleitung und Motivation

Das GeoBib Projekt entwickelt eine georeferenzierte Online-Bibliographie der frühen Holocaust- und Lagerliteratur zwischen 1933 und 1949 mit über 700 Werken und ca. 850 Autoren und Herausgebern. Anders als eine klassische Bibliographie werden auch handlungsrelevante Orte sowie biographische Daten inklusive einer schriftlichen Biographie zu Autoren und Herausgebern erfasst. Die bibliographischen Daten umfassen zusätzlich Informationen wie sie für ein Literaturlexikon nicht unüblich sind, z.B. Rezeptionen und Werksgeschichten. Die Kombination und Verlinkung dieser Entitäten, Personen, Werke und Orte, macht das Besondere und den Mehrwert der Online-Bibliographie aus.

2 Funktionen und Implementation

2.1 Funktionen und Implementation der Text-Suche

Um die entstehende Bibliographie und die dafür erstellten Texte (z.B. die Inhaltszusammenfassungen oder die Autorenbiographien) sowie die bibliographischen Daten (Autoren, Herausgeber, Verlag usw.) durchsuchbar zu machen wird Apache Solr [1] und das Open Source Projekt *glp4lucene*[2] zur Verarbeitung von Suchanfragen und Erstellung des Indexes verwendet.

Die natürlich Variabilität und Ambiguität einer Sprache machen Verarbeitungsschritte aus dem Bereich des Natural Language Processings (NLP) notwendig. Im Bereich des Information Retrievals (IR) hat sich das Stemming als einfaches, regelbasiertes Verfahren zur Vereinheitlichung unterschiedlicher Wortformen auf einen gemeinsamen Stamm durchgesetzt (vgl. [3,5]). Aus linguistischer

[1] https://lucene.apache.org/solr/.

[2] Zu finden unter `https://sourceforge.net/projects/glpforlucene/`. Das Paket umfasst die Ergänzung von Synonymen, eine Lemmatisierungsfunktion sowie eine Termgewichtungsmethode, die auf der Wortart der Terme basiert.

Sicht ist Stemming jedoch nicht so erstrebenswert wie eine Lemmatisierung, die Reduktion von verschiedenen Wortformen auf ein gemeinsames Lemma, da beim Stemming die Ambiguität einer Sprache erhöht wird. Das hier genutzte Verfahren basiert auf dem MATE Tool [2] und verwendet das in [9] beschriebene deutsche Modell.

Basierend auf einem Lemma können Synonyme in GermaNet [4] nachgeschlagen und dem Suchindex hinzugefügt werden. Das Hinzufügen der Synonyme geschieht schon während der Indexierung der Daten[3].

Die einfache textbasierte Suche durchsucht die wahrscheinlichsten Suchfelder nach einem Suchbegriff und nutzt dabei die Lemmatisierung des Indexes, um deklinierte oder konjugierte Formen zu finden. Zusätzlich sind im Index Synonyme vorhanden, so dass eine Suche nach *Gefängnis* auch Vorkommnisse von z.B. *Zuchthaus* findet. Die Suchergebnisse sind nach verschiedenen Personen-, Text- und Ortskategorien facettiert (s. Abb. 1).

Die Erweiterte-Suche liefert entweder Texte, Autoren/Herausgeber oder Orte als Ergebnis zurück. Wenn nach Texten gesucht wird, kann die Suche nach biographischen Daten der Autoren/Herausgeber (z.B. Name, Geburtsjahr oder Sterbeort), aber auch nach bibliographischen Daten (z.B. nach dem Verlag, dem Erscheinungsjahr oder -ort) gefiltert werden. Eine mögliche Suchanfrage wäre z.B. *Texte, deren Autoren weiblich sind.* Ähnliche Einschränkungen sind auch für Personensuchen möglich; z.B.: *Eine Autorin, die im Jahr 1939 einen oder mehrere Texte bei einem bestimmten Verlag veröffentlicht hat.*

2.2 Funktionen der Geo-Suche

Die Geo-Suche basiert auf einem Kartensatz Europas zur Zeit zwischen 1939 und 1945, der speziell für dieses Projekt aus verschiedenen Datensätzen kompiliert wurde (vgl. [6,7]). Für jedes Jahr wurde versucht, eine vollständige Karte mit den Grenzen Europas zu erstellen [8]. Die Jahre können über einen Slider unter der Karte ausgewählt werden. Auf der Karte dargestellte Datensätze können durch Klicken, Zoomen oder andere Werkzeuge ausgewählt werden.

Orte können in ein Suchfeld eingeben werden. Auf Grund der hohen Ambiguität von Toponymen wird dem User bei der Eingabe eines Ortsnamens eine Liste mit Vorschlägen angezeigt. Auf diese Weise gefundene Orte können dann mit einer Umkreissuche erweitert werden. Das ermöglicht zielgenaue regionale Recherchen, die für Heimatforscher und pädagogische Zwecke sinnvoll sind.

Ein Graph unterhalb der Karte (s. Abb 2) zeigt die Häufigkeit von Ereignissen (z.B. Anzahl von Handlungsorten) für jedes Jahr an. So können auf einen Blick Schwerpunkte ausgemacht werden. Ein Slider unter diesem Graph macht

[3] Das ist bei dem relativ kleinen Datensatz vertretbar. Im Vergleich zu einer Verarbeitung während des Suchvorgangs, sorgt dies für eine geringere Auslastung und weniger Wartungsbedarf des resultierenden Systems. Das verwendete Software Paket erlaubt allerdings beide Möglichkeiten.

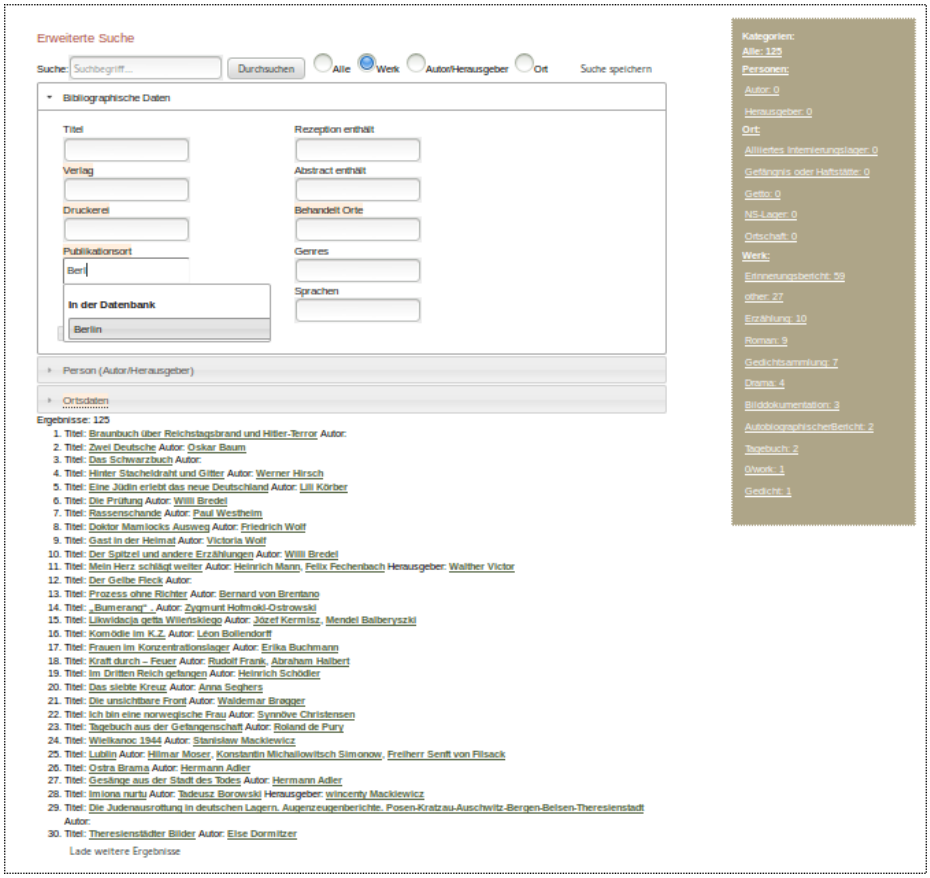

Abb. 1. Screenshot eines aktuellen Prototypen: Darstellung der Suchergebnisse und Vorschau der Eingabemaske.

es möglich sich nur Handlungsorte eines bestimmten Zeitraumes anzeigen zu lassen.

2.3 Verbindung von Text- und Geo-Suche

Die Verbindung der verschiedenen Entitäten in der Datenbank macht die Kombination der beiden Systeme möglich. Autoren/Herausgeber sind mit ihren Geburts- und Sterbeorten verbunden, außerdem mit den Orten in den von ihnen geschriebenen Texten. Werke sind mit ihren Erscheinungs- und Handlungsorten verbunden.

Die Beispiel-Suchanfragen können wie folgt ergänzt werden: *Texte, deren Autoren weiblich sind und in Berlin geboren wurden* und *Eine Autorin, die im Jahr*

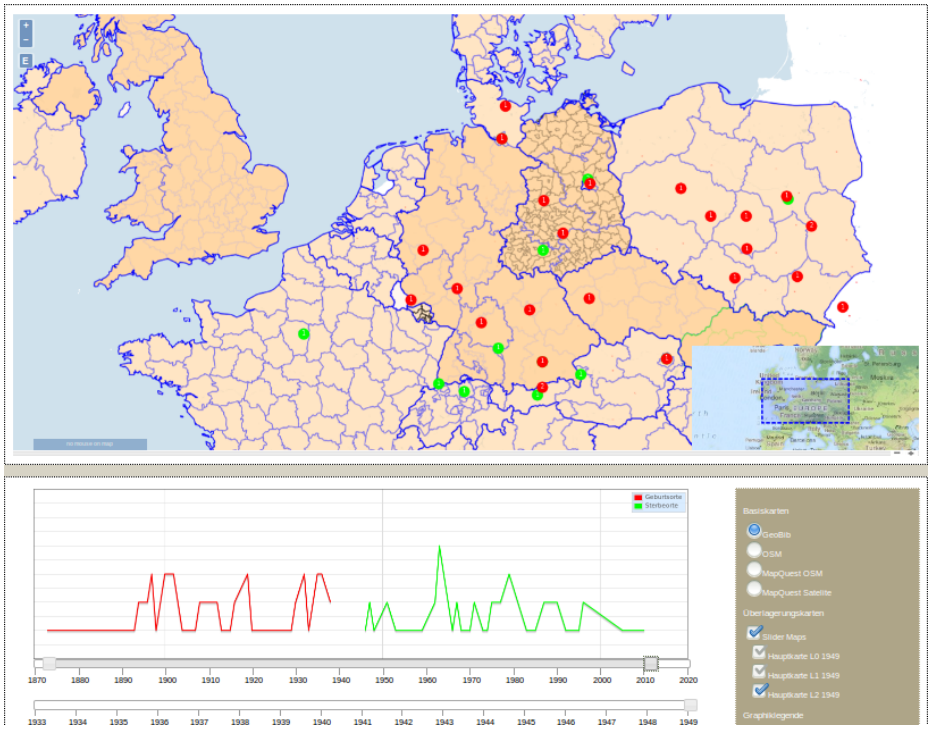

Abb. 2. Screenshot eines aktuellen Prototypen: Karte mit Angezeigten Geburts- (rot) und Sterbeorten (grün) basierend auf 125 Beispieltexte in den Grenzen von 1949.

1939 einen Text über Geschehnisse in Auschwitz bei einem bestimmten Verlag veröffentlicht hat. Auch Texte von Autoren, die in einer bestimmten Region geboren wurden oder Texte, die von einem bestimmten Lager handeln, sind so auffindbar.

Umgekehrt sind aber auch Orte auffindbar, die als Handlungsort zu bestimmten Zeiten eine Rolle spielen oder die Publikationsorte von bestimmten Werken sind. So lassen sich alle Orte finden, die in Werken eines bestimmten Autoren vorkommen.

3 Aussicht

Die Verbindung von Texten mit Geo-Daten ist nicht nur eine besondere Herausforderung an die Darstellung, die Organisation und die Suche nach Informationen, sondern bietet viele Möglichkeiten: Die Verteilung von Handlungsorten der Texte auf einer Karte bietet ein räumliches Verständnis eines Textes oder einer Sammlung von Texten. Besondere lokale Schwerpunkte können auf einen Blick erfasst werden.

Viele der im Projekt erfassten Texte gelten heute als vergessen. Sie werden nun das erste Mal systematisch durchsuchbar gemacht. Die Kombination von bibliographischen, biographischen, geographischen und inhaltlichen Daten ermöglicht einen völlig neuen (räumlichen) Zugang zu den Texten und den Ereignissen des Holocaust. So sind das Stellen und die Beantwortung neuer Forschungsfragen auf Grundlage einer breiten Textbasis und unter Berücksichtigung der geographischen Verteilung möglich.

Literatur

1. Binder, F., Entrup, B., Schiller, I., Lobin, H.: Uncertain about Uncertainty: Different Ways of Processing Fuzziness in Digital Humanities Data. In: Digital Humanities 2014, Book of Abstracts, pp. 97-100. Ecole Polytechnique Fédérale de Lausanne (EPFL) and The University of Lausanne (UNIL), Switzerland, 7-12 July 2014 (2014), http://dharchive.org/paper/DH2014/Paper-874.xml
2. Bohnet, B.: Very High Accuracy and Fast Dependency Parsing is Not a Contradiction. In: Proceedings of the 23rd International Conference on Computational Linguistics. pp. 89–97. COLING '10, Association for Computational Linguistics, Stroudsburg, PA, USA (2010)
3. Braschler, M., Ripplinger, B.: How Effective is Stemming and Decompounding for German Text Retrieval? Information Retrieval 7(3-4), 291–316 (2004)
4. Hamp, B., Feldweg, H.: GermaNet - a Lexical-Semantic Net for German. In: Proceedings of ACL Workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications. pp. 9–15 (1997)
5. Kraaij, W., Pohlmann, R.E.: Viewing Stemming as Recall Enhancement. In: In Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 40–48 (1996)
6. Schaarschmidt, S.: Bestandserhebung zu verfügbaren digitalen geographischen Grundlagenkarten (2013), http://geb.uni-giessen.de/geb/volltexte/2014/10572
7. Schaarschmidt, S.: Bedarfsanalyse zu weiterem Kartenmaterial (2014), http://geb.uni-giessen.de/geb/volltexte/2014/11102
8. Schiller, I., Entrup, B., Binder, F., Schaarschmidt, S., Lobin, H.: Using a GIS for Search and Visualization of Literary Works in the Digital Humanities. In: gis.SCIENCE - Die Zeitschrift für Geoinformatik 4 (to appear) (2014)
9. Seeker, W., Kuhn, J.: Making Ellipses Explicit in Dependency Conversion for a German Treebank. In: LREC. pp. 3132–3139 (2012)