Christian Thomas (BBAW, DTA & CLARIN-D) www.deutschestextarchiv.de, http://clarin-d.net; Mail: thomas@bbaw.de.

Abstract zur 3rd International Chamisso Conference "Worldwide Journeys: Recording, preserving, passing on – Forster, Humboldt, Chamisso" (February 24-26, 2016)

Werke Chamissos, Humboldts, R. & G. Forsters – und vieler weiterer im Deutschen Textarchiv

Hochwertige Volltexte, computerlinguistische Analysen und Tools als Grundlage kooperativer Forschung

Im Vortrag werden Primärquellen, die über das DFG-geförderte Projekt "Deutsches Textarchiv" (DTA, <u>www.deutschestextarchiv.de</u>) der Berlin-Brandenburgischen Akademie der Wissenschaften verfügbar gemacht wurden, und die vielfältigen Möglichkeiten zu deren Nutzung präsentiert. Die Texte des disziplinenübergreifenden, verschiedene Textsorten umfassenden DTA-Kernkorpus wurden nach der frühesten verfügbaren Ausgabe manuell erfasst. Als Ergänzungen wurden hochwertige Volltexte aus externen Quellen kuratiert, in das TEI/XML-basierte DTA-,Basisformat⁴¹ konvertiert und in das Gesamtkorpus integriert.² Alle Quellen stehen frei über das Internet in verschiedenen Formaten zum Herunterladen und auch zum maschinellen "Harvesten" über eine API bereit.³



Titelblätter einiger im DTA enthaltener Werke Georg Forsters, Reinhold Forsters und A. v. Humboldts

Die akkurate Wiedergabe der historischen Quellen in einem einheitlichen, standardisierten Format, die hohe Textqualität⁴ sowie die Zuverlässigkeit der die Objekte beschreibenden Metadaten zeichnen das DTA gegenüber anderen Volltextsammlungen aus. Neben dem Anspruch, vielseitig nutzbare und qualitativ hochwertige Primärquellen frei verfügbar zu machen, liegt der Fokus des DTA-Projekts auf der korpus- bzw. computerlinguistischen Analyse der elektronischen Volltexte. Die automatisierte Erschließung der Volltexte beinhaltet neben der Wortartenbestimmung⁵ und der Lemmatisierung der Wortformen⁶ auch deren orthographische Normierung. Dabei werden historische Schreibweisen auf ihr modernes Äquivalent abgebildet, was schreibweisentolerante Suchanfragen über das gesamte Korpus ermöglicht.

Dies eröffnet weitreichende und in dieser Granularität einzigartige Recherchemöglichkeiten nach einzelnen Stichwörtern, nach festen Wortgruppen, nach gemeinsam in bestimmter oder auch in beliebiger Reihenfolge auftretenden Begriffen. Die Suchanfragen können auf einzelne Texte bezogen, auf das Gesamtkorpus ausgeweitet oder im Vergleich verschiedener Subkorpora vorgenommen werden. Durch die Kombination mit den erfassten Metadaten können die Anfragen zusätzlich fokussiert und beispielsweise auf bestimmte Textsorten (z. B. wissenschaftliche Texte oder Reiseliteratur) und/oder Zeiträume eingeschränkt werden.⁷

1:	[dta:humboldt_gesetz_1806:5]	, welche unmittelbar mit dieser bald	niedrigern	bald höhern Schneegränze zulammen hängen.
2:	[dta:humboldt_gesetz_1806:8]	Zone, macht das Phänomen der	beftändigen	Schneegränze einen analogen, aber minder
3:	[dta:humboldt_gesetz_1806:9]	Erde, die verschiedenen Höhen der	untern	Schneegränze unterfucht hat.
4:	[dta:humboldt_gesetz_1806:11]	Phyfiker aus der richtigen Kenntnifs der	untern	Schneegränze für die Beftimmung der mittlern
5:	[dta:humboldt_bericht_1826:22]	Djebbehl, mit Seemuscheln bei Sanin	nahe	an der Schneegrenze, und mit
6:	(dta:humboldt_ursachen_1830:7)	Continent der reisende Physiker, der	ewigen	Schneegrenze nahe, sein Zelt aufschlägt
7:	[dta:humboldt_ursachen_1830:20]	Die	ewige	Schneegrenze liegt am nördlichen Abhange des
8:	[dta:humboldt_versuche_1837:4]	geringem wissenschaftlichen Interesse, wenn dieselben	weit	über der Schneegrenze liegen, und
9:	[dta:humboldt_versuche_1837:14]	, die man nicht mit der	ewigen	Schneegrenze verwechseln muss.
10:	[dta:humboldt_versuche_1837:14]	Meine Begleiter ritten noch bis zur	perpetuirlichen	Schneegrenze, das ist bis zur
11:	[dta:humboldt_versuche_1837:17]	der Andeskette, dass oberhalb der	ewigen	Schneegrenze weisse Menschen sich in den
12:	[dta:humboldt_versuche_1837:25]	Gipfels hat in dieser Höhe der	ewigen	Schneegrenze, also in 2460 Toisen
13:	[dta:humboldt_versuche_1837:26]	der Provinz Quito die Höhenunterschiede der	ewigen	Schneegrenze an den verschiedenen Nevados,
14:	[dta:humboldt_versuche_1837:31]	dass er am weitesten über die	ewige	Schneegrenze hinausreiche.
15:	[dta:humboldt_geognostisch_1837:13]	Gipfel, welche hoch über die	ewige	Schneegränze hinausreichen, Vulkane, und
16:	[dta:humboldt_geognostisch_1837:20]	einem Felskamme, der über die	ewige	Schneegränze hinausreichte, zu der Höhe
17:	[dta:humboldt_geognostisch_1838:23]	kaum 35 T. hoch über die	ewige	Schneegrenze hinaus, und einige Male
18:	[dta:humboldt_versuch_1842:10]	Streitigkeiten über die relative Höhe der	ewigen	Schneegrenze an dem indischen und tübetanischen
19:	[dta:humboldt_kosmos01_1845:63]	der sich am tübetanischen Abfall die	ewige	Schneegrenze zurückzieht, ist eine gleichzeitige
20:	[dta:humboldt_kosmos01_1845:64]	der Mitte der Hochebene, deren	ewige	Schneegrenze er zu 18000 bis 19000
21:	[dta:humboldt_kosmos01_1845:64]	i. 200 T. höher als die	ewige	Schneegrenze unter dem Aequator in Quito
22:	[dta:humboldt_kosmos01_1845:366]	der Erdoberfläche und die Höhe der	ewigen	Schneegrenze zuerst in Leopolds von Buch
23:	[dta:humboldt_kosmos01_1845:375]	drei andere Phänomene unterscheiden: die	jährliche	Schwankung der Schneegrenze; das Phänomen
24:	[dta:humboldt_kosmos01_1845:375]	Die	untere	Schneegrenze ist aber nicht bloß eine

Suchbeispiel: "Schneegrenze" in Kombination mit einem beliebigen Adjektiv, begrenzt auf Humboldt-Texte, sortiert nach Datum (aufsteigend), KWIC-Ansicht

Auf diesen Analysen aufbauende Tools ermöglichen zudem die Darstellung Wort- oder Dokument-basierter Zusammenhänge in Form einfacher Frequenzlisten (welche Begriffe tauchen am häufigsten auf?), komplexerer Kookkurrenzübersichten (welche Begriffe tauchen häufig gemeinsam mit anderen Begriffen auf?), frequenzbasierten Wortverlaufskurven und 'Tag Clouds' oder auch mehr explorativen, semantisch motivierten Begriffswolken:



Automatisch berechnete, semantisch korrelierende Begriffe im Umfeld "Magnetismus" (Top 50)

Im Vortrag werden diese und weitere Möglichkeiten anhand konkreter, auf das Konferenzthema bezogener Beispiele exemplarisch vorgeführt. Dabei wird vor allem deren Anwendbarkeit für historische – etwa im Bereich der Sprach-, Wissens- oder Wissenschaftsgeschichte – und philologische Forschungen illustriert. Als Beispiel dafür dienen die im DTA verfügbaren Publikationen Alexander von Humboldts.⁸ Dieses 'virtuelle Humboldt-Korpus' setzt sich aus

mehrbändigen Monographien im DTA-Kern- bzw. -Erweiterungskorpus, einer Auswahl von derzeit 164 unselbstständigen Schriften, 21 'Abhandlungen' nach Akademie-Vorträgen Humboldts sowie einem wachsenden Bestand von Hörernachschriften seiner berühmten 'Kosmos-Vorträge' zusammen.

Innerhalb dieses Korpus lässt sich etwa die Beschäftigung Humboldts mit dem Phänomen des (Erd-)Magnetismus nachverfolgen – von den frühesten Veröffentlichungen der 1790er Jahre⁹ über die noch zur Zeit seiner amerikanischen Reise veröffentlichten Beobachtungen in Südamerika¹⁰, seine 1806 in Berlin angestellten Messungen über den Einfluss des Nordlichts auf die Magnetnadel¹¹, die Darstellung des damaligen Wissensstandes im Bereich des Magnetismus in den Kosmos-Vorträgen 1827/28 in der Berliner Universität¹² und vor einem gemischten Publikum in der Singakademie¹³, der Akademievorlesung "Ueber die Mittel, die Ergründung einiger Phänomene des tellurischen Magnetismus zu erleichtern" aus dem darauffolgenden Jahr¹⁴, bis hin zur Behandlung dieses Themenfeldes im fünfbändigen *Kosmos*¹⁵.

Am Beispiel der im DTA verfügbaren Humboldt-Texte wird auch deutlich, dass es nur gemeinsam gelingen kann, die aus Sicht der Forschung in vielen Bereichen und für viele Autoren bzw. Fragestellungen noch zu geringe Zahl der frei im Netz verfügbaren, qualitativ hochwertigen Primärquellen zu erhöhen. Der notwendig kooperative Korpusauf- und -ausbau, die kollaborative Qualitätssicherung und gemeinsame weitere Annotation von Forschungsdaten wird im DTA durch die enge Zusammenarbeit mit mehreren Bibliotheken¹⁶, Dienstleistern¹⁷, Universitäten und Forschungsvorhaben¹⁸ umgesetzt.

Anmerkungen und Referenzen

- Deutsches Textarchiv DTA-Basisformat DTABf, <u>www.deutschestextarchiv.de/doku/basisformat</u> [Alle URLs in diesem Dokument abgerufen am 1.7.2015], siehe auch Susanne Haaf, Alexander Geyken, Frank Wiegand: The DTA "Base Format": A TEI Subset for the Compilation of a Large Reference Corpus of Printed Text from Multiple Sources, In: Journal of the Text Encoding Initiative [Online], Issue 8 PREVIEW | 2014–2015 [http://jtei.revues.org/1114; DOI: 10.4000/jtei.1114].
- ² Umfang des Gesamtkorpus, Stand 18.0.2016: 2825 Dokumente mit ca. 650 000 digitalisierten Seiten, ca. 220 Mio. Tokens, ca. 1,1 Mrd. Zeichen.
- ³ www.deutschestextarchiv.de/api. Die Volltexte sind jeweils in XML, HTML, plain text und TCF verfügbar.
- Im gesamten Gesamtkorpus bietet das DTA eine Zeichengenauigkeit von mehr als 99,95%. Vgl. dazu Susanne Haaf, Frank Wiegand, Alexander Geyken: Measuring the Correctness of Double-Keying: Error Classification and Quality Control in a Large Corpus of TEI-Annotated Historical Text. In: Journal of the Text Encoding Initiative (jTEI) 4, 2013 [http://jtei.revues.org/739; DOI: 10.4000/jtei.739].
- Part of Speech-Tagging (POS), vgl. dazu den Abschnitt Linguistische Analyse historischer Texte (CAB) in der Dokumentation zur Software im Deutschen Textarchiv, http://www.deutschestextarchiv.de/doku/software#cab.
- 6 D. h. jedem (normierten) Wort wird eine Grundform (Lemma) zugeordnet; damit wird gewährleistet, dass eine Suche nach dem Term "beilegen" auch Treffer für die Formen "beygelegt", "beyzulegen", "beilegte" usw. liefert.
- ⁷ Siehe zu weiteren Möglichkeiten die Hilfe zur Suche in den Korpora des Deutschen Textarchivs, http://www.deutschestextarchiv.de/doku/DDC-suche hilfe.

- ⁸ Vgl. die Übersicht der Humboldt-Texte im DTA unter http://www.deutschestextarchiv.de/dtae/avh.
- Vgl. z. B. Humboldt, Alexander von: Neue Entdeckungen [betr. Magnetberg am Fichtelgebirge.] In: Allgemeine Literatur-Zeitung. Intelligenzblatt. Bd. 2 (1797) Nr. 68, Sp. 564-568, www.deutschestextarchiv.de/humboldt neue 1797; Neue Entdeckungen. In: Allgemeine Literatur-Zeitung. Intelligenzblatt. Bd. 3 (1797) Nr. 87, Sp. 722, www.deutschestextarchiv.de/humboldt entdeckungen 1797; Erklärung. In: Allgemeine Literatur-Zeitung. Intelligenzblatt. Nr. 38 (1797) Sp. 323-326, www.deutschestextarchiv.de/humboldt erklaerung 1797.
- Vgl. z. B. Skizze einer Geologischen Schilderung des südlichen Amerika. In: Allgemeine Geographische Ephemeriden. Bd. 9 (1802) St. 4, S. 310-329, www.deutschestextarchiv.de/humboldt-skizze01-1802; Notizen Alex. von Humboldt's von seinen Reisen in der Kordillere der Anden und von seinen physikalischen Beobachtungen in Quito und Mexico. In: Annalen der Physik. Bd. 16 (1804) S. 450-493, www.deutschestextarchiv.de/humboldt-notizen-1804.
- [Gilbert, Ludwig Wilhelm]: Die vollständigste aller bisherigen Beobachtungen über den Einfluss des Nordlichts auf die Magnetnadel; angestellt von Herrn Alexander von Humboldt zu Berlin am 20sten Dec. 1806. In: Annalen der Physik. Bd. 29 (1808) S. 425-429, www.deutschestextarchiv.de/humboldt_nordlicht_1808.
- ¹² Parthey, Gustav: Alexander von Humboldt[:] Vorlesungen über physikalische Geographie. Novmbr. 1827 bis 1828. Nachgeschrieben von G. Partheÿ. [Berlin], www.deutschestextarchiv.de/parthey msgermqu1711 1828; [N. N.]: Alexander von Humboldts Vorlesungen über physikalische Geographie nebst Prolegomenen über die Stellung der Gestirne. Berlin Winter von 1827 bis 1828. [Berlin], [1827/28],www.deutschestextarchiv.de/ nn msgermqu2345 1827; [N. N.]: Die physikalische Geographie von Herrn Alexander v. Humboldt, vorgetragen im Semestre 1827/28. [Berlin], [1827/28],www.deutschestextarchiv.de/ nn oktavgfeo79 1828.
- Vorlesungen über physicalische Geographie von A. v. Humboldt. [G]eschrieben im Sommer 1829 durch Otto Hufeland. [Berlin], [ca. 1829], www.deutschestextarchiv.de/hufeland-privatbesitz 1829; [N. N.]: Physikalische Geographie. Vorgetragen von Alexander von Humboldt. [Berlin], [1828], www.deutschestextarchiv.de/nn-msgermqu2124 1827.
- Humboldt, Alexander von: Ueber die Mittel, die Ergründung einiger Phänomene des tellurischen Magnetismus zu erleichtern. (Auszug aus einer am 2. April 1829 vor der Königlichen Academie der Wissenschaften zu Berlin gehaltenen Vorlesung.) In: Annalen der Physik und Chemie, Bd. 15, St. 3, (1829), S. 319–336, http://www.deutschestextarchiv.de/humboldt_mittel_1829.
- Humboldt, Alexander von: Kosmos. Entwurf einer physischen Weltbeschreibung. Stuttgart u. a.,
 Bde., 1845–62, siehe www.deutschestextarchiv.de/search?q=Kosmos&in=metadata.
- ¹⁶ Die Zusammenarbeit mit den Bibliotheken findet v. a. im Bereich Bilddigitalisierung und ggf. auch dem der Metadatenerfassung statt. Vgl. http://www.deutschestextarchiv.de/doku/kooperationen.
- ¹⁷ In den Bereichen Bilddigitalisierung und Volltexterfassung.
- V. a. in den Bereichen kooperative Edition und tiefere Annotation von Volltexten. Zu nennen sind für die Zusammenstellung des Humboldt-Korpus vor allem die (inzwischen abgeschlossene) Alexander-von-Humboldt-Forschungsstelle der BBAW, in deren Nachfolge das kürzlich begonnene Akademienvorhaben "Alexander von Humboldt auf Reisen" (http://www.bbaw.de/forschung/avh-r), die Humboldt-Universität zu Berlin (Projekt Hidden Kosmos Reconstructing A. v. Humboldt's "Kosmos-Lectures", www.culture.hu-berlin.de/hidden-kosmos) sowie die Universität Potsdam (Lehrstuhl Prof. Dr. Ottmar Ette).