Textsuche und -speicherung für das PopScie-Korpus

Stand: 27.10.2011
Autor: Karin Maksymski

Diese Zusammenfassung beschreibt das Vorgehen bei der Textsuche und beim Abspeichern der Korpustexte für PopScie. Es dient hauptsächlich als Orientierung und Richtlinie für die am Korpusaufbau beteiligten wissenschaftlichen Hilfskräfte und bietet daher nur einen Kurzüberblick. Eine detaillierte Beschreibung des Korpusaufbaus folgt nach Abschluss desselben im Dokument Korpusaufbau.doc.

I. Material und Auswahlkriterien

1. Zeitschriften / Webseiten

- *bild der wissenschaft* (wissenschaft.de) => bdw
- *FAZ* (faz.net) => Ressort "Wissen" => faz
- *GEO* (geo.de) => geo
- *P.M.* (pm-magazin.de) => pm
- *Spektrum der Wissenschaft* (spektrum.de) => sdw
- *Süddeutsche Zeitung* (sueddeutsche.de) => => Ressort "Wissen" => sdz
- *Welt der Wunder* (weltderwunder.de.msn.com) => wdw
- *Wissenschaft Online* (http://www.wissenschaft-online.de/) => wol
- *ZEIT* (zeit.de) => Ressort "Wissen" => zeit
- *ZEIT Wissen* (zeit.de) => Magazin => zw

2. Themen

Grob gesagt wird alles aufgenommen, was mit Naturwissenschaft und Technik zu tun hat. Dazu zählen wir auch Artikel aus den Bereichen Geologie, Psychologie und Medizin. Themenbereiche wie z.B. Geschichte, Archäologie, Sprach- und Literaturwissenschaften etc. werden ausgeschlossen, da diese Bereiche über eigene, besondere Vertextungskonventionen verfügen.

Eine Unterscheidung in Fachbereiche wird insofern vorgenommen, als in den Metadaten zu den einzelnen Artikeln Stichworte zu den Fachbereichen, in die man den Artikel einordnen könnte, aufgenommen werden. Ein Artikel muss also nicht einem konkreten Fachbereich zugeordnet werden, sondern kann auch mit mehreren Stichworten versehen werden.

Die Stichworte, die den Artikeln zugewiesen werden, entsprechen in etwa den klassischen Bezeichnungen der jeweiligen Forschungsbereiche. Eine genauere Unterteilung dieser Bereiche wird nur dort vorgenommen, wo sich Schwerpunkte bei der Themenwahl der Zeitschriften zeigen; dies ist bei Astronomie und Genetik der Fall. Es werden keine Mischdisziplinen als eigene Rubrik aufgenommen; in solchen Fällen erhält der Text einfach beide Stichwörter (z.B. Psychologie und Biologie für einen Text zur Verhaltensforschung, Physik und Chemie für einen Text zur Meteorologie). Es ergibt sich die folgende Stichwortliste (vgl. Tabelle 1):

Stichwort	**Status**	**Themen (Beispielauswahl)**
Astronomie	Teilbereich der **Physik**	
Biologie	eigener Fachbereich	Neuro-/Mikrobiologie, Meeresbiologie, Molekularbiologie

Chemie	eigener Fachbereich	allgemeine Chemie, anorganische und organische Chemie
Genetik	Teilbereich der **Biologie**	Vererbungslehre, Gentechnik
Geologie	eigener Fachbereich	Petrologie, Mineralogie
Mathematik	eigener Fachbereich	Logik
Medizin	eigener Fachbereich	Pharmazie
Physik	eigener Fachbereich	Optik, Kernphysik
Psychologie	eigener Fachbereich	allgemeine Psychologie Neuropsychologie
Technik	eigener Fachbereich	Informatik, Ingenieurwesen

Artikel aus den genannten Teilbereichen und Mischdisziplinen erhalten nicht das Stichwort des / der dazugehörigen "Ober"-Bereiche(s); ein Artikel zu Genetik wird z.B. nicht mit dem Stichwort "Biologie" versehen, auch wenn noch andere Bereiche der Biologie in dem Artikel angesprochen werden.

3. Inhaltliche Kriterien

Durch das Vorgehen bei "Themen" ist auch die klassische Unterscheidung zwischen "harten" und "weichen" Disziplinen aufgehoben. Bei den Artikeln wird stattdessen zwischen "harter" und "weicher" Wissenschaft unterschieden: Bei "harter" Wissenschaft wird durch den Journalisten zusätzlich Grundlagenwissen vermittelt, bei "weicher" Wissenschaft reicht das Weltwissen des Rezipienten, um den Text zu verstehen. Daraus ergeben sich weitere Kriterien für die Textsuche; ein Text wird dann ins Korpus aufgenommen, wenn

a) dort eine neue Entdeckung oder ein sonstiges Phänomen erklärt wird
b) im Text diese Erklärung im Mittelpunkt steht und nicht ein Politikum, eine persönliche Geschichte o.ä.

4. Länge und Zeitraum

Jeder Text (außer den kurzen Online-Artikeln, s.u.) soll mindestens 600 Wörter umfassen. Nach oben hin gibt es zunächst keine Beschränkung.

Zunächst werden Artikel aus dem Zeitraum 2005-2010 berücksichtigt. Sollte dies bei einzelnen Zeitschriften nicht ausreichen, wird der Zeitraum um weitere fünf Jahre verlängert (2000-2005).

5. Autoren

Die Autoren sollten Wissenschaftsjournalisten oder freie Journalisten sein; bei der Auswahl ist darauf zu achten, dass nicht zu viele Artikel von ein und demselben Autor stammen. Nicht berücksichtigt werden Artikel, in denen Wissenschaftler über ihre eigene Forschung berichten (wie das z.B. oft bei *Spektrum der Wissenschaft* der Fall ist); in Ausnahmefällen können sie allerdings in anderer Form als Akteur im Text in Erscheinung treten (z.B. in einer GEO-Reportage). Ebenfalls nicht berücksichtigt werden Artikel, die nicht von deutschen Muttersprachlern stammen oder bei denen es sich um eine bearbeitete Übersetzung handelt (ebenfalls häufig bei *Spektrum der Wissenschaft*).

II. Textspeicherung und gesammelte Daten

1. Metadaten

Die Datei "Ueberblick_Artikel.xls" enthält die folgenden Metadaten für jeden Artikel:

- Dateiname: s. Punkt II 3 (Dateibenennung)
- Titel
- Thema: wo vorhanden, Text des Vorspanns
- Rubrik laut Quelle: Rubrik, unter der der Artikel erschienen ist
- Rubrik: eigene Zuordnung von Schlagwörtern (s. Punkt I 2)
- Autor(en): Verfasser/in des Artikels (inkl. Titel)
- Funktion: Informationen zu Ausbildung und Beruf der Autoren, sofern auffindbar
- Seiten: wenn vorhanden, die Seitenzahlen des Artikels in der Zeitschrift
- Link: Webseite, auf der die erste Seite oder der gesamte Artikel steht
- Wörter: Anzahl der Wörter in der .txt-Datei laut automatischer Zählung in Word
- W-Quantil: Bereich im Wort-Quantil
- Rating, LeseK: Angabe, ob Text ins Ratingkorpus oder ins Lesekorpus aufgenommen wurde
- Anmerkungen: sonstige Anmerkungen zum Text
- Zeitschrift: Zeitung oder Zeitschrift, in der der Text erschienen ist
- Erscheinungsdatum: Monat und Jahr des Erscheinens
- Version: s. Punkt II 2 (Aufmachung)
- gesp. am: Speicherdatum
- von: Bearbeiter/in (DM = Dimitar Molerov, KM = Karin Maksymski, LR = Lisa Rüth, UH = Uli Held)

2. Aufmachung

Es werden Artikel in insgesamt vier Aufmachungsarten elektronisch gespeichert:

- kurze reine Online-Artikel (eigenes Teilkorpus) ; Web-kurz => WK
- Online-Versionen gedruckter Artikel (nicht im Originallayout); Druck-Web => DW
- gedruckte Artikel im Originallayout; Druck-Original => DO
- längere reine Online-Artikel; Web-Lang => WL

3. Dateibenennung

Der Name der Zeitschrift / Zeitung, die Ausgabe und die Aufmachungsart erscheinen bereits als Teil des Dateinamens. Es ergibt sich die folgende Dateibenennung:

Zeitschriftenkürzel_Jahr_Monat_erste Seitenzahl/Ausgabennummer/Tag-Aufmachungsartskürzel

- ⇨ wenn die Seitenzahl nicht ersichtlich ist, stattdessen die Nummer der Ausgabe angeben;
- ⇨ falls es auch dazu keine Informationen gibt (z.B. bei reinen Online-Artikeln), den Tag des Erscheinungsdatums angeben

Bsp.: sdw_2007_02_90-DO

(gedruckter Artikel im Original-Layout aus Spektrum der Wissenschaft, erschienen im Februar 2007 auf den Seiten 90ff.)

zeit_2010_03_12-DW

(Online-Artikel aus dem Ressort "Wissen" der ZEIT, erschienen im März 2010 in der Ausgabe Nr. 12)

4. Format

Jeder Artikel wird zunächst in zwei Formaten gespeichert:

1. als ***PDF*** (Dateiendung: .pdf)
   - ⇨ Bei *Spektrum der Wissenschaft* gibt es PDFs im Originallayout auf der Webseite; von *ZEIT Wissen* haben wir sie aus dem Archiv erhalten. *GEO* hat uns ebenfalls einige PDFs geschickt, allerdings nicht im Originallayout. Bei *bild der wissenschaft*, *P.M.*, der *Zeit*, der *Süddeutschen* und der *FAZ* kann man sich die Online-Versionen bzw. die reinen Online-Artikel als PDF von der Webseite drucken. Bei *Welt der Wunder* und *Wissenschaft Online* müssen wir auf Nur-online-Artikel zurückgreifen.
   - ⇨ Achtung (bei bdw, faz, sdz, zeit): Es gibt manchmal Unterschiede zwischen der Online-Version und dem daraus entstandenen PDF (z.B. werden bei der *ZEIT* auf der Webseite oft Links zu ähnlichen Artikeln angegeben, aber nicht im PDF). Wir nehmen als Ausgangspunkt der Textspeicherung die PDF-Artikel.
   - ⇨ Die Artikel von *GEO* und *bild der wissenschaft* haben wir zum Teil zusätzlich noch gescannt, so dass wir zum einen den Text richtig anordnen können und zum anderen mehr Untersuchungsmaterial für Fragen der Text-Bild-Integration haben.
2. als ***Nur-Text-Datei*** (Dateiendung: .txt)
   - ⇨ Diese Version enthält den kompletten Text, der auch in der PDF-Datei vorhanden ist, d.h. der Text, den man erhält, wenn man sich den Artikel drucken lässt. Dabei kann es zu Unterschieden zur Web-Version der Artikel kommen. So erscheinen z.B. bei der *ZEIT* die Hinweise "Mehr zum Thema" und auch manche Bildbeschriftungen nicht im PDF, werden also auch nicht im Text gespeichert.
   - ⇨ <u>abspeichern</u>: Titel, "Obertitel", "Untertitel", "Vorspann", Autorenangabe (in der Regel nur der Name), Autoreninfo (mehr Infos zum Autor, oft am Ende des Textes), Infokästen u.ä., Haupttext, Bildbeschriftungen, Danksagungen, Literaturhinweise, Verweise auf Audio-Versionen u.ä.
   - ⇨ <u>ignorieren</u>: Angaben zum Veröffentlichungsdatum, Liste der Schlüsselwörter, komplettes Literaturverzeichnis, Grafiken etc., Werbung
   - ⇨ Der Text eines Artikels soll "am Stück" gespeichert werden, also nicht durch Kästen oder Bildbeschriftungen unterbrochen werden. Entsprechend wird Text, der nicht zum Haupttext gehört (d.h. der aus den Infokästen, Bildbeschriftungen, Fußnoten, Verweisen etc.) ans Ende des Dokuments gestellt. Diese "Zusatztexte" sind vom Haupttext dann immer mit einer kurzen Reihe kleiner Trennstriche abgetrennt.
   - ⇨ Wenn eine Tabelle oder Grafik ausschließlich Text enthält, wird dieser Text mit abgespeichert.
   - ⇨ Bei den Fachtexten fällt die Spalte "Funktion" bei den Metadaten weg. Wenn vorhanden, werden Angaben zur ISSN und der DOI gemacht.

Alle weiteren gespeicherten Versionen beinhalten eine Annotation des Textes und sind in der entsprechenden Dokumentation / den Richtlinien näher beschrieben.