Elisabeth Burr
elisabeth.burr@uni-leipzig.de
Universität Leipzig

Julia Burkhardt
jburk@rz.uni-leipzig.de
Universität Leipzig

Elena Potapenko
potep@rz.uni-leipzig.de
Universität Leipzig

Rebecca Sierig
rebecca.sierig@uni-leipzig.de
Universität Leipzig

Arámis Concepción Durán
acduran@uni-leipzig.de
Universität Leipzig

DAS DUISBURG-LEIPZIG KORPUS ROMANISCHER ZEITUNGSSPRACHEN UND SEIN TEXTMODELL

1. EINLEITUNG

Ziehen wir das 574 Seiten starke *Book of Abstracts* der Internationalen ADHO Konferenz *Digital Humanities 2014* vom Juli 2014 in Lausanne heran, dann wirft eine Suche nach *newspaper(s)* / *magazine(s)* / *periodical(s)* eine Vielzahl von Belegstellen aus. Eine Suche nach Zeitung(en) in den Abstracts von *DHd 2014*, der ersten Konferenz des deutschsprachigen Fachverbandes, die im März desselben Jahres in Passau stattgefunden hat, lässt sich leider nicht so komfortabel durchführen, da die Abstracts nicht zu einem digitalen Band zusammengeführt wurden, Zeitungen werden aber zumindest in einem Abstract genannt (cf. Fischer / Kirsten / Witt 2014). Schauen wir uns dann die Projekte, bei denen Digitalisierung und Erschließung von Zeitungen / Magazinen / Zeitschriften eine Rolle spielen, genauer an, so werden wir feststellen, dass es dabei entweder um Kuration oder um Sprachkorpora geht. So versuchen die einen Sammlungen historischer Zeitungen / Zeitschriften / Magazine in Form von digitalen Facsimilies (vgl. *Modernist Magazines Project*) oder hochwertigen digitalen Editionen (vgl. z. B. *The Modernist Journals Project* oder *Die Fa-*

ckel) verfügbar oder große Mengen an Zeitungsseiten durchsuchbar zu machen (vgl. etwa *Europeana Newspapers*). Die anderen nehmen dagegen Zeitungen / Magazine / Zeitschriften in Referenzkorpora auf, die der Untersuchung des schriftlichen Gebrauchs der jeweiligen Sprache dienen (vgl. z. B. *Das Deutsche Referenzkorpus* – DeReKo oder *Coris/Codis*) oder transformieren eine Vielzahl von Ausgaben eines bestimmten Blattes zu einem Zeitungskorpus (vgl. z.B. das *La Repubblica Corpus* oder *The Time Magazine Corpus*). Wird eine Verbindung zwischen beiden Formaten versucht, so beschränkt sich diese in der Regel auf die Integration von mehr oder minder ausgefeilten Werkzeugen, die eine Suche nach sprachlichen Phänomenen (zumeist lexikalischen) in den digitalen Editionen und das Aufrufen der den einzelnen Belegstellen entsprechenden Texte erlauben (vgl. etwa *Die Fakel*, *Der Brenner* oder *The Modernist Journals Project*).

Solche Werkzeuge sind allerdings nicht neutral, sondern es liegt ihnen eine bestimmte Konzeption bzw. Modellierung der jeweiligen Artefakten zugrunde. So spricht Harro Biber etwa von Magazinen als „container of texts“ (Ermolaeva et al. 2014: 52) und als erstes Ziel von *Europeana Newspapers* wird angegeben „Europeana Newspapers [...] will create full-text versions of about 10 million newspaper pages. It will also detect and tag millions of single articles with related metadata and named entities (information identifying people, locations etc.).“ (EurNews 2014). Die *Austrian Academy of Sciences* ist sich zwar bewusst, dass historische Zeitungen, wie der Sozialdemokrat „nicht nur anderer technischer Handhabung [bedürfen], sondern auch anderer Auszeichnungskriterien“ als die „literarischen Texten, die im AAC bearbeitet werden“, beschränkt die Unterschiede aber auf „Textsortenzuordnung, [...] spezifische[...] Annotierung von Pseudonymen, chiffrierten Autoren- und anderen Personennamen“, obwohl die Illustration zu den genannten Kriterien mit einer Zeitungsseite eine viel komplexere Realität abbildet (AAC o.J. c). Konzeptionen wie Textcontainer, Ansammlung individueller Texte oder Datenkonserve liegen auch vielen Korpora zugrunde, die Tageszeitungen als Komponenten aufweisen oder insgesamt aus Tageszeitungen bestehen. Wulfman und Ermolaev kontern solche Konzeptionen zurecht mit: „In order to discuss the rela-

tionship of content elements in a magazine, for example – such as the relationship of advertisements to articles – one must have a common language for expressing layout. Simple text transcription of a magazine's content is insufficient for many kinds of research; thus our ontology is based on an understanding of the historical language of page composition (columns, paragraphs, various forms of headings, publication metadata, and so on) that is vital to the useful encoding of magazine structure and the analysis of a magazine's meaning." (Ermolaev et al. 2014: 53).

In unserem Beitrag werden wir ein Textmodell vorstellen, das auf eine originalgetreue Abbildung nicht nur der Struktur von Tageszeitungen und ihrer einzelnen Seiten zielt, sondern auch der komplexen Beziehungen, die auf einzelnen Seiten oder über ganze Teile der Zeitung hinweg angelegt sind.

2. DAS KORPUS ROMANISCHER ZEITUNGSSPRACHEN

Mit der Erstellung des heute als *Korpus romanischer Zeitungssprachen* vorliegenden Korpus wurde 1989 in Duisburg mit einem allein der Untersuchung der italienischen Zeitungssprache gewidmeten Korpus begonnen (cf. Burr 1993). Mitte der neunziger Jahre des letzten Jahrhunderts kam, wiederum in Duisburg, ein Korpus der italienischen, französischen und spanischen Zeitungssprache hinzu (cf. Burr 1997). Seit 2011 arbeitet eine Projektgruppe am Leipziger Lehrstuhl für französische, frankophone und italienische Sprachwissenschaft an der Erstellung eines Korpus aus französischen, québequer und italienischen Tageszeitungen (cf. Burkhardt et al. 2014).

2.1 Zusammensetzung und Größe des ausgezeichneten Korpus

In das Korpus gehen seit Beginn grundsätzlich ganze Zeitungsausgaben ein. Während die ersten beiden Korpora nach COCOA ausgezeichnet wurden, folgt das Markup des sich derzeit in der Entstehung befindlichen Korpus dem TEI-Standard P5 (cf. TEI Consortium 2014). Das ausgezeichnete Korpus setzt sich aktuell wie folgt zusammen:

- **Italienische Zeitungen - “Die Wende 1989” (724.517 Wortformen)**

Blatt	**Ausgabe**	**Wortformen**
Corriere della Sera	19, 20. und 21.10.1989	258.287
Il Mattino	20. und 21.10.1989	171.501
La Repubblica	20. und 21.10.1989	174.958
La Stampa	20. und 21.10.1989	119.771

- **Französische, italienische und spanische Zeitungen – “Europawahlen 1994” (801.010 Wortformen)**

Blatt	**Ausgabe**	**Wortformen**
Le Monde	12./13., 14. und 15.06.1994	236.236
Corriere della Sera	13., 14. und 15.06.1994	303.641
La Vanguardia	13., 14. und 15.06.1994	261.133

- **Französische, québequer und italienische Zeitungen – “Frauenfußballweltmeisterschaft 2011” (836.977 Wortformen)**

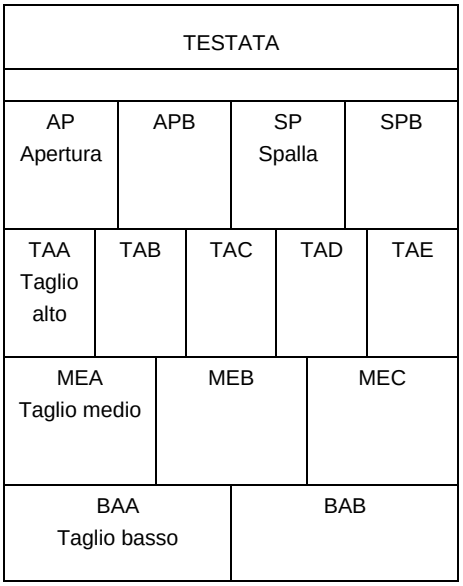
Blatt	**Ausgabe**	**Wortformen**
Le Monde	06. u. 20.07.2011	104.926
Libération	06. u. 20.07.2011	85.520
Le Parisien	20.07.2011	38.894
La Repubblica	06. u. 20.07.2011	260.277
La Stampa	06. u. 20.07.2011	347.360

2.2 Das Textmodell

Das Textmodell, das wie gesagt auf eine originalgetreue Abbildung der textuellen Inhalte und Strukturen der Quellen, also hier der Tageszeitungen zielt, wurde in seiner ersten Form (vgl. das *Korpus italienischer Zeitungssprache*) auf der Basis einer intensiven Auseinandersetzung mit der in den 70er und 80er Jahren nicht nur in der italienischen und deutschen Sprachwissenschaft, sondern gerade auch in der Medienwissenschaft und Publizistik geführten Diskussion um die Tagespresse, ihre Struktur und Sprache sowie ihre interne

Ausdifferenzierung mit Blick auf ein Massenpublikum erarbeitet (cf. Burr 1993: 125-174). Bei dem im Folgenden dargestellten Textmodell handelt es sich also um ein theorie- und forschungsbasiertes und das Medium als solches sowie auch seine Produktion (vgl. das Layoutschema unten) fokusierendes Modell:

Bezug	Kodierung	Beispiel
Zeitung als Fragment des Korpus	**<Z>**	<Z Stampa>
Ausgabe	**<E>**	<E 211089>
Sparte	**<S>**	<S Politica>
Autorenschaft	**<A>**	
unterschieden werden:		
a) signiert		<A firmato>
b) anonym		<A Non firmato>
c) Redaktion		<A Redazione>
Name des Autors	**<N>**	<N Ferrara Giovanni>
Positionierung des Textes	**<C>**	<C MEA01>



Textart	**<T>**	
unterschieden werden:		
a) die Typen von Überschriften		
Vorzeile		<T Occhiello>
Schlagzeile		<T Titolo>
Untertitel		<T Sottotitolo>
Zusammenfassung		<T Sommario>
Zwischenüberschrift		<T Catenaccio>
b) die festen journalistischen Formen		
Leitartikel		<T Fondo>
'Aufmacher'		<T Spalla>

Glosse		<T Corsivo>
Leserbrief		<T Lettera>
Antwort auf eine Leseranfrage		<T Risposta>
Kolumne		<T Rubrica>
Wetterbericht		<T Tempo>
Filmbesprechung		<T Film>
Kurznachricht		<T Breve>
Kurzmeldung		<T Flash>
Agenturmeldung		<T Agenzia>
Ankündigung		<T Riassunto>
c) die fließenden Formen		
Nachricht		<T Notizia>
Artikel		<T Articolo>
Kritik		<T Critica>
Spielbericht		<T Partita>
Interview		<T Intervista>
Darstellungsarten	**<P>**	
fortlaufender Text		<P Prosa>
direkte Rede		<P Discorso>
Zitat von schriftlich Geäußertem		<P Citazione>
Frage des Journalisten (Interview)		<P Domanda>
Antwort des Interviewten		<P Risposta>

(cf. Burr 1993:464-465)

Dieses Textmodell wurde im Zusammenhang mit der Erstellung des zweiten Korpus (vgl. das Korpus *Französische, italienische und spanische Zeitungen*) auf der Grundlage einer ausgiebigen Beschäftigung mit der englischen Korpuslinguistik und ihrer Diskussion um *Corpus Design* und *Sampling frames* weiter begründet und an einzelnen Stellen modifiziert (cf. Burr 1997).

Auch das Textmodell des aktuell im Entstehen begriffenen Korpus wurde unter Heranziehung der Forschung entwickelt. Das es umsetzende TEI Markup sollte nämlich zum einen den Forschungsinteressen, die den in den letzten 20 Jahren erschienenen korpusbasierten Untersuchungen zur französischen Pressesprache zugrunde liegen, gerecht werden, zum anderen den Elementen, die in den dem Zeitungsdesign gewidmeten linguistischen, pressewissenschaftlichen oder kommunikationswissenschaftlichen Arbeiten als konstitutiv für Tageszeitungen betrachtet werden und die ihrerseits den Sprachgebrauch determinieren oder für die Interpretation von Untersuchungsergebnissen bedeutsam sein können: Titelseite, Rubriken und Textsorten, Strukturierungselemente von Artikeln, Gestaltung des Textkörper, mögliche Komponenten des Anlaufs, Illustrationen und Legenden, synoptische Texte, Übersichtstexte und Textcluster (cf. Sierig

2013: 49-72). In dieses soll zu gegebener Zeit auch das COCOA-Markup der beiden früher schon erstellten und online verfügbaren Korpora (cf. Burr 1997-2004) überführt werden.

3. AUSBLICK

Die Entwicklung dieses Textmodells und des es umsetzenden TEI-Markups werden wir in unserem Beitrag begründen und diskutieren. Dass eine einfache Texttranskription nicht ausreicht und wir stattdessen Textmodelle brauchen, die der komplexen Struktur von Zeitungen und ähnlichen Artefakten Rechnung tragen, wenn wir Sprache tatsächlich in ihrem Gebrauch, das Wissen, das bei den Rezipierenden vorausgesetzt wird, oder die Semantik solcher Artefakten untersuchen wollen, können allein schon die folgenden Sierig (2013) entnommenen Beispiele zeigen:

Bausteine und Ebenen der *Une*

VOIX EXPRESS

Quel est votre pire souvenir avec un voisin de bureau ?

Alain Guichaoua
52 ans, chef de projet
Penmarc'h (29)

« Je partageais mon bureau avec un gros fumeur. Il était à plus d'un paquet par jour, alors que moi, je n'ai jamais fumé. A la maison, je demande à ma femme de fumer dehors. On avait un passé un accord avec ce collègue pour ouvrir la fenêtre. Mais il était frileux... Cela a duré deux-trois ans. A la fin nous avons eu des bureaux séparés. Et lorsque la loi antitabac a été votée, j'étais très heureux ! »

Nathalie Dupond
32 ans, en recherche d'emploi
Meaux (77)

« J'ai été victime de harcèlement moral de la part d'une collègue qui refusait de me former alors que je devais la remplacer pendant ses congés. Elle m'a accusé de faits que je n'avais pas commis. Elle était jalouse et se sentait menacée. Elle ne voulait pas qu'on montre qu'on bossait, alors qu'elle ne faisait rien. Tout le monde était pétrifié car elle avait plus de vingt-cinq ans de boîte. J'ai fini par partir. »

Marc Dumas
55 ans, chef d'entreprise
Versailles (78)

« J'avais une stagiaire qui sentait mauvais. Elle transpirait beaucoup et nous étions trois dans un petit bureau. Je l'adorais et je voulais l'embaucher. J'ai donc demandé à sa tutrice de stage de lui faire comprendre le problème en douceur. Mais elle lui a dit que la remarque venait de moi. C'était très gênant. Finalement, je l'ai embauchée et on a travaillé ensemble pendant quinze ans. »

Christiane Frazier
53 ans, assistante de direction
Aubervilliers (93)

« Je travaillais à côté de quelqu'un qui parlait tout le temps. Cela ne s'arrêtait jamais. C'était non-stop. Elle nous racontait sa vie, même si on cela ne nous intéressait pas. Il devenait impossible de se concentrer. Parfois, je perdais patience et je lui demandais d'arrêter de me parler, mais elle changeait d'interlocuteur. Comme on était dans un espace sans cloisons en open space... »

synoptische Texte

Le fait du jour — Aujourd'hui en France / Samedi 26 octobre 2013 — 02

Le fait du jour — Aujourd'hui en France / Samedi 26 octobre 2013 — 03

Le gouvernement face à la colère des épargnants

Votée en première lecture, une hausse de la taxation de certains produits d'épargne provoque la polémique.

L'ÉDITO — Mauvaise manière

Où peut-on placer son épargne

L'opposition lorgne vers le Conseil constitutionnel

VOIX EXPRESS — Cette nouvelle taxation va-t-elle vous pousser à épargner autrement ?

La réforme en trois cas concrets — PEL — PEA — Assurance vie — Ancien régime — Nouveau régime

« Il faut davantage de stabilité »

« Je ne suis pas un nanti »

Affolement dans les banques

Text-Cluster

Bei der Begründung und Diskussion des Textmodells sollte auch deutlich werden, dass die Zeitungen nicht als Textcontainer, Ansammlungen individueller Texte oder Datenkonserven verstanden werden, sondern das Medium mit samt seinen komplexen Strukturen und internen wie externen Relationen als Artefakt betrachtet wird und das Korpus romanischer Zeitungssprachen deshalb letztendlich dem Humanities Computing / den Digitalen Humanities näher steht als der Korpuslinguistik computerlinguistischer Prägung.

AUSGEWÄHLTE QUELLEN

[AAC] = Austrian Academy Corpus (o.J. a): *Der Brenner* <http://corpus1.aac.ac.at/brenner/> [08.11.2014].

[AAC] = Austrian Academy Corpus (o.J. b): *Die Fakel* <http://corpus1.aac.ac.at/fackel/> [08.11.2014].

[AAC] = Austrian Academy Corpus (o.J. c): „Der Sozialdemokrat", in: *Austrian Academy Corpus* <http://www.aac.ac.at/apps_digied_sozial.html> [08.11.2014].

AA.VV. (2014): *Book of Abstracts*. Digital Humanities 2014. Lausanne <http://dh2014.org> [08.11.2014].

Brooker, Peter / Thacker, Andrew (2006-2014): *Modernist Magazines Project* <http://www.modernistmagazines.com/about.php> [08.11.2014].

Brown University / The University of Tulsa (o.J): *The Modernist Journals Project* <http://modjourn.org/> [08.11.2014].

Burkhardt, Julia / Concepción Durán, Aramis / Potapenko, Elena / Sierig, Rebecca (2014): „FrItZ: Le corpus de la langue des journaux français et italiens de Leipzig – Développement et possibilité d'application d'un corpus de la presse écrite", 9. Frankoromanistenkongress *Schnittstellen / Interfaces*. Sektion *Les interfaces numériques* (Elisabeth Burr / Christoph Schöch). 24. bis 27. September 2014, Universität Münster.

Burr, Elisabeth (1993): *Verb und Varietät*. Ein Beitrag zur Bestimmung der sprachlichen Variation am Beispiel der italienischen Zeitungssprache (= Romanische Texte und Studien 5). Hildesheim: Olms.

Burr, Elisabeth (1997): *Wiederholte Rede und idiomatische Kompetenz*. Französisch, Italienisch, Spanisch. Habilitationsschrift, Gerhard-Mercator-Universität GH Duisburg, Fachbereich 3: Sprach- und Literaturwissenschaften (Manuskript 429 Seiten).

Burr, Elisabeth (1997-2004): *Korpus Romanischer Zeitungssprachen*. Duisburg / Bremen / Leipzig <http://www.uni-leipzig.de/~burr/CorpusLing/> [09.11.2014].

Ermolaev, Natalia / Wulfman, Clifford E. / Biber, Hanno / Crombez, Thomas (2014): „Remediating 20th-Century Magazines of the Arts: Approaches, Methods, Possibilities", in: AA.VV.: Book of Abstracts. Digital Humanities 2014. Lausanne: 52-55 <http://dh2014.org> [08.11.2014].

Davies, Mark (2007-): *TIME Magazine Corpus: 100 million words, 1920s-2000s* <http://corpus.byu.edu/time/> [08.11.2014].

[EurNews] = Staatsbibliothek zu Berlin – Preußischer Kulturbesitz (2014): *Europeana Newspapers* <http://www.europeana-newspapers.eu/> [08.11.2014].

Fischer, Jens / Kirsten, Jens / Witt, Andreas (2014): „Aufbau eines Korpus zur Beobachtung des Schreibgebrauchs im Deutschen", in: *DHd 2014*. Universität Passau <https://www.conftool.pro/dhd2014/index.php/Fischer-Aufbau

_eines_Korpus_zur_Beobachtung_des_Schreibgebrauchs-2711152.pdf> [08.11.2014].

Institut für Deutsche Sprache (o. J.): *Das deutsche Referenzkorpus – DeReKo* <http://www1.ids-mannheim.de/kl/projekte/korpora/> [10.11.2014].

Rossini Favretti, Rema / Grandi, Nicola / Nissim, Malvina / Tamburini, Fabio / Gagliardi, Gloris (1998-): *Coris / Codis*. Università di Bologna <http://corpora.dslo.unibo.it/coris_ita.html> [08.11.2014].

Sierig, Rebecca (2013): *Die Erstellung eines Zeitungskorpus*. Sampling und Markup. Leipzig: unveröffentlichte Masterarbeit.

SSLMIT (2004): *La Repubblica Corpus*. Università di Bologna <http://dev.sslmit.unibo.it/corpora/corpus.php?path=&name=Repubblica> [08.11.2014].

TEI Consortium (16.09.2014): *TEI P5: Guidelines for Electronic Text Encoding and Interchange* 2.7.0. <http://www.tei-c.org/release/doc/tei-p5-doc/en/Guidelines.pdf> [10.11.2014].

Zu den für die Erstellung des Textmodells relevanten Quellen vgl. Burr (1993 u.1997) sowie Sierig (2013).

1486 Wörter (ohne Bibliographie und Autor_innen)