Natural Language Processing for Cultural Heritage Domains

Caroline Sporleder
Saarland University

Abstract

Museums, archives, libraries and other cultural heritage institutes maintain large collections of artefacts which are valuable knowledge sources for both experts and interested lay persons. Recently, more and more cultural heritage institutes have started to digitise their collections, for instance to make them accessible via web portals. However, while digitisation is a necessary first step towards improved information access, to fully unlock the knowledge contained in these collections, users have to be able to easily browse, search and query these collections. This requires cleaning, linking and enriching the data, a process that is often too time-consuming to be performed manually. Information technology can help with (partially) automating this task. Since data processing and enrichment typically involve the textual metadata level, natural language processing has a key role to play in this endeavour. At the same time cultural heritage domains pose significant challenges for language technology and call for the development of very robust and flexible solutions. Consequently, cultural heritage data can also serve as a good test-bed for the development of robust natural language processing tools.

1 Introduction

The term 'cultural heritage' (CH) refers broadly to any item relating to the culture or history of a group of people. It includes monuments and historical sites, artefacts, manuscripts, and even abstract phenomena with cultural significance, such as folk tales or the language used by a society. Under a wider definition the term is also applied to natural heritage, which encompasses the natural environment including all aspects of biodiversity.

Determining what is worth preserving is the task of CH institutes such as museums, libraries, or archives. However, as mere assemblages of objects the collections housed by such institutes are of limited value. Without additional information, e.g., on where and when an artefact was found or created, its significance is reduced to its material or aesthetic value. It is these *metadata* which enable researchers to place artefacts in their right context and thereby shed some light on the history and culture of a society (see Section 2.1).

Collection management was traditionally a pen-and-paper business, with artefacts being catalogued in (often hand-written) register books or on file card systems. Finding information was a rather laborious process which involved going to the institute's archive and ploughing through thick volumes of catalogues. It also often required expert knowledge of the collection and its organisation. This has started to change. More and more CH institutes have begun to digitise their collections and the available metadata, often as part of larger-scale national or even multi-national initiatives. Two prominent examples are Europeana[1] and CultureSampo[2] (Hyvönen et al., 2007).

One motivation for digitisation is that it can be a safeguard against data loss. If the original artefacts or the associated metadata are destroyed, digital copies can ensure that at least some knowledge is saved. That even objects housed in CH institutes are not completely safe from damage or destruction is evidenced by recent disasters such as the 2004 fire in the Anna Amalia Library in Weimar, Germany, which is thought to have destroyed around 50,000 books and manuscripts, or the collapse of the City Archive of Cologne, Germany, in 2009, which buried more than 65,000 documents, partly dating back to Roman times, under a heap of rubble. Digitisation could have ensured that the content of these documents was preserved. It would also have made restoration considerably easier, with conservators now facing the daunting task of reassembling documents that have been torn into hundreds of pieces. However, to ensure data longevity, digitisation needs to be complemented by careful data management (see Section 2.2).

The main reason for most large-scale digitisation projects, however, is the fact that governments worldwide have come to view cultural heritage as a valuable asset, both ideationally and economically. Cultural heritage is considered important for national and cultural identity. It also stimulates tourism and attracts visitors. Digitisation allows cultural heritage institutes to provide continuous, simultaneous and world-wide access to their collections, including those objects which are usually not on public display due to a lack of exhibition space or due to their fragility (Battro, 2010).[3]. Having an engaging web interface which provides a preview of the collection helps to reach out to new visitor groups. Scientific collaboration also stands to benefit from digitisation as it is much easier to share data across institutes. Fragile objects no longer have to be sent out to interested researchers but can be viewed instantly via the internet. Moreover, digitisation helps researchers and curators at the cultural heritage institute themselves, as information can be found more quickly and the curation process can be partly automated (Alex et al., 2008; Karamanis et al., 2007). Finally, information technology (IT) can help museum staff with their research by automat-

[1] `http://www.europeana.eu/portal/`

[2] `http://www.kulttuurisampo.fi/?lang=en`

[3] It is estimated that only 2-5% of objects are on display in public exhibitions, the vast majority of artefacts is stored in depots (Fabrikant, 2009)

ically organising and presenting information in a way that makes it easier to spot interesting interdependencies and discover new knowledge (see Section 2.3).

Digitisation is only a necessary first step, however. To enable intelligent search and navigation, the digitised data need to be processed and enriched. For instance, related entities from different collections, such as a coin from Chios and an old map of the island, should be linked. Likewise, information retrieval should not be restricted to simple key-word based search but allow for more complex queries on semantic representations (see Section 2.1). Since this enrichment process tends to take place on the (typically textual) metadata level, natural language processing (NLP) has a large role to play (see Section 4).

Information access has long been an active research topic in NLP but researchers have only recently started to look at CH domains, which provide considerable challenges (see Section 2.4), for instance:

- language that is archaic and not standardised
- noise introduced during the digitisation process
- content that is multi-lingual, multimodal and multi-medial
- few natural language processing tools for this domain and limited availability of resources to develop new tools from scratch
- large variety of data formats and knowledge representation standards

There is a growing awareness of the need to develop IT and NLP solutions for CH domains, as witnessed by conferences and workshops such as *Museums and the Web*,[4] the *International Cultural Heritage Meetings (ICHIM)*[5] or *Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH)*,[6] dedicated journals such as the *ACM Journal on Computing and Cultural Heritage (JOCCH)*[7] and international networks, such as *CLARIN*.[8] This article will provide an overview of areas in which language technology has been successfully applied to CH data and of the challenges one needs to be aware of when developing such technology.

2 The Situation at the Cultural Heritage Institutes

To better understand how language technology can be applied to the cultural heritage domain it is first useful to look in more detail at the potential users of this technology. This section discusses the most commonly encountered data sources in CH institutes (Section 2.1),

[4] http://www.archimuse.com/conferences/mw.html
[5] http://www.archimuse.com/conferences/ichim.html
[6] http://ilk.uvt.nl/LaTeCH2010/
[7] http://jocch.acm.org/
[8] http://www.clarin.eu/

the digitisation and data management process (Section 2.2), the user's perspective (Section 2.3) and the typical challenges and demands (Section 2.4).

2.1 Primary Data and Metadata

In many disciplines it is customary to distinguish between *primary data*, one the one hand, and *secondary* or *meta*data, on the other. In practice, this distinction is not always entirely clear-cut; the same data source can be viewed either as primary data or metadata, depending on the context. For example, a history book would normally be considered metadata, providing second hand information about a particular time in history. However, Thukydides' *History of the Peloponnesian War* would nowadays be more likely to be viewed as primary data, which is analysed by historians to find out more about a specific period in ancient Greek history, just like excavated artefacts are analysed by archaeologists. In the context of cultural heritage collections the distinction is easier to make: primary data are the objects in the collection, be they archaeological excavates, manuscripts, audio-recordings, or books in a library; metadata are everything else. Under a narrow definition the term 'metadata' pertains to archival information that provides *contextual information* about an object in the collection. Library catalogues, field books, or transcripts of archived audio-recordings fall in this category. In a wider sense the term 'metadata' can also be applied to other data sources that are related to (or derived from) primary data, for example, a research paper on specimens from a natural history collection.

Another distinction that can be drawn is between metadata supplied by humans and metadata generated by machines. Traditionally, metadata have been created by and for humans (library catalogues, register books etc.). These metadata are nearly always textual. Since natural language is highly ambiguous, this causes problems if computer programs have to interpret the data (see Section 2.4). Hence, manually created metadata are often automatically enriched with a second metadata layer which disambiguates and structures the original metadata, e.g., by providing information about word-senses or co-references between named entities (see Section 4.2). Sometimes metadata are automatically generated directly from primary data, for instance in the case of automatic speech recognition applied to recorded interviews (Byrne et al., 2004).

Metadata are typically textual, primary data, on the other hand, come in a variety of formats, though they, too, can be textual (e.g., old manuscripts). While textual primary data often pose even greater problems for automatic analysis than textual metadata, e.g., in terms of archaic language, non-standard orthography or the lack of explicit structure in the document, the fundamental challenges are the same. Hence, in this paper I will not always distinguish between processing textual primary data and textual metadata.

Typical metadata sources in CH institutes include:

Registers or Catalogues contain basic information about an object, for example the author, title, and publication year of a book in a library. They also assign a unique identifier to each object, such as a registration number or barcode, and encode information about where the object can be found, e.g., a shelf number. Cataloguing systems are a major research focus in library science and libraries tend to use fairly elaborate and largely standardised cataloging systems. Museums, on the other hand, often employ more ad hoc systems, partly due to the fact that their catalogues are only meant to be consulted by experts and partly due to the heterogeneity of their collections, which makes standardisation more difficult (Chenhall and Vance, 2010).

Field Books or Field Notes are common in those cultural heritage institutes whose collections are largely assembled during expeditions or excavations, such as archaeological or natural history museums. Field books are semi-structured texts, written during an expedition, which record, often in note form, detailed information on the circumstances of the collection of an object. They tend to be significantly more detailed than catalogues.

Databases are manually created by curators from field books. They contain information in a more structured and thus better searchable form. As database population is a rather time-consuming task, databases typically only cover small parts of a collection. However, the task can be automated to some extent (see Section 4.1). Because cultural heritage databases are normally extended on an ad hoc basis, they are often not optimally structured. During their life span, new columns might be added, e.g., because it suddenly becomes obvious that a certain piece of information needs to be included, and old columns may also fall into disuse, e.g., because the database passes from one curator to another who does not fully understand the motivation for (or meaning of) that column. This can lead to errors and inconsistencies which can significantly reduce the usability of a database (Chapman, 2005). Error detection and correction in databases is therefore an active research area (see Section 3.2). It is also common for curators at the same institute to use different database formats.

To address these problems, many cultural heritage institutes are now moving to a unified database management, i.e., using one database system throughout the institute, possibly with slightly different parametrisations according to the demands of different subcollections. However, integrating existing resources into the new system is non-trivial and leads to problems such as detecting and merging near-duplicate entries (Hernández and Stolfo, 1998; Doerr et al., 2004) or performing co-reference resolution and name disambiguation (Chen and Martin, 2007; Mann and Yarowsky, 2003; Bagga, 1998).

2.2 Digitisation and Data Management

Digitisation of textual data is usually done by taking high quality digital photographs of the original manuscripts and then applying optical character recognition (OCR), which converts digital images to computer-readable text. For hand-written documents, OCR is still a largely unsolved problem, with average word-error rates of 50% and more (Rath et al., 2004). Manual transcription is expensive and therefore usually not a viable alternative. To make handwritten documents searchable even if no transcripts are available, information retrieval techniques have been developed that work directly on the digital image, e.g., by matching visual features (Zinger et al., 2007; Srihari et al., 2005; Rath et al., 2004). Typed documents, on the other hand, can in principle be OCRed but historic documents are difficult to deal with. Problems can, for example, arise from the use of old-fashioned scripts (e.g., Fraktur) or the document quality (stains, faded letters). OCRed documents therefore typically need to be post-processed to correct errors (see Section 3.1). Even when digitisation is technically unproblematic it can be relatively time-consuming and costly. Old documents are fragile and thus have to be scanned manually page by page.

Digitisation on its own is also not sufficient to ensure data longevity. Traditional media are generally much more durable than digital media. For example, while some Egyptian papyri have survived for more than two-thousand years, the life-span of an average CD-Rom is only two to five years due to physical decay (Horlings, 2003). Apart from this so-called 'bit-rot', there are a number of other reasons that can render digital data unusable: outdated media which cannot be read by modern devices (e.g., floppy disks), outdated formats (e.g., proprietory word processing formats that are no longer in use) or a loss of context (e.g., lost decryption keys) (Baker et al., 2006). Prominent examples of digital data loss include data pertaining to NASA's 1976 Viking Mars mission. The data were stored on magnetic tape which was later found to have deteriorated so much that a significant amount of data were not usable anymore (Besser, 2000). Likewise a large amount of multi-media data collected by the BBC's Domesday Project[9] in the mid-1980s, was found to be virtually unreadable fifteen years later due to an outdated storage format.

Digital data preservation is still a challenging problem. However, a number of initiatives have been launched to define standards and good practice. Examples are the *National Digital Information Infrastructure and Preservation Program*[10] of the Library of Congress or the *Kopal* project.[11] Strategies against digital data loss include: adherence to open standards such as XML, *refreshing* (i.e., periodically copying data to new storage media, which are ideally kept in different locations), *migration* (of data to newer storage formats), and *emu-*

[9] `http://www.domesday.org.uk/`

[10] `http://www.digitalpreservation.gov/`

[11] `http://kopal.langzeitarchivierung.de/index.php.en`

lation (of outdated software on modern hardware)[12] (Besser, 2000; Broeder et al., 2008).[13] However, digital data management is expensive and therefore often not systematically implemented by CH institutes.

2.3 The User's Perspective

Information technology can be used in two different ways in CH institutes: first, it can be employed to preprocess the data (cleaning, enriching, structuring) and thereby make the data more accessible; second, it can be used in the information access process itself.

Preprocessing often has to be done semi-automatically, e.g., to correct errors (see Section 3), provide metadata descriptions (Section 4.2) or convert (semi-)structured texts into searchable databases (Section 4.1). This can require considerable domain expertise. The users are thus typically the CH experts themselves, possibly supported by IT experts. It has been shown that NLP technology can significantly reduce the workload of curators when preparing the data (see e.g. Alex et al., 2008; Karamanis et al., 2007; Malaisé et al., 2006). However, as it is sometimes not clear how a system should interact with CH experts in order to provide the best possible support, it is advisable to conduct user studies. This is especially true since staff at CH institutes are often reluctant to apply language technology. There is a general feeling that information technology is not mature enough to produce reliable and useful results. Moreover, curators are frequently concerned that their data will be irreversibly altered. Both points are important and have to be taken into consideration. Software developers need to ensure data provenance, i.e., alterations should always be additive, making it possible to roll-back to earlier stages of the data. In addition, acceptance of language technology can be increased by giving users full control of the system, i.e., by employing a semi-automatic set-up. It often also helps if the system provides additional information on why a particular decision was taken (e.g., why two records were linked or why a piece of information was classified as erroneous).

Information access is relevant for both the general public and CH experts. CH researchers who work with a collection on a daily basis can potentially benefit enormously from improved data access and visualisation but relatively few studies have looked into this aspect. However, studies in other areas demonstrate how experts can be supported in their research activities, for example in the context of recommender systems (Bogers and van den

[12] The Domesday Project data, for example, were recovered by emulation as part of the *CHAMILEON* project: `http://www.si.umich.edu/CAMILEON/`

[13] See also the *Digital Preservation Management Tutorial* hosted by *the Inter-university Consortium for Political and Social Research (ICPSR)*: `http://www.icpsr.umich.edu/dpm/dpm-eng/eng_index.html`.

Bosch, 2008; McNee, 2006). Most work on improved information access to CH collections focuses on the general public, though. There is a long tradition of work on information retrieval, both of textual and audio-visual data. For studies that deal specifically with CH data see, e.g., Hollink et al. (2009); Olsson and Oard (2007); Byrne et al. (2004). There is also a large body of work on personalisation in the CH domain. Several researchers have developed systems that recommend objects to users based on their previous browsing history. Wang et al. (2009) and Ruotsalo and Hyvönen (2007), for instance, identify semantic relations between metadata terms that can be used to make meaningful recommendations to a user. There are also systems that are able to generate personalised descriptions for different target audiences (children, adults, experts) (Konstantopoulos et al., 2009; Isard et al., 2003).

For space reasons, I will mainly discuss the first usage scenario, using language technology to preprocess CH data. Preprocessing has to address a number of challenges that are specific to CH domains (see Section 2.4), while the second scenario is either relatively domain independent (e.g., information retrieval) or is only tangentially related to NLP (e.g., website accessibility). One exception is historic document retrieval, which is discussed in Section 4.2.

2.4 Challenges for Linguistic Processing

Linguistic Imprecision and Ambiguity While textual data and metadata provide useful information to humans, they are often difficult to interpret by a machine due to the fact that natural language is both highly ambiguous and frequently too underspecified. This may lead to suboptimal results when searching for information. *False positive* or *precision* errors arise when information is retrieved that is not relevant to a query, while *false negative* or *recall* errors refer to a failure to find all relevant information. Lexical ambiguity in the form of synonyms or homonyms is one reason for suboptimal results. For example, the surname *Breugel* can point to several different painters, while the species *Leptophis ahaetulla* is also known under the name *Dendrophis liocercus*. Hence, there is a many-to-many mapping between linguistic form and entities or concepts in the real world. Noise in the data such as spelling errors also has a negative effect. Precision and recall errors can be avoided if the data are enriched with meta-information on synonyms, homonyms and corrected spellings.

Missing structural information can also hamper search. For example, searching for all specimens of the Lesser Tree Frog (*Hyla minuta*) in a digitised natural history field book may wrongly return the entry in Example (1) which actually describes a snake (*Leptophis ahaetulla*). To avoid the retrieval of such entries, it is necessary to know that *Hyla minuta* is part of a side remark and does not specify the genus and species of the specimen described

in this entry. Explicitly structuring the available information via a second metadata layer (see Example (2)) permits more sophisticated queries and avoids retrieval errors. Curators at cultural heritage institutes are well aware of this and manually created databases are one attempt to provide additional structural information.

(1) Leptophis ahaetulla, road to Overtoom, in bush above water in the process of eating Hyla minuta 20-VI-1972, RMNH 38290.

(2) [Leptophis]$_{Genus}$ [ahaetulla]$_{Species}$, [road to Overtoom]$_{Location}$, [in bush above water]$_{Biotope}$ [in the process of eating Hyla minuta]$_{Remarks}$ [20-VI-1972]$_{DateOfCollection}$, [RMNH 38290]$_{ID}$.

To address the problem of ambiguity, some cultural heritage institutes restrict metadata, especially keywords, to a well defined subset of terms. Using such *controlled vocabularies* for metadata annotation has the advantage of enforcing a one-to-one mapping between form and meaning. Controlled vocabularies are often organised hierarchically into a thesaurus or knowledge base which can aid automatic query expansion. Enforcing the use of such vocabularies for data indexing and querying can improve retrieval performance, especially if the querying is done by users who are familiar with the indexing scheme. However, there are also situations in which free-text searching leads to better results (Svenonius, 1986).

Over the years a number of metadata standards and schemes have evolved (National Information Standards Organization, 2004). One example is the Dublin Core Metadata Element Set, which was originally geared towards web resources but has since been adapted to a number of domains, including cultural heritage. Other standards for the cultural heritage domain include the MIDAS Heritage standard (Lee, 1998) and the CIDOC Conceptual Reference Model (CIDOC-CRM) (Crofts et al., 2009). Some cultural heritage institutes are starting to adopt these standards, while others make use of ontologies or controlled vocabularies developed in-house or do not adhere to rigid standards at all.

Metadata requirements differ for domains such as arts and natural history. As a consequence a plethora of different, domain-specific standards have developed, hampering *semantic interoperability*, i.e., the easy exchange of data across institutes. A particular problem are different name and concept spaces. To alleviate this problem, some institutes have looked into the use of Semantic Web technology (Berners-Lee et al., 2001). The standard metadata representation model of the semantic web, the Resource Description Framework (RDF), provides support for the integration of diverse name spaces. However, this requires automatic alignment of different schemes. For some schemes, manually defined mapping rules (so called *crosswalks*) are available (National Information Standards Organization, 2004). To align different vocabularies, automatic methods using ontology matching have been proposed (Isaac et al., 2008).

Domain Adaptation and Portability Some basic linguistic preprocessing (such as sentence splitting, part-of-speech tagging, or syntactic parsing) is usually necessary before applying more sophisticated enrichment or text mining tools. While tools for these linguistic preprocessing tasks exist for the major languages, they are normally based on models that were trained on manually annotated texts from one specific domain and genre, usually news wire. These tools often do not work well when applied to texts from a different domain, e.g., archaeological field reports, as domain changes can come with significant variations in vocabulary and lexical and syntactic structures (Gildea, 2001; Roland and Jurafsky, 1998). Another problem are differences in linguistic register. Consider the three excerpts from a natural history field book shown in (3). Field books rarely contain full sentences, the text is usually in note form (e.g., *Mother from Paramaribo*, example (3a) below). There may also be add-ons and comments in brackets (*e-mail to T. M. Pauls January 2004*, example (3b)), ungrammatical chunks (*according information*, example (3b)) or unusual abbreviations (*ad.* for *adult*, example (3c)).

(3) a. Mother from Paramaribo (?). Laid egg in captivity. Hatched 03-07-1969. Offspring died 20-07-1969.

b. according information from P. Taylor, NHM, Washington, this is likely to be L. knudseni, considering the short dorso-lateral folds and chest spines (e-mail to T. M. Pauls January 2004)

c. 1 halfgrown ad.

In some fields, such as biomedicine, this problem is addressed by manually annotating relatively large amounts of text in the new target domain and then re-training the model (Tsuruoka et al., 2005). However, this is expensive and typically not financially viable for CH institutes. Fortunately, there is a growing body of work on domain adaptation (Rimell and Clark, 2008; Daumé III, 2007; McClosky et al., 2006).

Older Language Varieties A further problem for linguistic processing arises if the data are written in an older language variety. This is more a problem for primary data, though some metadata sources are also old enough to contain, for example, out-dated orthography.[14] Pennacchiotti and Zanzotto (2008) show that on these data, the performance of modern NLP tools can drop noticeably. To address this problem, some researchers have used manually created rules to re-tune a tool to older language varieties (Borin and Forsberg, 2008; Rocio et al., 1999) and others have re-trained on a small amount of manually labelled data from the target variety (Rögnvaldsson and Helgadóttir, 2008). An interesting

[14] A related problem is that 'old' metadata may contain references to place names which are no longer in use, such as *Rhodesia* for *Zimbabwe*.

alternative is proposed by Moon and Baldridge (2007) who model the task as a cross-lingual projection problem.

3 Data Cleaning

An important first step in processing CH data involves data cleaning. Error-prone digitisation techniques often introduce noise. Manually created databases can also be noisy due to the fact that they grow organically (see Section 2.1).

3.1 Error Correction and Normalisation in Text

Given that most digitisation efforts involve OCR, it is not surprising that a big challenge in data cleaning involves the detection and correction of OCR errors. While the performance of state-of-the-art OCR systems reach word accuracies of up to 99% (Lopresti, 2005), i.e., one mis-spelled word in a hundred, the error rate is still much higher than for manually typed text. For comparison, Reynaert (2005) found only one non-word error for 400 words of text in the typed Reuters RCV corpus (Lewis et al., 2004). Moreover, high accuracies are only obtainable in favourable circumstances, i.e., for high-quality scans and modern fonts. The error rate for OCR-scans from historical documents is likely to be significantly higher.

OCR errors can lead to significant problems further down the line. First, they have a potentially negative effect on keyword-based information retrieval, as erroneous forms will not be found, thus lowering recall. Second, OCR errors hinder further processing. A particular problem is that OCR errors, unlike typos, often affect word segmentation, e.g., by replacing a letter with a space or punctuation mark, which has a detrimental effect on tokenisation. Missing full-stops are also problematic because they affect sentence splitting. Both erroneous tokenisation and sentence splitting can have serious knock-on effects (Lopresti, 2008, 2005).

OCR errors also often deviate more from the correct form in terms of edit distance (Levenshtein, 1965) than typos made by humans. In particular letter-to-letter mappings between the correct and the erroneous form are less likely to be one-to-one. For instance, the sequence ‘iii’ can be misread as the single letter ‘m’ and vice versa (Taghva and Stofsky, 2001). Errors are also much more likely to affect multiple positions in a word, e.g., when the letter ‘e’ is systematically misread as ‘c’ (Reynaert, 2008). These properties of OCR errors make it difficult to directly apply standard spell checkers, particularly because these tend to rely on correct word boundaries. A number of studies have focussed specifically on error detection in OCRed text. Taghva and Stofsky (2001), for instance, propose an interactive system that incorporates multiple knowledge sources such as domain-specific lexicons, user feedback or mapping rules derived from a gold standard of aligned pairs of

error and correct form. Kolak and Resnik (2005) propose a noisy channel model that does not require a lexicon or human interaction but also needs to be trained on a small amount of hand-aligned data. Reynaert (2008) introduces a fully automatic, unsupervised system that tries to find variants of a dominant form. Finally, Boschetti et al. (2009) discuss a method that detects errors by aligning the outputs of different OCR systems.

3.2 Error Correction in Databases

It has been estimated that even well-maintained databases contain up to 5% noise, i.e., fields with incorrect or partially incorrect values (Redman, 1996). Van den Bosch et al. (2009) identified three types of errors commonly found in cultural heritage databases and computed the proportion of entries affected by these error types in a sample database from the natural history domain: *spelling errors* (1.23% of entries), which are orthographically erroneous forms such as *Linneus* instead of *Linnaeus*; *content errors* (34.8% of entries), which are incorrect values such as *Mexico* instead of *Brasil* in the location column; and *wrong-column errors* (4.37%), which are values that are in principle correct but were entered in the wrong database column, such as *Mexico* in a column entitled BIOTOPE. The high proportion of content errors arises from the use of dispreferred synonyms in the taxonomic columns, a slip-up that is relatively common for natural history databases since many species are known by multiple names and the preferred names can change over time. What is surprising is the relatively large proportion of wrong-column errors. This is often a consequence of a suboptimal database structure in which the individual columns are not clearly differentiated. Wrong-column errors have a particularly detrimental effect on column-based database querying; information that is in the wrong column will simply not be found.

Fully manual database cleaning is time-consuming and usually infeasible. Because error correction requires a certain amount of expert knowledge it cannot normally be fully automated. However, there have been a number of semi-automatic approaches. For example, Sporleder et al. (2006a) propose a method to semi-automatically detect wrong-column errors. They cast the problem as a text classification task and train a classifier to propose the most likely column for a given text string; if the predicted column deviates from the original column, the system flags an error and suggest the predicted column as a correction.

Content error detection in databases is typically modelled as outlier detection. However, many traditional approaches are not suited for cultural heritage databases because they treat cell values as numeric (e.g., Hawkins 1980) or categorical without internal structure (Marcus and Maletic, 2000; Knorr and Ng, 1998). However, this may not be the best approach, since cultural heritage databases contain large proportions of free text values, e.g., in columns such as BIOTOPE or SPECIAL REMARKS which should not be treated as atomic (Sporleder et al., 2006a).

A semi-automatic error detection tool called *Timpute* that is specifically geared towards cultural heritage data is presented by Van den Bosch et al. 2009. Similarly to previous approaches, the system exploits interdependencies between different columns. For example, if an artifact was collected near 'Madrid' the value of the country column is likely to be 'Spain'. The tool learns to predict the value of a cell based on the values of the other cells for a given instance. An error is flagged if the predicted value deviates from the original value. Timpute also specifies its confidence in the proposed value to help the user to assess its suggestion. Figure (1) shows an example. Here a Striped Sand Snake was accidentally entered into the database as belonging to the class 'Amphibia'. The tool detects this error and suggests the correction 'Reptilia' with a high confidence of 99% (due to the fact that the taxonomic class is relatively easy to infer given the species and genus information). While taxonomic errors could also be detected by checking the database against published taxonomies, this error detection method is general enough to also work for other fields, like PLACE or BIOTOPE, which are more difficult to check against existing resources.

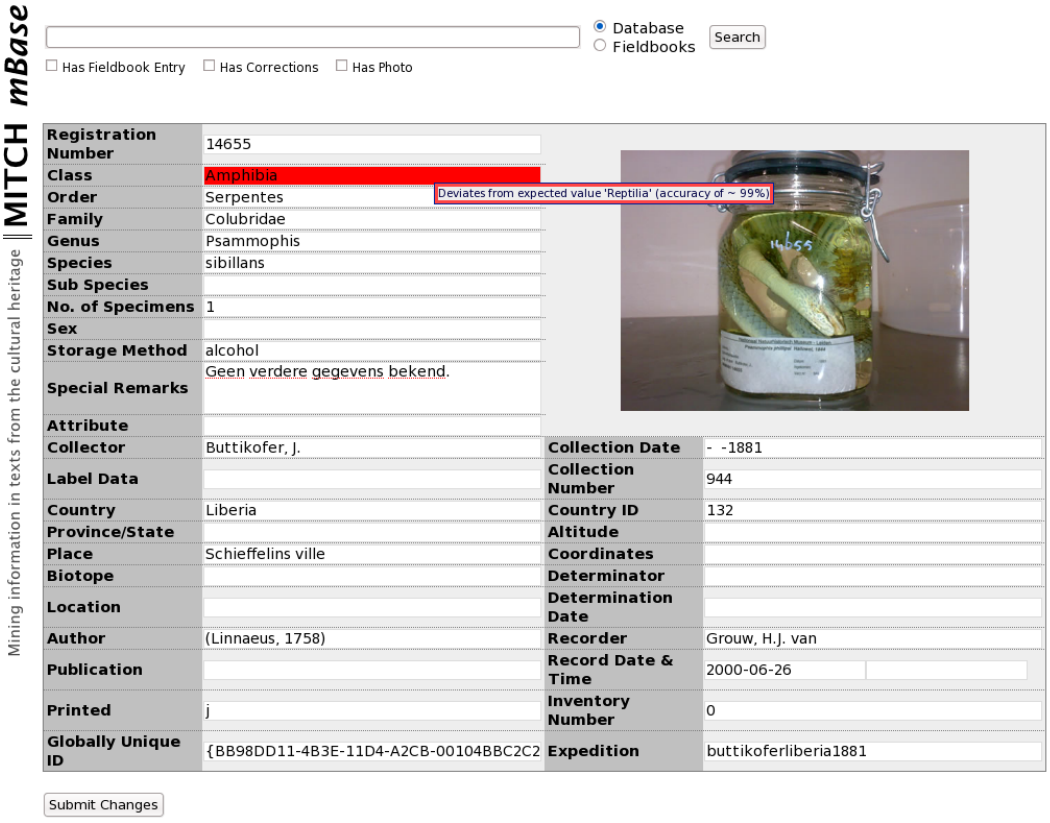

Figure 1: The Timpute Error Detector.

In addition to these generic methods, there has also been work on approaches that utilise domain-specific background knowledge to detect outliers. For example, geographic errors in species databases can be detected by applying automated georeferencing tools (Chapman, 2005).

4 Data Enrichment

Once the data have been cleaned, enrichment is the next step. Free or semi-structured texts, for example, can be enriched with structural information and ultimately turned into easily searchable databases (Section 4.1). Another common enrichment task involves automatically inferring descriptive metadata, for instance enriching non-textual data with key-words or descriptions in natural language (Section 4.2).

4.1 From Texts to Knowledge Bases

Digitised data in their raw form typically only permit key-word based search (see Section 2.1). To enable more sophisticated search strategies the data need to be represented in a more structured form in a knowledge base.

Converting Field Books to Databases If field books are available in computer-readable form, they can be semi-automatically segmented and turned into databases (Example (1) on page 9). Segmenting semi-structured texts is a well-known problem in NLP (often referred to as *field segmentation* (Grenager et al., 2005)). It can be modelled as a sequence labelling task in which each token in a field book entry is labelled with the database field it belongs to or with O ('other') if it should not be entered in the database (see the token-level annotations in example (4a) and the corresponding segmentation in (4b)). Consecutive tokens with identical labels are then entered into the corresponding database field.[15] Several supervised machine learning methods exist to do sequence labelling, along with some unsupervised methods (e.g. Grenager et al., 2005). Supervised methods are typically ruled out for CH domains due to a lack of annotated training data but a number of researchers have employed semi-supervised or bootstrapping techniques. Lendvai and Hunt (2008) propose an active learning approach in which a classifier is trained on a seed set of manually labelled examples and then applied to unlabelled data. Those examples about whose label the classifier is least confident are then passed on to a human expert for manual labelling before they are added to the training data and the classifier is re-trained. Using active learning typically means that fewer examples have to be labelled for a given performance level than if the labelled examples were chosen randomly. Abandoning the need for manual data annotation

[15] Field book entries tend not to split information of a given type, i.e., there are usually no discontinuous segments.

completely, Canisius and Sporleder (2007) propose a bootstrapping approach that exploits an existing, partially filled database to generate training examples automatically.

(4) a. Leptophis/GEN ahaetulla/SPEC ,/O road/LOC to/LOC Overtoom/LOC ,/O in/BIO bush/BIO above/BIO water/BIO in/REM the/REM process/REM of/REM eating/REM Hyla/REM minuta/REM 20-VI-1972/DATE ,/O RMNH/ID 38290/ID ./O

b. [Leptophis]$_{Gen(us)}$ [ahaetulla]$_{Spec(ies)}$, [road to Overtoom]$_{Loc(ation)}$, [in bush above water]$_{Bio(tope)}$ [in the process of eating Hyla minuta]$_{Rem(arks)}$ [20-VI-1972]$_{Date(OfCollection)}$, [RMNH 38290]$_{ID}$.

Event Recognition Often the type of information curators are interested in centers on the notion of an 'event', e.g., who collected or created an artefact when and where, or when was an archaeological site surveyed by whom. Field books are already semi-structured in that each record typically refers to one event (i.e., a collection event). However, in resources such as excavation reports, events are described in free texts. Extracting event information from those texts is a more difficult task. This problem is well-known from the field of information extraction. It can be split into three sub-tasks: event recognition, argument recognition and role assignment. For example, the sentences in example (5) refer to an EXCAVATION event with the roles: DATE: *December 2005*, LOCATION: *1.5 km east of Lough Neagh* and FIND: {*a large amount of post-medieval pottery sherds*, *several hundred flint pieces*}. Event extraction also needs to address coreference resolution, i.e., identifying descriptions that refer to the same event or the same argument. Both event (Bagga and Baldwin, 1999; Humphreys et al., 1997) and entity coreference (Chen and Martin, 2007; Mann and Yarowsky, 2003; Bagga, 1998) have attracted a lot of attention in NLP.

(5) A site 1.5km east of Lough Neagh was excavated in December 2005. A large amount of post-medieval pottery sherds and several hundred flint pieces were retrieved.

In languages like English events are usually evoked by verbs, so event recognition can be approximated by finding all verb phrases in a text (Ruotsalo et al., 2009). Alternatively, it is also possible to annotate data and train a classifier to identify event evoking expressions such as "was excavated" (Byrne, 2009).

The arguments of an event are often named entities such as locations or temporal expressions. While there has been a considerable body of work on named entity recognition (NER), generic NER systems typically cannot be applied directly to cultural heritage data because the inventory of named entity classes is to some extent domain dependent. For instance, in her work on relation extraction from free text archaeological descriptions, Byrne

(2009) uses 11 entity classes: organisation (e.g., *Ordnance Survey*), person name (e.g., *Vere Gordon Childe*), role (e.g., *architect*), sitetype (e.g., *chambered cairn*), artefact (e.g., *bronze axe*), place (for administrative place names, e.g., *River North Esk*), sitename (e.g., *Stones of Stenness*), address (e.g., archaeological text grid references, such as *HU 3754 3380*), period (e.g., *late Neolithic*), and date (e.g., *1st Jan 1980*). This is quite different from the typical named entity inventory for news texts. Location information, for instance, tends to be much more important and fine-grained for CH data. Moreover, the contexts in which specific named entities occur are different. For instance, in news texts the abbreviation *Ltd.* is typically a good cue for an organisation, while in CH texts it is of limited use. Consequently, generic NER systems tend to perform poorly on cultural heritage data even for known named entity classes (Sporleder et al., 2006b).

Some approaches for named entity recognition in cultural heritage data have circumvented this problem by manually annotating data and training a domain-specific named entity tagger (Byrne, 2007; Paijmans and Wubben, 2007). However, this is only viable if relatively large sets of similar data have to be processed. To avoid the need for manually annotated data, other studies have made use of domain-specific background knowledge. For instance, Paijmans and Brandsen (2009) and Sporleder et al. (2006b) extract lists of named entities (so-called gazetteer lists) from entity-specific columns in collection databases, such as the COLLECTOR column which contains expressions of type PERSON. These lists are then used to initiate a bootstrapping process in which the input text is matched against the gazetteer list and all unambiguous occurrences of known named entities are automatically labelled and then used as training material for a second-stage supervised classifier, which can then also exploit contextual clues. In a similar vein, Ruotsalo et al. (2009) use background knowledge in the form of domain-specific thesauri and name lists to improve argument recognition.

Finally, role assignment is typically modelled as a supervised machine learning task in which a classifier has to assign the correct role to each argument identified in the previous step (Byrne and Klein, 2009; Ruotsalo et al., 2009). To assign the correct role, most systems rely on linguistic cues, such as the part-of-speech and named entity tags of the arguments or their syntactic relation to the event evoking element. A final problem involves the merging of related events, e.g., "was excavated" and "retrieved" in example (5) (Bejan and Harabagiu, 2008; Humphreys et al., 1997).

4.2 Information Retrieval and Metadata Descriptions

Information access to non-textual CH data such as archaeological objects, paintings or historical audio recordings typically requires textual meta-information to index the data, for example in the form of key words, titles or short descriptions. For archaeological or natu-

ral history collections textual metadata are usually available in the form of field books or excavation reports; for other collections, metadata can be sparse. This applies especially to archives of audio-visual material, such as the Shoah Foundation Institute's Visual History Archive[16] or the Video Active portal.[17] While some basic, extrinsic information is typically available, e.g., who made a recording when and where, information access on a wider scale also requires precise and detailed information on the content of a recording. However, manual metadata creation can take up to three times the duration of the recordings for keyword annotation (Gazendam et al., 2009) and up to 15 hours for each recorded hour if the recordings are annotated in detail with (time-coded) descriptions (Gustman et al., 2002).

Consequently, there has been a significant amount of research on the automatic creation of metadata. Obtaining content information directly from the audio-video stream remains challenging. If a recording involves spoken language, automatic speech recognition can provide content information. However, the obtained transcriptions are often noisy, especially for historical recordings (Byrne et al., 2004). Concept detection in videos (Snoek et al., 2006) is also not yet robust enough for large scale semantic video indexing.

Several studies have exploited the availability of accompanying linguistic data to automatically create metadata content descriptions. Data sources can be subtitles (Netter, 1998), production scripts, commentaries and match reports for sports events (Saggion et al., 2002), online TV guides (Gazendam et al., 2006) or transcripts of speeches (van der Werff et al., 2007). Some data sources, such as subtitles or live commentaries provide very detailed information of what is happening or being said at a given moment. They can be automatically aligned with the audio-video stream to obtain time-coded descriptions for fine-grained multimedia retrieval (Christel et al., 2006). However, in some situations these low-level content descriptions may not be suitable. For example, a user might be interested in retrieving all speeches that were made in a specific historical context regardless of the precise content of the speech. To obtain high-level metadata descriptions other data sources, such as TV guides, are more appropriate because they abstract away from the details and also provide context information (Gazendam et al., 2006).

Sometimes metadata creation can be necessary even for collections of textual data sources, for example if the primary data consists of historical manuscripts which may employ unusual spelling and archaic vocabulary. *Historic document retrieval* deals with the task of retrieving relevant documents written in an older variant of the language. Typically users formulate their queries in modern language. The queries are then either automatically converted into the older form (*query translation*) (e.g. Ernst-Gerlach and Fuhr, 2007) or the historical manuscripts (or at least their index terms) are automatically translated into their modern equivalents (*document translation*) (e.g. Koolen et al., 2006). The second approach

[16] `http://college.usc.edu/vhi/`

[17] `http://www.videoactive.eu/`

has the advantage that the translated documents are easier to process by modern NLP tools and that the user can choose to read the modernised version of the document rather than the original one (Koolen et al., 2006). Most approaches to historic document retrieval are restricted to deviations in spelling rather than changes in vocabulary (e.g., semantic shifts, words falling out of or coming into use) or variations in syntax. For a recent approach on detecting diachronic semantic drift in a non-CH context see Cook and Stevenson (2010).

Rules for converting historical forms to their modern equivalents can be hand-crafted (Braun et al., 2002) or they can be automatically learned from parallel word lists (Ernst-Gerlach and Fuhr, 2007) or from historical and modern corpora which are automatically processed to find spelling variants (Koolen et al., 2006).

5 Language Technology and Cultural Heritage

In this article, I have outlined several areas in which language technology can help to improve information access to the vast collections of museums, libraries, and archives around the world. While the development of language technology tools for CH domains is still an emerging area, it is already clear that this combination offers enormous promise. Both areas stand to benefit from it. The work of a researcher or curator at a CH institute will be made significantly easier if all the relevant information can be found instantly.

For the natural language processing community the cultural heritage domain can serve as a challenging test-bed for the development of robust technology. While mainstream NLP recently tended to concern itself with specific processing tasks for small and relatively well-defined domains such as news wire, the focus is now shifting to new domains and genres, such as cultural heritage, biomedicine or weblogs. With this also comes a new awareness of the challenges and promises that these new data sources bring with them. For example, the standard supervised machine learning paradigm that has worked so well for the news domain cannot be transferred directly to other domains due to a lack of annotated data. Language technology researchers therefore have to develop more creative solutions, for example by incorporating available resources and domain-specific background knowledge, such as thesauri, ontologies or existing databases. This is already happening now and will continue to drive the field forward in future.

Acknowledgements

The author has been supported by the German Research Foundation DFG within Saarland University's Cluster of Excellence "Multimodal Computing and Interaction". I would also like to thank the anonymous reviewers, who provided valuable feedback on the first version

of this article.

References

Beatrice Alex, Claire Grover, Barry Haddow, Mijail Kabadjov, Ewan Klein, and Xinglong Wang. Assisted curation: Does text mining really help? In *Pacific Symposium on Biocomputing*, pages 556–567, 2008.

Amit Bagga. *Coreference, Cross-Document Coreference, and Information Extraction Methodologies*. PhD thesis, Deptartment of Computer Science, Duke University, 1998.

Amit Bagga and Breck Baldwin. Cross-document event coreference: Annotations, experiments, and observations. In *Proceedings of the ACL-99 Workshop on Coreference and Its Applications*, pages 1–8, 1999.

Mary Baker, Mehul Shah, David S. H. Rosenthal, Mema Roussopoulos, Petros Maniatis, TJ Giuli, and Prashanth Bungale. A fresh look at the reliability of long-term digital storage. In *Proceedings of the 1st ACM SIGOPS/EuroSys European Conference on Computer Systems*, pages 221–234, 2006.

Antonio M. Battro. From Malraux's imaginary museum to the virtual museum. In Parry (2010), pages 136–147. Reprinted from: *Xth World Congress Friends of Museums*, Sydney, September 13–18, 1999.

Cosmin Adrian Bejan and Sanda M. Harabagiu. A linguistic resource for discovering event structures and resolving event coreference. In *Proceedings of the Sixth International Language Resources and Evaluation(LREC-08)*, pages 2881–2887, 2008.

Tim Berners-Lee, James Hendler, and Ora Lassila. The semantic web. *Scientific American*, 284:34–43, 2001.

Howard Besser. Digital longevity. In Maxine Sitts, editor, *Handbook for Digital Projects: A Management Tool for Preservation and Access*, pages 155–166. Northeast Document Conservation Center, 2000.

Toine Bogers and Antal van den Bosch. Recommending scientific articles using CiteULike. In *Proceedings of the 2008 ACM Conference on Recommender Systems (RecSys-08)*, pages 287–290, 2008.

Lars Borin and Markus Forsberg. Something old, something new: A computational morphological description of Old Swedish. In *Proceedings of the LREC-08 Workshop on Language Technology for Cultural Heritage Data (LaTeCH-08)*, pages 9–16, 2008.

Federico Boschetti, Matteo Romanello, Alison Babeu, David Bamman, and Gregory Crane. Improving OCR accuracy for classical critical editions. In *Proceedings of the 13th European Conference on Research and Advanced Technology for Digital Libraries (ECDL-09)*, pages 156–167, 2009.

Loes Braun, Floris Wiesman, and Ida Sprinkhuizen-Kuyper. Information retrieval from historical corpora. In *Proceedings of the 3rd Dutch-Belgian Information Retrieval Workshop (DIR)*, pages 106–112, 2002.

Daan Broeder, Eric Auer, Marc Kemp-Snijders, Han Sloetjes, Peter Wittenburg, and Claus Zinn. Managing very large multimedia archives and their integration into federations. In *First Workshop in Very Large Digital Libraries (VLDL-08)*, 2008.

Kate Byrne. Nested named entity recognition in historical archive text. In *Proceedings of the IEEE International Conference on Semantic Computing (ICSC-07)*, pages 589–596, 2007.

Kate Byrne. *Populating the Semantic Web - Combining Text and Relational Databases as RDF Graphs*. PhD thesis, Edinburgh University, School of Informatics, 2009.

Kate Byrne and Ewan Klein. Automatic extraction of archaeological events from text. In *Computer Applications in Archaeology (CAA-09)*, 2009.

William Byrne, David Doermann, Martin Franz, Samuel Gustman, Jan Hajic, Douglas Oard, Michael Picheny, Josef Psutka, Bhuvana Ramabhadran, Dagobert Soergel, Todd Ward, and Wei-Jing Zhu. Automatic recognition of spontaneous speech for access to multilingual oral history archives. *IEEE Transactions on Speech and Audio Processing*, 12(4):420–435, 2004.

Sander Canisius and Caroline Sporleder. Bootstrapping information extraction from field books. In *Proceedings of the Joint Meeting of the Conference on Empirical Methods on Natural Language Processing and the Conference on Natural Language Learning (EMNLP-CoNLL-07)*, pages 827–836, 2007.

Arthur D. Chapman. Principles and methods of data cleaning. Technical report, Global Biodiversity Information Facility, Copenhagen, 2005. Version 1.0.

Ying Chen and James Martin. Towards robust unsupervised personal name disambiguation. In *Proceedings of the Joint Meeting of the Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL-07)*, pages 190–198, 2007.

Robert Chenhall and David Vance. The world of (almost) unique objects. In Parry (2010), pages 39–47. Reprinted from: Robert Chenhall and David Vance, *Museum Collections and Today's Computers*, Greenwood Press, New York, Westpoint, Connecticut, London: 1988, pages 3-13.

Michael G. Christel, Julieanna Richardson, and Howard D. Wactlar. Facilitating access to large digital oral history archives through Informedia technologies. In *Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries (JCDL-06)*, pages 194–195, 2006.

Paul Cook and Suzanne Stevenson. Automatically identifying changes in the semantic orientation of words. In *Proeceedings of the Seventh International Conference on Language Resources and Evaluation (LREC-10)*, pages 28–34, 2010.

Nick Crofts, Martin Doerr, Tony Gill, Stephen Stead, and Matthew Stiff, editors. *Definition of the CIDOC Conceptual Reference Model*. ICOM/CIDOC CRM Special Interest Group, 2009.

Hal Daumé III. Frustratingly easy domain adaptation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL-07)*, pages 256–263, 2007.

Martin Doerr, Kurt Schaller, and Maria Theodoridou. Integration of complementary archaeological sources. In *Proceedings of the Computer Applications and Quantitative Methods in Archaeology Conference*, pages 13–17, 2004.

Andrea Ernst-Gerlach and Norbert Fuhr. Retrieval in text collections with historic spelling using linguistic and spelling variants. In *Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital libraries (JCDL-07)*, pages 333–341, 2007.

Geraldine Fabrikant. The good stuff in the backroom. New York Times, March 12, 2009. `http://www.nytimes.com/2009/03/19/arts/artsspecial/19TROVE.html` (last accessed 30.01.2010).

Luit Gazendam, Véronique Malaisé, Guus Schreiber, and Hennie Brugman. Deriving semantic annotations of an audiovisual program from contextual texts. In *First International Workshop on Semantic Web Annotations for Multimedia (SWAMM-06)*, 2006.

Luit Gazendam, Christian Wartena, Véronique Malaisé, Guus Schreiber, Annemieke de Jong, and Hennie Brugman. Automatic annotation suggestions for audiovisual archives: Evaluation aspects. *Interdisciplinary Science Reviews*, 34(2-3):172–188, 2009.

Daniel Gildea. Corpus variation and parser performance. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-01)*, pages 167–202, 2001.

Trond Grenager, Dan Klein, and Christopher D. Manning. Unsupervised learning of field segmentation models for information extraction. In *Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL-05)*, pages 371–378, 2005.

Samuel Gustman, Dagobert Soergel, Douglas Oard, William Byrne, Michael Picheny, Bhuvana Ramabhadran, and Douglas Greenberg. Supporting access to large digital oral history archives. In *Proceedings of the Joint Conference on Digital Libraries*, pages 18–27, 2002.

Douglas M. Hawkins. *Identification of Outliers*. Chapman & Hall, London, 1980.

Mauricio A. Hernández and Salvatore J. Stolfo. Real-world data is dirty: Data cleansing and the merge/purge problem. *Journal of Data Mining and Knowledge Discovery*, 2(1): 9–37, 1998.

Laura Hollink, Bouke Huurnink, Michiel van Liempt, Johan Oomen, Annemieke de Jong, Maarten de Rijke, Guus Schreiber, and Arnold Smeulders. A multidisciplinary approach to unlocking television broadcast archives. *Interdisciplinary Science Reviews*, 34(2): 257–271, 2009.

Jeroen Horlings. CD-R's binnen twee jaar onleesbaar. PC Active, 2003. `http://www.pc-active.nl/component/content/article/10508` (last accessed 27.01.2010).

Kevin Humphreys, Robert J. Gaizauskas, and Saliha Azzam. Event coreference for information extraction. In *Proceedings of the Workshop On Operational Factors In Practical Robust Anaphora Resolution For Unrestricted Texts*, pages 75–81, 1997.

Eero Hyvönen, Tuukka Ruotsalo, Thomas Häggström, Mirva Salminen, Miikka Junnila, Mikko Virkkilä, Mikko Haaramo, Eetu Mäkelä, Tomi Kauppinen, and Kim Viljanen. CultureSampo-Finnish culture on the semantic web: The vision and first results. In K. Robering, editor, *Information Technology for the Virtual Museum*, pages 25–36. LIT Verlag, 2007.

Antoine Isaac, Stefan Schlobach, Henk Matthezing, and Claus Zinn. Integrated access to cultural heritage resources through representation and alignment of controlled vocabularies. *Library Review*, 57(3):187–199, 2008.

Amy Isard, Jon Oberlander, Ion Androutsopoulos, and Collin Matheson. Speaking the users' languages. *IEEE Intelligent Systems*, 18(1):40–45, 2003.

Nikiforos Karamanis, Ian Lewin, Ruth Seal, Rachel Drysdale, and Ted Briscoe. Integrating natural language processing with FlyBase curation. In *Proceedings of the Pacific Symposium on Biocomputing (PSB-07)*, pages 245–256, 2007.

Edwin M. Knorr and Raymond T. Ng. Algorithms for mining distance-based outliers in large datasets. In *Proceedings of the 24th International Conference on Very Large Data Bases (VLDB-98)*, pages 392–403, 1998.

Okan Kolak and Philip Resnik. OCR post-processing for low density languages. In *Proceedings of the Human Language Technology Conference and the Conference on Empirical Methods in Natural Language Processing(HLT-EMNLP-05)*, pages 867–874, 2005.

Stasinos Konstantopoulos, Vangelis Karkaletsis, and Dimitris Bilidas. An intelligent authoring environment for abstract semantic representations of cultural object descriptions. In *Proceedings of the ACL-09 Workshop on Language Technology and Resources for Cultural Heritage, Social Sciences, Humanities, and Education (LaTeCH-SHELT&R-09)*, pages 10–17, 2009.

Marijn Koolen, Frans Adriaans, Jaap Kamps, and Maarten de Rijke. A cross-language approach to historic document retrieval. In *Proceedings 28th European Conference on Information Retrieval (ECIR-06)*, pages 407–419, 2006.

Edmund Lee, editor. *MIDAS: A Manual and Data Standard for Monument Inventories*. English Heritage, Swindon, 1998.

Piroska Lendvai and Steve Hunt. From field notes towards a knowledge base. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC-08)*, 2008.

Vladimir I. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. *Cybernetics and Control Theory*, 10(8):707–710, 1965.

David Lewis, Yiming Yang, Tony Rose, and Fan Li. RCV1: a new benchmark collection for text categorization research. *Journal of Machine Learning Research*, 5:361–397, 2004.

Daniel Lopresti. Performance evaluation for text processing of noisy inputs. In *Proceedings of the 20th Annual ACM Symposium on Applied Computing (Document Engineering Track)*, pages 759–763, 2005.

Daniel Lopresti. Optical character recognition errors and their effects on natural language processing. In *Proceedings of the ACM SIGIR Workshop on Analytics for Noisy Unstructured Text Data*, pages 9–16, 2008.

Véronique Malaisé, Lora Aroyo, Hennie Brugman, Luit Gazendam, Annemieke de Jong, Christain Negru, and Guus Schreiber. Evaluating a thesaurus browser for an audio-visual archive. In *Proceedings of the 15th International Conference on Knowledge Engineering and Knowledge Management (EKAW-06)*, pages 272–286, 2006.

Gideon S. Mann and David Yarowsky. Unsupervised personal name disambiguation. In *Proceedings of the Seventh Conference on Natural Language Learning (CoNLL) at HLT-NAACL 2003*, pages 33–40, 2003.

Andrian Marcus and Jonathan I. Maletic. Utilizing association rules for identification of possible errors in data sets. Technical Report TR-CS-00-04, The University of Memphis, Division of Computer Science, 2000.

David McClosky, Eugene Charniak, and Mark Johnson. Reranking and self-training for parser adaptation. In *Proceedings of the Association for Computational Linguistics (COLING-ACL-06)*, pages 337–344, 2006.

Sean M. McNee. *Meeting User Information Needs in Recommender Systems*. PhD thesis, University of Minnesota, 2006.

Taesun Moon and Jason Baldridge. Part-of-speech tagging for Middle English through alignment and projection of parallel diachronic texts. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL-07)*, pages 390–399, 2007.

National Information Standards Organization. Understanding metadata. NISO Press, 2004.

Klaus Netter. POP-EYE and OLIVE - Human language as the medium for cross-lingual multimedia information retrieval. In *Proceedings of the 2nd International Conference on Quality and Standards in Audiovisual Language Transfer. Languages and the Media*, 1998.

J. Scott Olsson and Douglas W. Oard. Improving text classification for oral history archives with temporal domain knowledge. In *30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 623–630, 2007.

Hans Paijmans and Alex Brandsen. What is in a name: Recognizing monument names from free-text monument descriptions. In *Proceedings of the 18th Annual Belgian-Dutch Conference on Machine Learning*, pages 2–6, 2009.

Hans Paijmans and Sander Wubben. Memory based learning and the interpretation of numbers in archaeological reports. In *Proceedings of the 7th Dutch-Belgian Information Retrieval Workshop (DIR 2007)*, pages 51–56, 2007.

Ross Parry, editor. *Museums in a Digital Age*. Leicester Readers in Museum Studies. Routledge, London / New York, 2010.

Marco Pennacchiotti and Fabio Massimo Zanzotto. Natural language processing across time: an empirical investigation on Italian. In *Proceedings of the 6th International Conference on Natural Language Processing (GoTAL-08)*, pages 371–382, 2008.

Toni M. Rath, R. Manmatha, and Victor Lavrenko. A search engine for historical manuscript images. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR-04)*, pages 369–376, 2004.

Thomas C. Redman. *Data Quality for the Information Age*. Artech House Inc., 1996.

Martin Reynaert. *Text-Induced Spelling Correction*. PhD thesis, Tilburg University, 2005.

Martin Reynaert. Non-interactive OCR post-correction for giga-scale digitization projects. In *Proceedings of the Computational Linguistics and Intelligent Text Processing 9th International Conference (CICLing-08)*, pages 617–630, 2008.

Laura Rimell and Stephen Clark. Adapting a lexicalized-grammar parser to contrasting domains. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-08)*, pages 475–484, 2008.

Vitor J. Rocio, Mário Amado Alves, Gabriel Pereira Lopes, Maria Francisca Xavier, and Graça Vicente. Automated creation of a partially syntactically annotated corpus of Medieval Portuguese using Contemporary Portuguese resources. In *Proceedings of the 1999 ATALA workshop on Treebanks*, pages 59–67, 1999.

Eiríkur Rögnvaldsson and Sigrún Helgadóttir. Morphological tagging of Old Norse texts and its use in studying syntactic variation and change. In *Proceedings of the LREC Workshop on Language Technology for Cultural Heritage Data (LaTeCH-08)*, pages 40–46, 2008.

Douglas Roland and Daniel Jurafsky. How verb subcategorization frequencies are affected by corpus choice. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (ACL-COLING-98)*, pages 1122–1128, 1998.

Tuukka Ruotsalo and Eero Hyvönen. A method for determining ontology-based semantic relevance. In *Proceedings of the International Conference on Database and Expert Systems Applications (DEXA-07)*, pages 680–688, 2007.

Tuukka Ruotsalo, Lora Aroyo, and Guus Schreiber. Knowledge-based linguistic annotation of digital cultural heritage collections. *IEEE Intelligent Systems*, 34(2):64–75, 2009.

Horacio Saggion, Hamish Cunningham, Kalina Bontcheva, Diana Maynard, Cris Ursu, Oana Hamza, and Yorick Wilks. Access to multimedia information through multisource and multilanguage information extraction. In *Proceedings of the 6th International Conference on Applications of Natural Language to Information Systems*, pages 160–171, 2002.

Cees G. M. Snoek, Marcel Worring, Jan C. van Gemert, Jan-Mark van Geusebroek, and Arnold W. M. Smeulders. The challenge problem for automated detection of 101 semantic concepts in multimedia. In *Proceedings of the 14th annual ACM international conference on Multimedia*, pages 421–430, 2006.

Caroline Sporleder, Marieke van Erp, Tijn Porcelijn, and Antal van den Bosch. Correcting 'wrong-column' errors in text databases. In *Proceedings of the Annual Machine Learning Conference of Belgium and The Netherlands (Benelearn-06)*, pages 49–56, 2006a.

Caroline Sporleder, Marieke van Erp, Tijn Porcelijn, Antal van den Bosch, and Pim Arntzen. Identifying named entities in text databases from the natural history domain. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC-06)*, pages 1742–1745, 2006b.

Sargur N. Srihari, Chen Huang, and Harish Srinivasan. Search engine for handwritten documents. In *Document Recognition and Retrieval (DRR-05)*, pages 66–75, 2005.

Elaine Svenonius. Unanswered questions in the design of controlled vocabularies. *Journal of the American Society for Information Science*, 37(5):331–340, 1986.

Kazem Taghva and Eric Stofsky. OCRSpell: an interactive spelling correction system for OCR errors in text. *International Journal on Document Analysis and Recognition*, 3(3): 125–137, 2001.

Yoshimasa Tsuruoka, Yuka Tateishi, Jin-Dong Kim, Tomoko Ohta, John McNaught, Sophia Ananiadou, and Jun'ichi Tsujii. Developing a robust part-of-speech tagger for biomedical text. In *Advances in Informatics - 10th Panhellenic Conference on Informatics*, pages 382–392, 2005.

Antal van den Bosch, Marieke van Erp, and Caroline Sporleder. Making a clean sweep of cultural heritage. *IEEE Intelligent Systems*, 34(2):54–63., 2009.

Laurens van der Werff, Willemijn Heeren, Roeland Ordelman, and Franciska de Jong. Radio Oranje: Enhanced access to a historical spoken word collection. In *17th Meeting of Computational Linguistics in the Netherlands*, pages 207–218, 2007.

Yiwen Wang, Natalia Stash, Lora Aroyo, Laura Hollink, and Guus Schreiber. Using semantic relations for content-based recommender systems in cultural heritage. In *Proceedings of the Workshop on Ontology Patterns (WOP) at ISWC*, pages 16–28, 2009.

Svitlana Zinger, John Nerbonne, Lambert Schomaker, and Henny van Schie. Content-based text line comparison for historical document retrieval. In *Proceedings of the RANLP-07 Computational Phonology Workshop*, pages 79–84, 2007.